ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis in Quantitative Finance

3 July, 2022

Autoregressive Extremal Random Forests for Value at Risk

Name student:
Bas Peeters

Student number:
479569

Second assessor:
(Aishameriane) A Venes Schmidt

Supervisor:
prof.dr. (Chen) C Zhou

Abstract

Estimating Value at Risk (VaR) in commodity futures markets is essential for risk managers. However, not much research has been conducted on estimating extreme left-tail quantiles of returns in commodity markets. This paper first, shows that the Extremal Random Forest (ERF) method of Gnecco et al. (2022) outperforms other quantile estimation methods in an extensive simulation study. Secondly, the ERF method is used to predict extreme quantile levels of U.S. wages using a real-life data set. Finally, the Autoregressive ERF method is introduced and used to predict the daily VaR of five commodity futures. The estimates of this model, a GARCH(1,1) model, and a Conditional Autoregressive Value at Risk are tested for correct conditional coverage using Christoffersen tests. We conclude that Autoregressive ERF model is not a recommended method for estimating extreme level VaR of commodity futures.

Contents

1 Introduction

Quantile regression is an econometric method that differs from regular regressions by modelling conditional quantiles instead of the conditional mean (Koenker & Bassett, 1978). This method can be applied to many situations such as medicine, survival analysis, finance (Value at Risk), economics, and environmental modelling (Yu et al., 2003).

In this paper, we are interested in estimating extreme quantiles of a response variable, conditional on a set of predictors, where the predictor space is large. One of the main problems arises when the expected number of observations in the tail is close or equal to 0. This could lead to a large bias when using an empirical estimator. The second potential problem is the high-dimensional predictor space. If we have many predictors, there might not be any observations close to the value on which the response variable is conditioned. Then simple regression models will lead to additional bias. Using methods from extreme value theory (de Haan & Ferreira, 2006) can counter this problem. These methods approximate the tails of data by extrapolation to quantile levels beyond the data range. The first extreme quantile regression methods made use of linear functions (Chernozhukov, 2005), additive models (Chavez-Demoulin & Davison, 2005), or non-parametric approaches (Beirlant et al., 2004). These however, do not scale well when using a large number of predictors or are not flexible enough to model complex response surfaces.

This second problem can be tackled with specific machine learning approaches. Most of these methods are extensions of the original random forests of Breiman (1996). The Quantile Random Forest (QRF) is a non-parametric method for estimating conditional quantiles in high-dimensional predictor spaces (Meinshausen & Ridgeway, 2006). Another method is from Athey et al. (2019), who developed Generalized Random Forests (GRF) and used it for different tasks, including non-parametric quantile regressions. However, these forest-based methods do not work well when the quantile level of interest is extreme.

Recent techniques have been introduced which bring together extreme value theory and machine learning methods to tackle the problems of extreme quantile regressions with high dimensional predictor spaces (Gnecco et al., 2022; Velthoen et al., 2021). Gnecco et al. (2022) introduce a new method for extreme quantile estimation, which they call Extremal Random Forests (ERF). This method is used to estimate extreme quantiles in heavy-tailed distributions. When the sample size is not large enough, there are often few or no data points beyond the extreme quantile. If this is the case, regular quantile regressions might lead to biased estimations. Thus, a different approach is needed. The method of Gnecco et al. (2022) extrapolates data beyond a certain intermediate threshold. It is assumed that the exceedances over this intermediate threshold follow a Generalized Pareto Distribution (GPD) (de Haan & Ferreira, 2006). We will rely on an approximation of the GPD to extrapolate beyond the data range. Thus, to estimate the extreme quantile, Gnecco et al. (2022) first estimate the parameters of this GPD conditional on the predictor vector. The method they use for this estimation is weighted maximum likelihood, where the weights are obtained from GRF with a quantile loss function. The intermediate quantile also needs to be estimated for the GPD function. To do this, they use GRF with a

quantile loss function again. Gnecco et al. (2022) show that ERF leads to consistent estimations of the GPD parameters under certain assumptions. In a simulation study, they also show that their method outperforms six other methods for quantile estimation. Finally, they estimate the extreme quantiles of U.S. wage data using a real-life data set.

There are many possible applications for extreme quantile regressions, and Gnecco et al. (2022) have applied the ERF method to only one. Another possible application is the estimation of Value at Risk (VaR). VaR is one of the most used measures of risk of financial returns (Duffie & Singleton, 2012). It is defined as the minimum return that could occur over a given holding period with a specified confidence level.

The most classic approaches for estimating VaR are the historical simulation method, the variance-covariance method and the Monte Carlo method. Other methods based on quantile regressions have also gained empirical support (Engle & Manganelli, 2004; Gaglianone et al., 2011; Taylor, 2008). Furthermore, multiple methods using extreme value theory have been used to estimate quantiles of returns in case of extreme levels of τ (Odening & Hinrichs, 2002; Gencay & Selçuk, 2004).

The Conditional Autoregressive Value at Risk (CAViaR) uses quantile regressions to estimate VaR. Engle & Manganelli (2004) argue that because asset return volatility is clustered over time their distribution must be correlated. They introduce a generic model with asset returns as dependant variable and lagged values of the estimated VaR as one of the regressors. They implement their methodology on a real-life data set and find that the CAViaR method performs well on quantile levels of 5% and 1%. They however, do not test their method on more extreme quantile levels.

Commodity futures are used for hedging purposes and by traders to make directional price bets. The purpose of hedging is to prevent losses from potential unfavourable price changes of physical commodities that a company is exposed to. Estimating the risk associated with holding physical commodities and commodity futures is essential for these companies to decide their hedging position. Speculative traders often use a high degree of leverage on futures contracts, hence estimating the risk involved in trading futures contracts is also crucial to them. In historical periods like the first half of 2022, the commodity futures markets exhibited enormous volatility due to multiple economic and political factors (war in Ukraine, global supply chain problems and high inflation levels). In such periods, it is even more crucial for companies exposed to commodities to understand their VaR.

Martins-Filho et al. (2018) introduce a new non-parametric approach to measure extreme VaR, which relies on extrapolation of the tail with a GPD. They also apply this new method to a data set of daily commodity futures returns. While Gnecco et al. (2022) show ERF outperforms other extreme quantile regression methods, this method has not yet been applied to measure financial risk. This leads to the following research question:

Is the Extremal Random Forest method of Gnecco et al. (2022) an acceptable method for estimating daily extreme Value at Risk of commodity futures?

To answer this research question, this paper will first use multiple simulations of heavy-tailed distributions from Gnecco et al. (2022) to evaluate the ERF method against other methods of quantile estimation. After that, the ERF method is used to estimate extreme quantiles of U.S. wage using a real data set. Finally, the ERF method will be used to measure the VaR of certain commodity futures.

In section 2, the methodology of the entire paper is explained. First, the ERF model of Gnecco et al. (2022) is elaborated on including his cross-validation method for hyperparameter tuning and two simulation setups. Then, two existing methods for estimating VaR are explained, and the new Autoregressive ERF method for VaR is introduced. In section **??** a data set consisting of U.S. wage data is discussed where the ERF method will be applied on. In section 3.2 the data set used to estimate VaR with the Autoregressive ERF method, consisting of 5 commodity futures prices, is explained. After this, the results of the two simulation setups, the U.S. wage experiment and the VaR experiment are discussed. The paper finishes with a conclusion and recommendation for further research.

2 Methodology

2.1 Extremal Random Forests

In the first part of this paper, we try to find the best estimator of the quantile $Q_x(\tau)$, $\tau \in (0,1)$, for a high level of the conditional distribution $Y|X = x$, with x a fixed predictor value. To achieve this, the ERF method from Gnecco et al. (2022) is first applied. Their notation is also used.

Let (X, Y) be a random vector of predictors $X \in \chi \subset \mathbb{R}^p$ and response $Y \in \mathbb{R}$ with χ a compact set. To tackle the problem of the τ-quantile potentially being out of range of the data, the tail of the conditional response variable is extrapolated. This is done by using asymptotic theory to fit a localized GPD. Let τ_0 be an intermediate quantile level. The distribution function of $Y - Q_x(\tau_0)$ conditional on $Y > Q_x(\tau_0)$ and $X = x$ has approximately a GPD (Gnecco et al., 2022), thus for any z,

$$P(Y - Q_x(\tau_0) \leq z | Y > Q_x(\tau_0), X = x) \approx G(z; \theta(x)), \tag{1}$$

where $Q_x(\tau_0)$ is the quantile function with intermediate quantile τ_0. $\theta(x) = (\sigma(x), \xi(x))$ are the scale and shape parameters respectively, which depend on the value of x. The right-hand side

of equation (1) is the GPD function given by,

$$G(z;\theta(x)) = 1 - \left(1 + \frac{\xi(x)}{\sigma(x)}z\right)_+^{-\frac{1}{\xi(x)}}, z > 0, \tag{2}$$

where the scale ($\sigma(x)$) and shape ($\xi(x)$) parameters both are continuous functions. If we combine the GPD function with Bayes' theorem, we obtain a final expression for the approximation of the τ-quantile of Y conditional on the predictor vector (Gnecco et al., 2022),

$$Q_x(\tau) \approx Q_x(\tau_0) + \frac{\sigma(x)}{\xi(x)}\left[\left(\frac{1-\tau}{1-\tau_0}\right)^{-\xi(x)} - 1\right]. \tag{3}$$

To obtain the extreme quantile on the left-hand side in equation (3), the intermediate quantile $Q_x(\tau_0)$ is first estimated. This is not an extreme quantile so the τ_0-quantile potentially being out of range of the data is not an issue here. However, the problem of a high-dimensional predictor space still holds. Thus, the GRF with quantile loss of Athey et al. (2019) is used. GRF is a method which uses quantile regressions originally from Koenker & Bassett (1978). Quantile regressions minimize the expected quantile loss. An estimator for this minimizer is given by

$$\hat{Q}_x(\tau_0) = \arg\min_{q\in\mathbb{R}} \textstyle\sum_{i=1}^n w_n(x, X_i)\rho_\tau(Y_i - q), \tag{4}$$

where $w_n(x, X_i)$ for i = i,...,n, are localizing similarity weights around the predictor. The quantile loss is given by $\rho_\tau(Y_i - q) = (Y_i - q)(\tau - 1_{Y_i<q})$, $q \in \mathbb{R}$, with $1_{Y_i<q}$ an indicator function which is 1 if $Y_i < q$ and 0 otherwise. Y_i are observations of the response variable. Multiple methods are available to obtain these weights. To tackle the problem of complex quantile surfaces in larger predictor dimensions, obtaining the weights from a GRF with quantile loss function is the most efficient method (Athey et al., 2019).

The second estimation is of the conditional GPD parameter $\theta(x)$. To do this, Gnecco et al. (2022) define the exceedances of training data over the estimator of the conditional quantile function $\hat{Q}_x(\tau_0)$ as

$$Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+, i = 1, ..., n. \tag{5}$$

We use the localizing weight functions $w_n(x, X_i)$ for $i = 1, ..., n$ to give more weight to the exceedances that carry more information on the conditional distribution of the response variable. These localizing weight functions are estimated using GRF. Finally, we estimate the conditional GPD parameters with weighted maximum likelihood. This leads to the following estimator

$$\hat{\theta}(x) = \arg\min_{\theta\in\Theta} L_n(\theta; x), \tag{6}$$

where $\hat{\theta}(x)$ is the estimator of the conditional GPD parameter $\theta(x)$, Θ is an arbitrarily large compact set and

$$L_n(\theta; x) = \textstyle\sum_{i=1}^n w_n(x, X_i)\ell_{\theta(x)}(Z_i)1_{Z_i>0}, x \in \chi. \tag{7}$$

With $1_{Z_i>0}$ an indicator function which is 1 if $Z_i > 0$ and 0 otherwise. $\ell_\theta(Z_i)$ is the negative log-likelihood of the i^{th} exceedance which is given by

$$\ell_{\theta(x)}(Z_i) = \log\sigma + \Big(1 + \frac{1}{\xi(x)}\Big)\log\Big(1 + \frac{\xi(x)}{\sigma(x)}\Big), \theta(x) \in (0, \infty) \times \mathbb{R}, \tag{8}$$

if $Z_i > 0$, and zero otherwise.

2.2 Hyperparameter tuning

Generalized random forests have several tuning parameters. The cross-validation approach presented by Gnecco et al. (2022) to tweak such hyperparameters is used in this research. The quantile loss is not a credible scoring function when $\tau \approx 1$ because there may be few or no test observations above this threshold. So instead, the GPD deviation is used as a cross-validation metric. Within each training set of n observations, M equally sized folds N_m are partitioned. The optimal tuning parameter α^* is then the minimizer of

$$CV(\alpha_j) = \sum_{m=1}^{M} \sum_{i \in N_m} \ell_{\hat{\theta}(X_i;\alpha_j)}(Z_i) 1_{Z_i>0}, \tag{9}$$

where α_j is a tuning parameter, ℓ_θ(z) is the deviance of the GPD and $1_{Z_i>0}$ is an indicator function with value 1 if $Z_i > 0$ and 0 otherwise. The optimal tuning parameter α^* is then found by minimizing $CV(\alpha_j)$. The minimum node size κ is the most important tuning parameter for ERF because it controls the complexity of the individual trees and thus also the similarity weights (Gnecco et al., 2022). Therefore, we will start by tuning this hyperparameter.

The shape parameter is notoriously difficult to estimate, and maximizing the GPD log-likelihood can lead to convergence issues with small sample sets. In general, penalization can help lower an estimator's variance at the cost of increased bias. The estimates $\hat{\xi}(x)$ are shrunk to a constant shape parameter ξ_0, which is the unconditional fit derived by minimizing the GPD deviation in equation (7) with constant weights. The following penalized negative log-likelihood proposed by Gnecco et al. (2022) is used

$$\hat{\theta}(x) = \underset{(\sigma,\xi)=\theta\in\Theta}{\arg\min} \frac{1}{(1-\tau_0)} L_n(\theta; x) + \lambda(\xi - \xi_0)^2, \tag{10}$$

where $\lambda \geq 0$ is the second parameter to be tuned. When λ is large, we have a more simple model with a more constant shape and when λ is small we have a more varying shape over the predictor space and thus a more complex model.

2.3 Simulation

The ERF method is evaluated in a simulation study following the same steps as Gnecco et al. (2022). This method is also used by Athey et al. (2019), only they simulate the tail of the response variable with a Gaussian distribution instead of a heavy-tailed distribution. We consider 2 setups. Within each setup, we repeat a simulation $m = 50$ times, with one simulation having $n = 2000$ training observations.

The setup for the first experiment is as follows, $X \sim U_p$ with U_p a uniform distribution with p dimensions, $[-1, 1]^p$. The response variable conditional on the predictor vector is simulated from a Student's t-distribution with $v = 4$ degrees of freedom multiplied by a step function. So, $Y|X = x \sim s(x)T_4$ with $s(x) = 1 + 1_{X_1>0}$ for $x \in \mathbb{R}^p$. This leads to the shape parameter being constant, $\xi(x) = 1/v(x) = 0.25$. The setup for the second experiment only differs in the shape parameter, which takes on multiple values $\xi(x) = 0$ (light-tailed Gaussian), $\xi(x) = 1/4$, and $\xi(x) = 1/3$ (heavy-tailed Gaussian).

We evaluate ERF against 6 other methods. Two other forest-based methods are considered, QRF (Meinshausen & Ridgeway, 2006) and GRF (Athey et al., 2019). Their tuning parameters are set to the default values. Two other extrapolation methods are also considered, EGAM uses generalized additive models to estimate GPD parameters (Youngman, 2019), and GBEX uses Gradient Boosting Machines to estimate these parameters (Velthoen et al., 2021). A fifth evaluation method is the unconditional GPD, which assumes the entire distribution is GPD. Finally, the EGP tail method of Taillardat et al. (2019), which assumes the entire distribution follows an extended Pareto distribution, is also used as evaluation method.

2.4 Value at Risk for commodity futures

In this section, the ERF method of Gnecco et al. (2022) is tested on a real financial data set. The Conditional extreme left tail quantiles for 5 series of commodity futures returns are estimated. These quantiles are commonly known as Value at Risk (VaR). The exact data set used is elaborated on in section 3.2. Besides the ERF method, two other well-known methods are used to predict VaR. The generalized autoregressive conditional heteroskedasticity (GARCH) model (Lamoureux & Lastrapes, 1990) and, the Conditional Autoregressive Value at Risk (CAViaR) model (Engle & Manganelli, 2004) are used as methods to evaluate against. We define one-day VaR of commodity futures as follows:

$$P[r_{t+1} \leq VaR_t(1-\tau)|I_t] = F_{t+1|t}(VaR_t(1-\tau)) = \tau, \tag{11}$$

where $VaR_t(1-\tau)$ is the τ-th quantile of the conditional distribution of the return at time $t+1$, r_{t+1}. $F_{t+1|t}$ is the cumulative distribution function of r_{t+1} conditional on the information set up to time t, I_t. We are interested in predicting extreme left-tail quantiles of returns corresponding to $\tau \in \{0.005, 0.01, 0.05\}$.

2.4.1 GARCH

GARCH estimates volatility in financial markets and translates that to VaR by making the assumption of Gaussian distributed standardized errors. We assume the following equation for the return at time t, $r_t = \mu + \epsilon_t$, where μ is the unconditional mean. The unexpected return ϵ_t follows the following equation, $\epsilon_t = z_t\sigma_t$, where z_t are the standardized returns which follow a normal distribution and σ_t is the conditional standard deviation. We specify the GARCH(1,1) as,

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{12}$$

The conditional variance at time t is σ_t^2 and the lagged unexpected squared return is ϵ_{t-1}^2. Var estimates at time t can then be obtained using the Gaussian distribution of the standardized errors by the following equation,

$$VaR_t(1-\tau) = \hat{\mu} + z_\tau \hat{\sigma}_{t+1|t}, \tag{13}$$

where $\hat{\mu}$ is the estimated unconditional mean of returns, z_τ is the τ-th quantile of the normal distribution and $\hat{\sigma}_{t+1|t}$ are one step ahead volatility estimates form equation (12). In practice, GARCH(1,1) can be modelled using the multiple R packages. Hill & McCullough (2019) evaluate the three most used packages and conclude that the package rugarch is preferred. Thus, the rugarch package is also used in this analysis.

2.4.2 CAViaR

Another model which will be used to evaluate against is the CAViaR model of Engle & Manganelli (2004). They created a general conditional autoregressive quantile model with multiple specifications. The following model is the most simple case of CAViaR. Let $r_{t=1}^T$ be a vector of asset returns and $Q_t(\tau)$ the τ-th quantile of the conditional distribution of returns formed at time $t-1$, so $Q_t(\tau) = VaR_t(1-\tau)$. We then have the following 'symmetric absolute value' expression for this conditional quantile,

$$Q_t(\tau) = \beta_0 + \sum_{i=1}^{q} \beta_i Q_{t-i}(\tau) + \sum_{j=1}^{s} \beta_j |r_{t-j}|, \tag{14}$$

where $Q_{t-i}(\tau)$ is the lagged τ-th quantile of returns, estimated at time $t-i$. The absolute value of returns at time $t-j$ is $|r_{t-j}|$. Furthermore, β_i is a vector of parameters for $i = 0, ..., q$ with β_0 an intercept. The regressor $Q_{t-i}(\tau)$ lets the predicted quantile change smoothly over time, and the regressor $|r_{t-j}|$ lets the VaR react equally to positive and negative return shocks (Engle & Manganelli, 2004). The dimension of the parameter β is then $q+s+1$ and is estimated by minimizing the regression quantile loss function as introduced by Koenker & Bassett (1978). In their most simple symmetric absolute value model, Engle & Manganelli (2004) let $q = 1$ and $s = 1$, which is the model we will use. In practice, we estimate VaR with CAViaR using the CAViaR package in R.

2.4.3 Autoregressive ERF

To answer the research question, the ERF method is used to predict the VaR of commodity futures. The response variable is the return of a commodity futures price r_t, multiplied by -1, which we define as Y_t. We do this because the ERF method estimates conditional quantiles in the right tail, but we want to estimate conditional quantiles in the left tail. So $VaR_t(1-\tau)$ is the τ-th quantile of r_t and the $(1-\tau)$-th quantile of Y_t. The predictors are multiple lags of the returns Y_{t-i} together with the predicted quantile at $t-1$. Because the predicted quantile at $t-1$ already holds information of the previous values, it seems logical to only take one lag. For the

actual return, however more lags than 1 could hold information and thus be a relevant predictor.

Per commodity, the data set with n observations is split in half. The first half is used for hyperparameter tuning, and in the second half the hyperparameters are fixed. Within each half we use the first 100 observations to estimate VaR for observation 101 with the historical simulation method. Then, the remaining observations $(0.5n - 100)$ are split in $M = 3$ folds N_m. The first and second fold both contain $1/5$ of the remaining observations and the third fold N_3 contains $3/5$ of the remaining observations. For the first fold (N_1), VaR is calculated with the historical simulation method with a rolling window of 100. Then, these estimates are used as the lagged quantile estimates (predictors) to fit the ERF object. This fitted ERF object is then used to estimate quantiles in an autoregressive way in the second fold (N_2). Then, the estimated quantiles of the second fold are used as lagged estimated quantiles (predictors) to fit ERF again. This fitted ERF object is then, used to estimate quantiles in the third fold (N_3). The VaR estimates obtained from this fold are used for evaluation. We thus have estimates for $\frac{3}{5} * (0.5n - 100)$ observations.

For hyperparameter tuning, the GPD deviance is used, but with a slightly different metric than the metric in equation (10). We do not use a cross-validation scheme because our data set is a time series, and we are interested in predicting the future with past information. We only use the most extreme quantile of Y_t for hyperparameter tuning, $\tau = 0.99$. The reason we do this is as follows. Fitting the Autoregressive ERF model happens per quantile level because one of the predictors is the lagged quantile. Tuning could be done separately for all three quantile levels, but this significantly increases computation time, so we choose the intermediate quantile level $\tau = 0.99$. The metric we use is the following:

$$F(\alpha_j) = \textstyle\sum_{i \in N_m} \ell_{\hat{\theta}(X_i;\alpha_j)}(Z_i) 1_{Z_i > 0}, \tag{15}$$

where α_j is a tuning parameter, $\ell_\theta(z)$ is the deviance of the GPD, and $1_{Z_i>0}$ is an indicator function with value 1 if $Z_i > 0$ and 0 otherwise. We then find the optimal tuning parameter α^* by minimizing $F(\alpha_j)$. The minimum node size κ is the most important tuning parameter for ERF because it controls the complexity of the individual trees (Gnecco et al., 2022). Thus, this is the only actual hyperparameter which will be tuned. We will however also tune the number of lags for the lagged return Y_{t-i}. For $\kappa \in \{10, 40, 100\}$ and $i \in \{1, \{1, 2\}, \{1, 2, 3\}\}$ we then have a grid with 9 combinations to choose from. Furthermore, the penalty value λ is fixed at 0.01 to reduce the shape parameters variance.

2.4.4 Christoffersen tests

The Christoffersen tests are used to check if the estimates of the models are acceptable. The first test for correct unconditional coverage evaluates whether the fraction of observations violating the estimated quantile level is equal to the coverage probability (Christoffersen, 1998). For daily returns, a violation is defined as a day where an observation exceeds the estimated quantile level. In the following two models, we let state $\{0\}$ indicate a day where a violation did not occur and state $\{1\}$ a day where a violation did occur. The likelihood ratio test statistic for correct

unconditional coverage is

$$LR_{UC} = -2ln\{[(1-p)^{T_0}p^{T_1}]/[(1-\pi)^{T_0}\pi^{T_1}]\} \sim \chi^2(1), \tag{16}$$

where p is the coverage probability and π is the percentage of violations. T_1 is the number of days where a violation occurred and T_0 the number of days where a violation did not occur. The second Christoffersen test is the test for independence. This test compares the probability of having state $\{1\}$ follow state $\{1\}$ with the probability of state $\{1\}$ follow state $\{0\}$. Ultimately, this tests if the VaR violations are clustered. The likelihood ratio statistic of the independence test is,

$$LR_{IND} = -2ln\{[(1-\pi)^{T_0}\pi^{T_1}]/[(1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}]\}, \tag{17}$$

where T_{ij} represents the number of occurrences that state i follows state j where $i, j \in \{\{0\}, \{1\}\}$. Furthermore, we have $\pi_{01} = T_{01}/(T_{00}+T_{01})$ and $\pi_{11} = T_{11}/(T_{10}+T_{11})$. Both LR_{UC} and LR_{IND} have an asymptotic chi-squared distribution with one degree of freedom under the null hypothesis of independence and correct unconditional coverage (Christoffersen, 1998). If the null hypothesis for both tests are not rejected, the estimates have correct conditional coverage (Christoffersen, 1998). This means the VaR model is acceptable.

3 Data

3.1 U.S. wage structure

To evaluate ERF, we consider a data set used by J. Angrist et al. (2006), who applied quantile regressions on the U.S wage structure. The data from Harvard Dataverse (J. D. Angrist et al., 2009) consists of 1980 U.S. census microdata of 65,023 U.S.-born black and white men aged 40–49, with five to twenty years of education, positive annual income, and a positive amount of hours worked in the year preceding the census. Because of the large number of observations, this data set is excellent for evaluating the performance of the different estimation methods. We compare the performance of ERF, GBEX, GRF, and the unconditional GPD.

The response is weekly wage, stated in 1989 U.S. dollars and is calculated by dividing the annual income by the number of weeks worked. The predictors consist of the numerical variables, age and years of schooling and a categorical variable, whether the person is black or white. We consider two setups. The first is the same setup Gnecco et al. (2022) follows. They generate 10 extra random predictors independently from a uniform distribution on the interval [-1,1], resulting in a predictor space with a dimension of $p = 13$. Adding these 10 random predictors seems counter-intuitive because the ERF method proposedly is a more suitable method in high-dimensional predictor spaces than other methods. We are then creating an artificial environment where ERF more likely outperforms other methods. This is of course not the intention of a field experiment. Due to this, we will apply ERF in a setup with the extra predictors and in a setup without the extra predictors. Furthermore, hyperparameter tuning of κ and λ will be done with the cross-validation scheme from 2.2, depending on the performance of this scheme.

3.2 Commodity futures

To answer the research question, the ERF method is used to predict the VaR of commodity returns. the data set used consists of 5 series of historical daily returns of commodity futures prices. The data is obtained from the database Global Financial Data. This database contains the daily commodity futures prices of the most liquid commodity futures traded on the Chigaco Board of Trade (CBOT), New York Mercantile Exchange (NYMEX) and Commodity Exchange Inc. (COMEX) which are all part of the Chicago Mercantile Exchange (CME) Group.

The first two commodities we analyze are CBOT Wheat Futures and CBOT Corn Futures. These are two of the most traded agricultural futures, and two of the most used agricultural products for human consumption. This highlights the importance of risk management in commodity markets. Russia is the worlds largest wheat exporter and Ukraine the fifth largest. Together, they account for nearly a third of global wheat supplies (Observatory of Economic Complexity, 2020). The war in Ukraine in 2022 lead to much uncertainty in wheat prices which translated to an increase in volatility in commodity (futures) markets. This makes estimating the risk in holding these futures even more important.

The second and third commodities investigated are NYMEX Crude Oil Futures and NYMEX Natural Gas Futures. These are the two most traded energy futures and thus, important for all parts of the world. Especially in the first half of 2022, the risk which comes with trading or holding these commodity futures became clear. The war in Ukraine resulted in import restrictions of these two physical commodities by many countries. This lead to a lot of uncertainty which translated to a lot of volatility in the energy commodity futures markets. The price of most energy commodity futures, including oil and gas, rose to exceptionally high levels but holding such a commodity in this period also comes with potential large daily losses. Thus estimating the VaR and more specifically the extreme level VaR for these commodity futures is very relevant.

The last commodity used in this research is COMEX Gold Futures. Gold is one of the most if not most important metal futures. Gold is historically seen as a safe haven asset and an inflation hedge. It is a safe haven asset because it is one of the world's first forms of currency due to its nice physical properties and finite supply. It is seen as an inflation hedge because it is a physical asset and cannot be printed like cash money. Gold also has a negative beta historically, meaning it is negatively correlated with the overall stock market (Ratner et al., 2014). Due too the characteristics of gold and its fundamental relationship to other financial markets, the risk of holding this commodity futures is interesting and very different from the other commodities.

For every series we will have roughly 10,000 observations which corresponds to 40 years of data (there are approximately 250 trading days in a year). Two commodity futures do not have data available from 1980 onwards. For NYMEX Crude Oil Futures data is only available from 1983 onwards and, for NYMEX Natural Gas Futures from 1990 onwards.

Table 1: *Descriptive statistics for 5 commodity futures daily returns*

Name	Obs	Mean	Std.Dev.	K	S	Min	Max	$Q_{0.05}$	$Q_{0.01}$	$Q_{0.005}$
Corn	10747	0.000	0.017	32.007	0.074	-0.241	0.300	-0.024	-0.044	-0.054
Wheat	10772	0.000	0.018	11.432	-0.143	-0.270	0.218	-0.026	-0.044	-0.055
Oil	9909	0.000	0.038	2118.557	-33.433	-2.543	0.377	-0.036	-0.067	-0.088
Gas	8174	0.001	0.035	11.765	0.858	-0.313	0.465	-0.050	-0.086	-0.102
Gold	10764	0.000	0.012	14.315	-0.106	-0.174	0.131	-0.018	-0.034	-0.043

Note. In the first column Corn is CBOT Corn Futures, Wheat is CBOT Wheat Futures, Oil is NYMEX Crude Oil Futures, Gas is NYMEX Natural Gas Futures and Gold is COMEX Gold Futures. The second column Obs is the total number of observations, Std.Dev. is the standard deviation, Min is the minimum value, Max is the Maximum value, $Q_{0.05}$ is the 5% empirical quantile, $Q_{0.01}$ is the 1% empirical quantile and $Q_{0.005}$ is the 0.5% empirical quantile.

Table 1 shows some descriptive statistics of the commodity futures returns. The total sum of all commodities is $n = 50,366$ observations. The mean is (nearly) zero for all commodities. The standard deviation is largest for the energy commodities, in addition to this, oil and gas have the lowest minimum return (-254.3% and -31.3% respectively) and the highest maximum return (37.7% and 46.5% respectively). The return of -254.3% of oil was on 04-20-2020, on this day the commodity futures obtained a close price of -37.63. This was the beginning of the corona pandemic and the demand for oil had dropped dramatically due to worldwide lockdowns. At the same time, the oil cartel OPEC+ could not agree on reducing their production immediately. Physical commodity traders ended with oil storage filling up. They tried to sell this oil but as their were no buyers, the price dropped below 0 (Myra P. Saefong, 2021). This is an example of extreme risk in commodity futures markets. We also observe that all commodities are heavy tailed (kurtosis > 3). Oil is extremely heavy tailed (kurtosis = 2118.557) and also negatively skewed. This indicates the oil futures has extreme values in the left tail. The standard deviation is lowest for gold as we expected. Due to golds stable properties described before it is less volatile than the other commodities. The empirical quantiles of the commodities differ largely, for $\tau = 0.05$ it ranges from -0.018 (-1.8%) for COMEX Gold Futures to -0.050 (-5.0%) for NYMEX Natural Gas Futures. For $\tau = 0.01$ the range is -0.034 (-3.4%) for COMEX Gold Futures to -0.086 (-8.6%) for NYMEX Natural Gas Futures. For $\tau = 0.05$ the range is -0.043 (-4.3%) for COMEX Gold Futures to -0.102 (-10.2%) for NYMEX Natural Gas Futures. Due to these large differences it is clear that a separate model is needed for every commodity futures.

4 Results

4.1 ERF

4.1.1 Hyperparameter tuning

Before evaluating the performance of ERF against other methods to estimate extreme quantiles, the proposed cross-validation scheme from section 2.2 is evaluated for both κ and λ. The intermediate quantile is fixed at $\tau_0 = 0.8$, the dimension of the predictor space is $p = 40$, and the number of training observations per simulation is $n = 2000$. Over 50 simulations, the performance is quantified as the square root of the mean integrated squared error (MISE)

between the estimated and true quantile function. The integrated squared error is given as follows,

$$ISE = \frac{1}{n'} \sum_{i=1}^{n'} \left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau) \right)^2. \tag{18}$$

The test data set $\{x_i\}_{i=1}^{n'}$, with $n' = 1000$, is generated with a Halton sequence on the cube $[-1, 1]^p$ (Halton, 1964). Figure 1 shows the root of MISE for various levels of κ on three levels of τ. The dashed line in this graph is the root of MISE for the cross-validated ERF. More precisely, for 50 simulations, ERF is applied with a fixed level of $\kappa \in 5, 10, 40, 100, 200, 300, 350, 400, 450, 500$ and ERF is applied with cross-validation using a grid of $\kappa \in \{5, 10, 40, 100, 200, 300, 350, 400, 450, 500\}$. For λ, a fixed level of 0 is used.

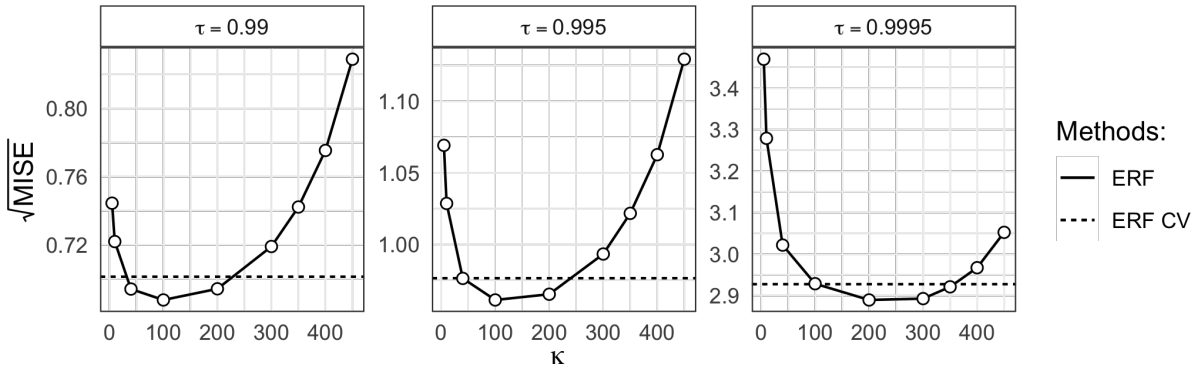

Figure 1: *The performance of ERF, measured as the root of the MISE, over 50 simulations for fixed values of κ (ERF), and a cross-validated κ (ERF)*

.

We notice that ERF's cross-validated performance is near the minimum square root MISE, implying that the proposed cross-validation approach is effective for κ. Next, we will evaluate the cross-validation scheme for λ. The same approach as before is used. We fix $\kappa = 100$ because in the previous cross-validation, this value was most often the minimizer of $CV(\alpha_j)$ in equation (10).

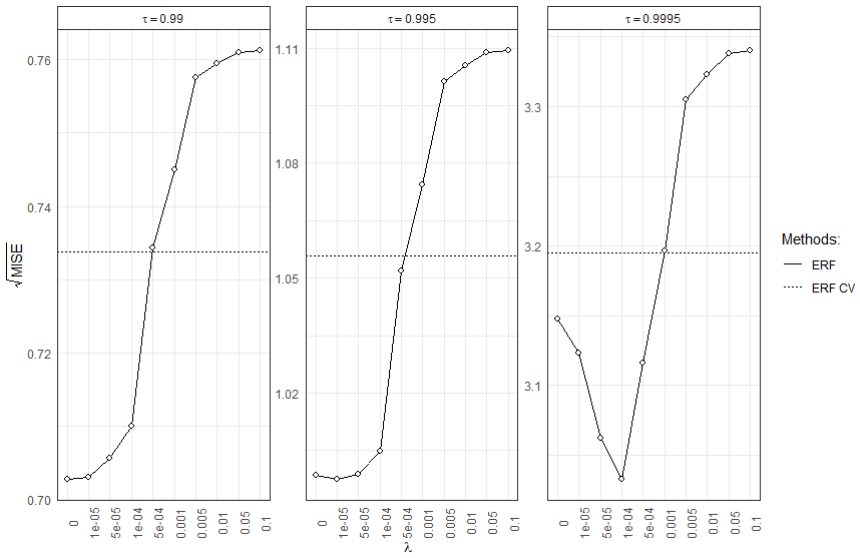

Figure 2: *The performance of ERF, measured as the root of the MISE, over 100 simulations for fixed values of λ (ERF), and for a cross-validated λ (ERF CV)*

Figure 2 shows the performance over 100 simulations for a fixed value of $\lambda = 0, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1$ and for a cross-validated value of λ over the following grid $\{0, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1\}$. The dashed line in figure 2 is roughly in between the fixed values of λ. This indicates that the cross-validation scheme does not work well for all three values of τ. Throughout the rest of this analysis we will thus, not use this cross-validation scheme for λ. To still stabilize the variance of the shape parameter we set $\lambda = 0.01$.

4.1.2 Simulation

For the first experiment, ERF is evaluated over 50 simulations using the settings explained in 2.3. For each simulation, the ERF model is first cross-validated over the grid $\kappa \in \{10, 40, 100\}$ and $\lambda = 0.01$ using the training set of $n = 2000$ observations. Then, the optimal hyperparameters are used to fit ERF on the same training set. Finally, estimations are made using the test set of $n' = 1000$ observations.

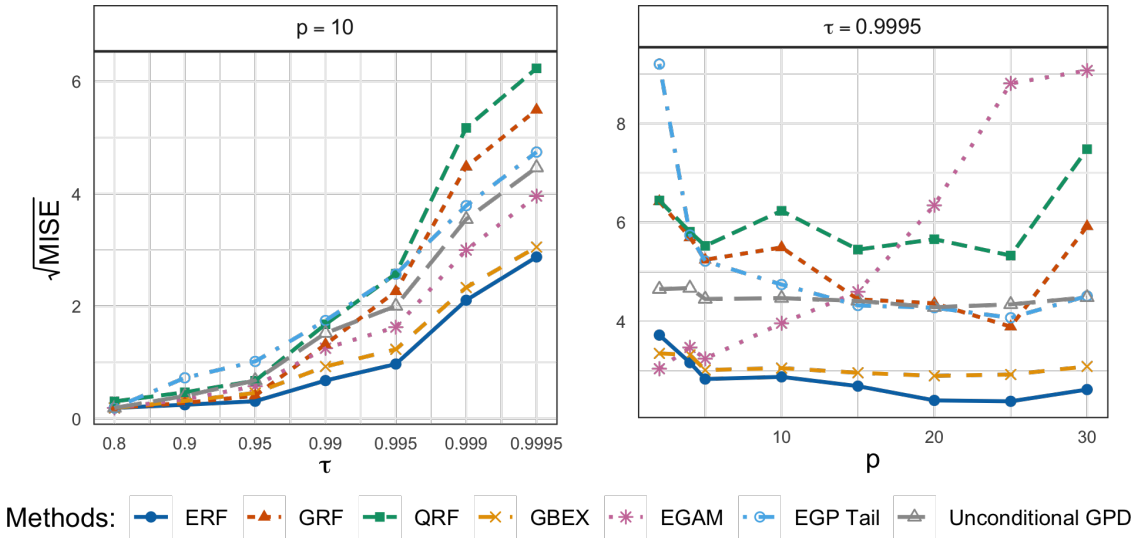

Figure 3: *The performance of ERF compared to 6 other models for different levels τ with fixed predictor dimension $p = 10$ (left panel), and for different levels p with fixed quantile level $\tau = 0.9995$ (right panel) over 50 simulations.*

Figure 3 shows the results of the first simulation setup. In the left panel, the number of predictors is fixed at $p = 10$, and the square root MISE is investigated for different values of τ. All methods have similar performance at less extreme quantile levels. The extrapolation-based methods (ERF, GBEX and EGAM) naturally coincide at these levels because they use the same estimator (GRF) to estimate the intermediate quantile τ. The methods that do not extrapolate to the tail (GRF, QRF, EGP tail and Unconditional GPD) do not perform well for high quantile levels as expected. ERF and GBEX outperform all other methods in the higher quantile levels, this is due too the fact that they suffer less from high quantile levels.

The right panel of figure 3 shows the results of the first simulation setup where different values of p are investigated and the quantile level is fixed at $\tau = 0.9995$. EGAM performs relatively well when $p \leq 10$ but fails entirely when the dimension of the predictor space increases. This method is thus not flexible enough to cope with many noise variables. The two other forest-based methods (GRF and QRF) do not seem to suffer much from a large number of predictors, except when $p > 25$. Again, ERF and GBEX outperform all other methods, they do not seem to suffer at all from the higher dimensional predictor space.

For the second experiment, the data is simulated using different values of the shape parameter $\xi = 0$ (Gaussian distribution) or $\xi = 1/4, 1/3$ (Students's t distribution) as explained in section 2.3. In this setup the quantile level is set at $\tau = 0.9995$ and the dimension of the predictor space at $p = 40$.

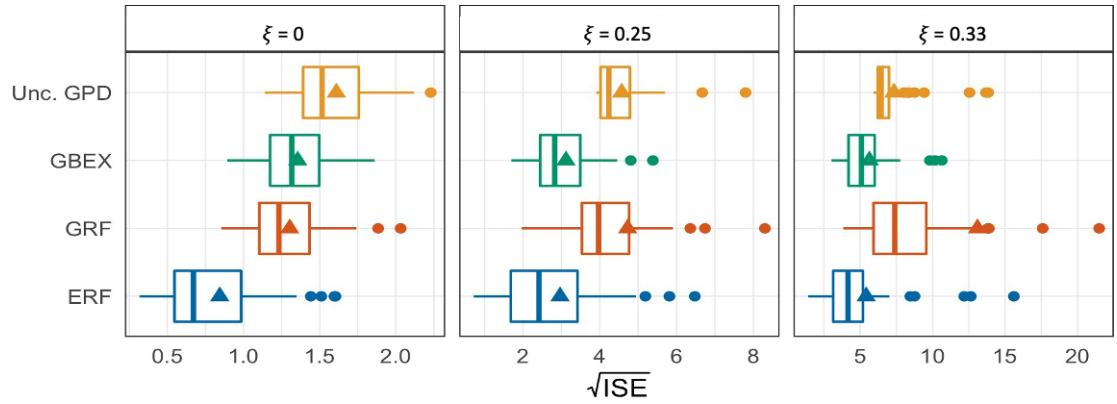

Figure 4: *Boxplots of the square root ISE of 4 different methods over 50 simulations for three values of ξ, triangles correspond to average values.*

Figure 4 shows the results of the second simulation setup. The results are displayed as boxplots where the triangles correspond to average values, the vertical line is the median, and the square box is the second to third quantile of values of the square root ISE. Large outliers of GRF are removed to make the boxplots easier observable. When the noise becomes more heavy tailed (e.g. a larger value of ξ), the performance of all models deteriorates because predicting becomes more difficult. For GRF, the performance deteriorates more in comparison to the rest of the models. Especially, the mean of the square root ISE becomes higher and has more outliers with large square root ISE. This is less the case for ERF, GBEX and the unconditional GPD. ERF has the best performance, even outperforming GBEX in the three scenarios, indicating the importance of tail extrapolation.

4.1.3 U.S. wage structure

To estimate extreme quantiles of the U.S. wage, we compare ERF to GRF, GBEX and the Unconditional GPD. To tweak the minimum node size κ, the 5-fold cross-validation scheme explained in 2.2 is applied three times on the grid $\kappa \in \{5, 40, 100\}$. The penalty variable is set to $\lambda = 0.01$ because the cross-validation scheme did not seem to work well for λ as shown in section 4.1.1. We do not set it to 0 because we still want to reduce the variance of the shape parameter.

The first half $(32,511$ observations) of the total data set is used for an exploratory data analysis. The other half of the (32,512 observations) is used to evaluate the performance of ERF. Within the first half, a random set of 10% is used to first tune the hyperparameters and then fit the ERF object with optimal parameters. The other 90% is then used to estimate GPD parameters and predict quantiles.

For the second half, a cross-validation scheme is applied to evaluate the performance. This is a different scheme than the cross-validation scheme to tune the hyperparameters. The data is split into 10 random folds. We take one fold as training data which is first used for hyper-parameter tuning, and then used for fitting the ERF object with the optimal parameters. The other 9 folds are used for testing. Much data is left for testing, so there are enough observations

to evaluate the performance.

To evaluate the performance of the different models, the following metric introduced by Wang & Li (2013) is used,

$$R_n(\hat{Q}_\cdot(\tau)) := \frac{\sum_{i=1}^{n} 1\{Y_i < \hat{Q}_{X_i}(\tau)\} - n\tau}{\sqrt{n\tau(1-\tau)}} \tag{19}$$

where $\hat{Q}_\cdot(\tau)$ is the conditional τ-th quantile estimated on the training data set and n is the number of test observations. Equation (19) compares the theoretical quantile level with the normalized estimated proportion of observations with $Y_i < \hat{Q}_{X_i}$. This metric is asymptotically standard normal. We consider two cases. In the first we follow the same setup as Gnecco et al. (2022) where we add an extra 10 randomly generated noise variables to increase the predictor space. We also consider a case where there are no extra generated noise variables.

Figure 5 shows the estimated GPD parameters as a function of education using the first half of the data set. We do not observe any clear difference between the black and white subgroups for both the estimated scale $\hat{\sigma}(x)$ and shape $\hat{\xi}(x)$ parameters. The scale parameter is increasing with respect to years of education as expected. Especially, from year 12 to 16 the scale parameter increases sharply and after year 16 it does not increase anymore. In the U.S., 12 to 16 years of education correspond to the period after high school until the end of the undergraduate study. This indicates the following: for people who finished their undergraduate study, the quantile of wage is higher than for people who did not, according to the ERF model. Furthermore, the shape parameter is decreasing with respect to years of education but this difference seems very small, roughly between 0.23 and 0.22, indicating heavy tails.

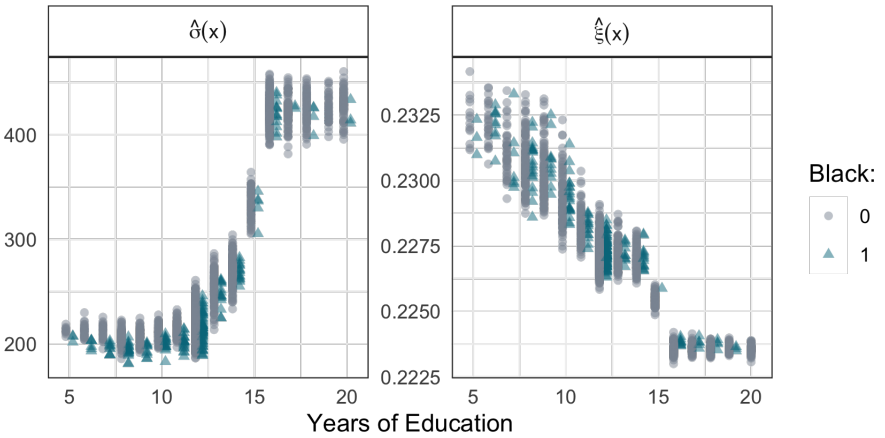

Figure 5: *Estimated GPD parameters as a function of years of education for black (triangles) and white (circles) subgroups with 10 extra generated noise variables.*

When considering the case without extra generated noise variables in figure 12 in the appendix, we see roughly the same pattern for the scale parameter, only more concentrated and with larger vales of $\hat{\gamma}(x)$. However, the scale parameter is larger for whites than for blacks when years of education is 12 or higher. This indicates that according to the ERF model, high-educated whites have a higher quantile wage level than high-educated blacks. For the shape parameter, the values are roughly twice as large but the difference in values with respect to years of education is still

very small. Figure 13 in the appendix shows the estimated GPD parameters with respect to age. The parameters do not seem to depend on age but a very clear relation can be seen with respect to years of education. People with more than 15 years of education have a higher scale parameter, indicating a higher quantile wage level. We also observe a slightly higher shape parameter for people with less than 15 years of education, possibly indicating more outliers in the tail. The same relations hold for the case without extra generated predictors which can be seen in figure 14.

Figure 6 shows the predicted quantiles for $\tau = 0.9$ and $\tau = 0.995$ with respect to years of education using the first half of the data set. From figure 5 we see that the value of the shape parameter varies with respect to years of education thus, the predicted quantile should also vary in a similar manner. For $\tau = 0.9$ all four methods seem to model this well. For $\tau = 0.995$ only the extrapolation methods (ERF and GBEX) model this variability well. For the case without extra generated noise variables in figure 15 the overall pattern is similar, only the curves are all flatter. This seems counter-intuitive because the curve in the left panel in figure 12 is not flatter. We do observe higher estimated quantiles for quantile level $\tau = 0.9$ for all 4 methods. This indicates that the quantiles might be overestimated in the case with 10 extra generated noise variables.

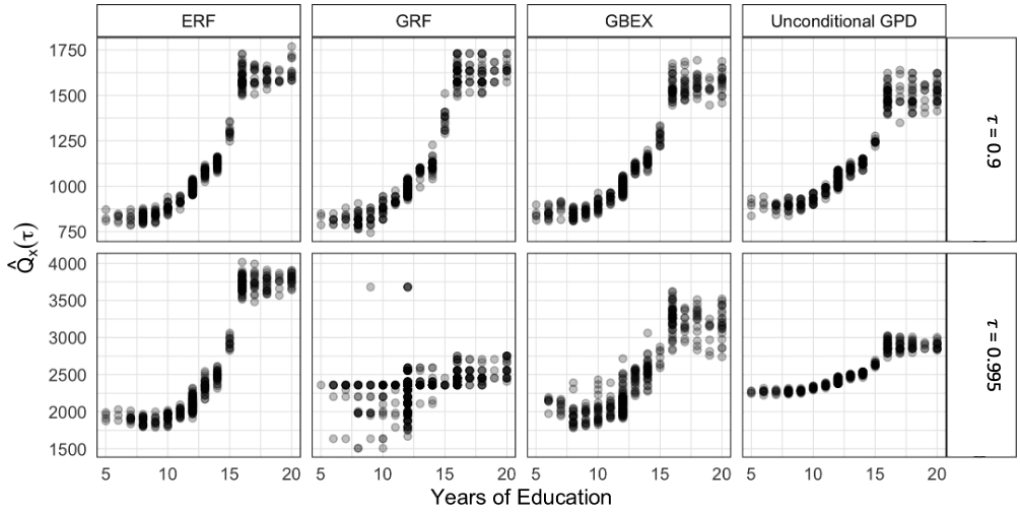

Figure 6: *Predicted quantiles for* $\tau = 0.9, 0.995$ *for ERF, GRF, GBEX and Unconditional GPD with 10 extra generated noise variables*

Figure 16 in the appendix shows boxplots of the loss metric in equation (19) for ERF, GRF, GBEX and the Unconditional GPD. All quantiles above 6000 predicted by GRF are removed to make the graph more clear. The shaded area shows the 95% interval of the absolute value of the standard normal distribution. Because the metric in equation (19) is asymptotically standard normal, the shaded area also corresponds to the 95% confidence level of this metric with true quantile value $Q.(\tau)$. GRF performs well for smaller quantile level $\tau = 0.9$ but not wel for higher quantile levels which is due to the fact that this method does not extrapolate the tail. The unconditional GPD method does not perform well for all quantile levels and especially not for higher quantile levels. Again, ERF and GBEX have the best performance.

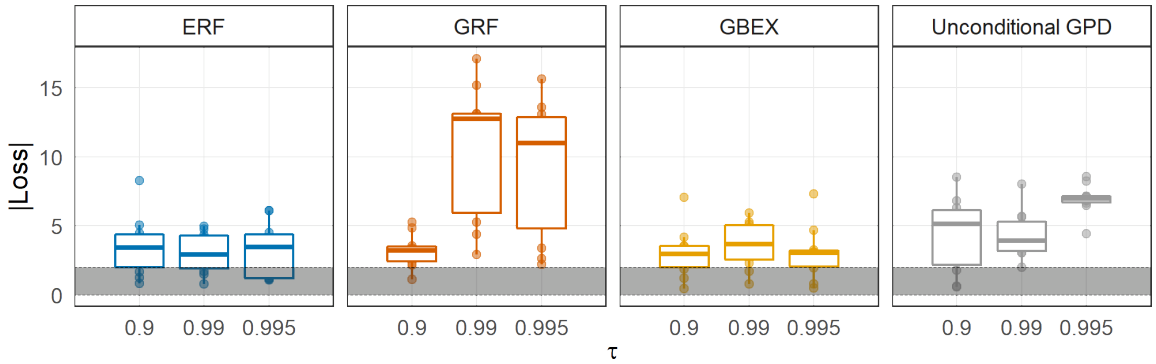

Figure 7: *The absolute value of the loss metric in equation* (19) *for ERF, GRF, GBEX and the Unconditional GPD fitted to the original response of U.S. wage data. The shaded region shows the 95% confidence interval of a standard normal distribution's absolute value. With 10 extra generated noise variables.*

Figure 16 shows boxplots of the loss metric in equation (19) for ERF, GRF, GBEX and the Unconditional GPD with no extra generated noise variables. The overall results do not seem very different. The value of the loss metric increases for GRF for $\tau = 0.99, 0.95$, while the variability in the loss metric is less.

4.2 Autoregressive ERF for VaR

To evaluate the performance of the Autoregressive ERF for VaR, the results for the CBOT Corn Futures returns, are analyzed in detail before we consider all 5 commodities together. Using the first data half, we find the optimal hyperparameters. Then, the results for the Christoffersen tests are shown using the second data half.

4.2.1 Hyperparameter tuning

In this section we tune the hyperparameter κ and the optimal number of lags p of Y_{t-i} to include for every commodity futures. We do this using the first half of data.

Table 2: *Values of $F(\alpha_j)$ for 3 levels of p and three levels of κ with $\tau = 0.99$ for the estimation of VaR of 5 commodity futures returns using the first half of observations for every commodity*

Commodity futures	$\kappa \setminus p$	1	2	3
CBOT Corn Futures	10	**-2537.412**	-2356.524	-2436.888
	40	-2529.564	-2389.103	-2458.805
	100	-2507.462	-2505.804	-2440.661
CBOT Wheat Futures	10	-2614.62	-2676.39	**-2704.897**
	40	-2642.286	-2651.76	-2608.778
	100	-2581.184	-2604.193	-2668.726
NYMEX Crude Oil Futures	10	-1642.229	-1623.176	-1664.512
	40	**-1682.14**	-1651.586	-1675.117
	100	-1636.407	-1660.658	-1665.492
NYMEX Natural Gas Futures	10	-1421.578	**-1691.432**	-1463.325
	40	-1101.285	-1003.303	-1029.315
	100	-1312.332	-1112.967	-1601.101
COMEX Gold Futures	10	-2165.380	-2146.130	**-2200.372**
	40	-2188.332	-2188.198	-2115.243
	100	-2186.461	-2199.721	-2202.264

Table 2 shows the values of the tuning metric $F(\alpha_j)$ from equation (15) for the estimated quantiles of all 5 commodity futures returns on the grid $\kappa \in \{10, 40, 100\}$ and $p \in \{1, 2, 3\}$. The first half of the entire data set is used to calculate these values. The values of κ and p which minimize $F(\alpha_j)$ are used for testing Autoregressive ERF on the second half of data. These values of $F(\alpha_j)$ are shown in **bold**.

4.2.2 Value at Risk

Figure 8 shows the historical returns of corn futures for the period 2010-02-01 to 2022-05-02 (3162 observations) with estimated VaR using the Autoregressive ERF method for three quantile levels $\tau = 0.005, 0.01, 0.05$. To obtain this figure, the estimated right-tail quantiles of Y_t (inverse return) are multiplied by -1. In the discussion of the results, when we say quantile levels $\tau = 0.005, 0.01, 0.05$ we refer to the quantile levels of the actual return r_t. When observing figure 8 it seems that the model is not very flexible in predicting the quantiles. The range of the estimations seems not to change much over time. This could indicate that the model does not take past information enough into account. Figure 9 zooms in on the last 200 observations of figure 8. When looking at the three most negative returns (below 0.05) we see that for quantile levels $\tau = 0.01, 0.05$ the model predicts a lower quantile the day after this negative return and then bounces back the next day. For $\tau = 0.01$ the estimation range is roughly -0.035 to -0.06 and for $\tau = 0.05$ it is -0.06 to -0.08. For quantile level $\tau = 0.001$ the estimations seem arbitrary in the range of -0.07 to -0.09.

Figure 10 and figure 11 show the quantile estimates using the GARCH(1,1) and CAViaR model respectively. These estimations are a lot more flexible. The large volatility spike in 2021 is captured well and the low volatility period of 2015 to 2020 has lower estimated quantiles for both models.

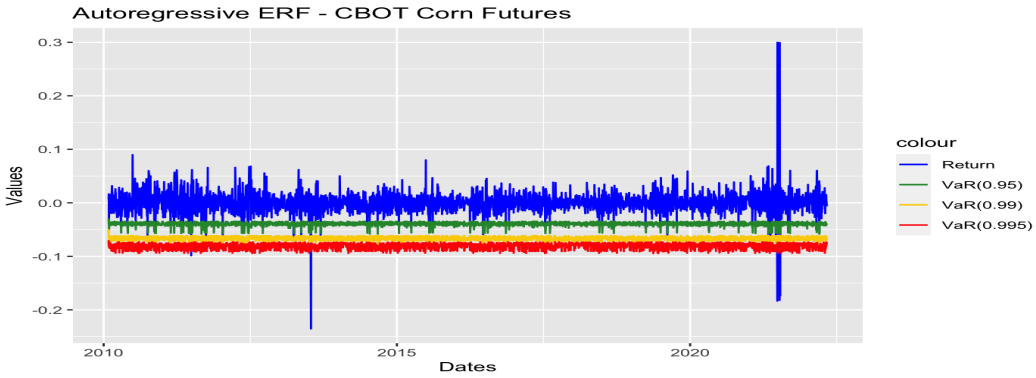

Figure 8: *Historical daily returns of CBOT Corn Futures for the period 2010-02-01 to 2022-05-02 (3162 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.01$.

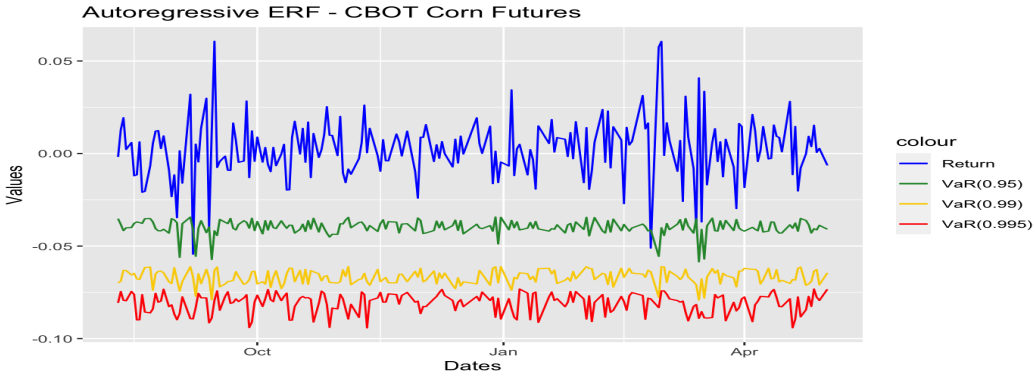

Figure 9: *Historical daily returns of CBOT Corn Futures for the period 2021-08-10 to 2022-04-29 (200 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.01$.

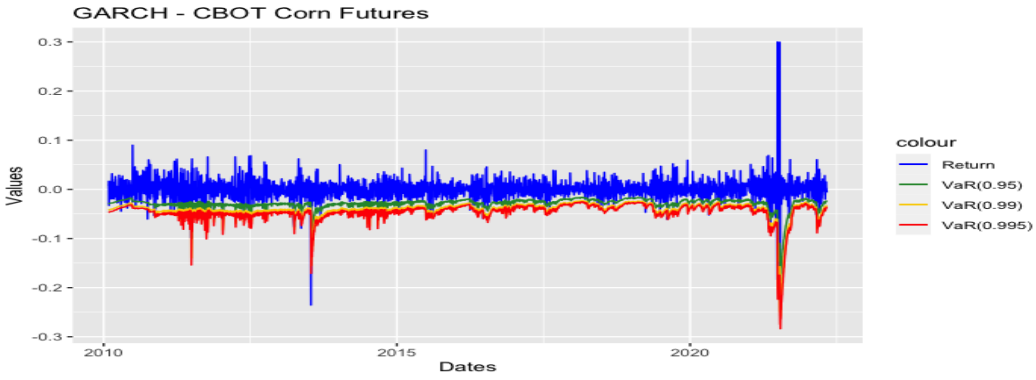

Figure 10: *Historical daily returns of CBOT Corn Futures for the period 2010-02-01 to 2022-05-02 (3162 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

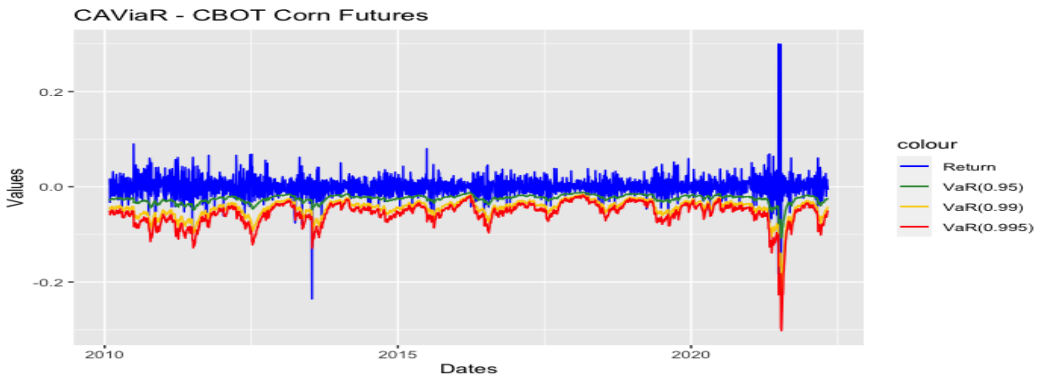

Figure 11: *Historical daily returns of CBOT Corn Futures for the period 2010-02-01 to 2022-05-02 (3162 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

Table 3: *Christoffersen test results for unconditional and conditional coverage of the VaR of CBOT Corn Futures returns using the Autoregressive ERF, GARCH and CAViaR method for the period 2010-02-01 to 2022-05-02 (3162 observations).*

Model	τ	EE	T_1	LR_{UC}	P_{UC}	LR_{CC}	P_{CC}
Autoregressive ERF	0.05	158	53	97.984	0.000	97.998	0.000
	0.01	31	7	28.323	0.000	28.354	0.000
	0.005	15	45	10.145	0.001	10.161	0.006
GARCH	0.05	158	110	17.162	0.000	18.301	0.000
	0.01	31	41	2.570	**0.109**	2.908	**0.234**
	0.005	15	30	10.117	0.001	11.234	0.004
CAViaR	0.05	158	159	0.005	**0.942**	1.958	**0.376**
	0.01	31	30	0.085	**0.770**	1.202	**0.548**
	0.005	15	15	0.042	**0.837**	3.588	**0.166**

Table 3 shows the results of the tests for unconditional and conditional coverage for the estimated VaR of corn futures. EE are the expected number of exceedances based on the quantile level τ, T_1 is the actual number of exceedances, LR_{UC} is the likelihood ratio test statistic for unconditional coverage, P_{UC} is the p-value for unconditional coverage, LR_{CC} is the likelihood ratio statistic for conditional coverage and P_{CC} is the p-value for conditional coverage. The p-values which exceed the level of statistical significance $\alpha = 0.05$ are shown in **bold**. A p-value exceeding α implies that the null hypothesis of correct unconditional coverage or correct conditional coverage is not rejected. For the VaR estimates of corn futures using the Autoregressive ERF model, these null hypotheses are rejected for all quantile levels. For the GARCH(1,1) model, the estimates have correct conditional coverage for $\tau = 0.01$. For CAViaR, all estimates have correct conditional coverage.

Figure 17, 18 and 19 in the appendix show plots of the estimated VaR for Wheat futures using the Autoregressive ERF, GARCH and CAViaR methods respectively. The Autoregressive ERF estimates look a bit more variable for Wheat than for Corn. The volatility spike in 2022 is captured lightly and the estimates for the more volatile first 2 years obtain lower values (more negative). The estimations however, still bounce back and forth and are all around a certain

range. When observing the VaR estimates using GARCH(1,1) and CAViaR, we see that two models clearly capture the volatility spike in 2022. Apart from this spike, the range of the estimates of the three models is close too each other. Only the estimates are smoother for CAViaR and even more smooth for GARCH(1,1). Table 4 shows the Christoffersen test results for wheat futures. The CAViaR model is the only model for which the estimates have correct conditional coverage for all quantile levels. The estimates of the GARCH(1,1) model have correct conditional coverage for quantile levels $\tau = 0.99, 0.995$.

Figure 20, 21 and 22 in the appendix show plots of the estimated VaR for Gas futures returns using the Autoregressive ERF, GARCH and CAViaR methods respectively. In all three figures, the large negative return on 04-20-2020 stands out. The GARCH(1,1) and CAViaR model capture this volatility spike and the Autoregressive ERF model does not. This spike makes the rest of the figure unclear so we examine the first 2000 observations in figure 23 for Autoregressive ERF. We observe that the estimates are more variable for oil than for corn and wheat. In the more volatile period of 2015 to 2016 the estimates obtain lower values and in the less volatile period of 2013 to 2014 the estimates obtain values closer to 0. The estimates do however, bounce back after obtaining a low value as was the case for corn and wheat. Figures 24 and 25 show the last 400 observations of estimated VaR for oil futures returns using GARCH(1,1) and CAViaR respectively. For the extreme quantile level $\tau = 0.995$ we observe much lower estimates than for GARCH(1,1). The estimates for GARCH(1,1) are the also the most smooth. Table 5 shows the Christoffersen tests for the oil futures. The tests indicate correct conditional coverage for CAViaR for all quantile levels. The estimates of GARCH(1,1) have correct conditional coverage for quantile level $\tau = 0.95$. For Autoregressive ERF we have correct conditional coverage for quantile level $\tau = 0.95, 0.99$ and correct unconditional coverage for $\tau = 0.995$.

Figure 26, 27 and 28 in the appendix show plots of the estimated VaR for Gas futures returns using the Autoregressive ERF, GARCH and CAViaR methods respectively. We observe a few volatility spikes which are captured well by the GARCH(1,1) and CAViaR models. These spikes are not captured by the Autoregressive ERF model. The estimates for Autoregressive ERF again are not very variable. Table 6 shows the Christoffersen test results for gas futures. The estimates of the CAViaR model are the only estimates with correct conditional coverage for all quantile levels. The estimates of the GARCH(1,1) model have correct conditional coverage for quantile levels $\tau = 0.95, 0.99$ only.

The last commodity futures which we investigate is gold. Figure 29, 30 and 31 in the appendix show plots of the estimated VaR using the Autoregressive ERF, GARCH and CAViaR methods respectively. We observe the same patterns as the other commodity futures. The estimates for Autoregressive ERF are in a small range and bounce back and forth. The estimates for GARCH(1,1) and CAViaR are more variable and capture volatility spikes. Table 7 show the estimates for CAViaR have correct conditional coverage for all three quantile levels and the estimates for the GARCH(1,1) model have conditional coverage for $\tau = 0.95$.

5 Discussion and conclusion

Estimating VaR is of great importance for asset managers. There are many existing methods to do this and some existing methods to estimate extreme level VaR. However, not much research has been conducted to estimate extreme level VaR in commodity futures markets. The recent extreme volatility in commodity markets makes this even more important. During the corona pandemic, daily returns below -100% occurred for oil futures. In 2021, global supply chain issues and rising inflation increased volatility in commodity markets as well. In 2022, The war in Ukraine amplified these problems leading to even more volatility. These examples stress the importance of finding a suitable method to estimate extreme level Value at Risk for commodity futures.

This paper estimated extreme level quantiles in multiple scenario's using the ERF method. First, the hyperparameter tuning scheme for this method was tested in a simulation study. Based on these simulation results, we conclude that the scheme is efficient for the minimum node size κ but not for the penalty value λ. Next, the ERF model was tested in a more extensive simulation study together with six other models. Two simulation setups were considered. In the first setup we found that the methods are comparable for intermediate quantile levels. ERF however outperforms the other methods for higher quantile levels and is competitive to GBEX. In the first setup we also found that ERF outperforms the other methods for different dimensions of the predictor space and is again competitive to GBEX. In the second setup we simulated data using different shape parameters of the noise distribution. In the second simulation setup we found that the ERF method is most flexible to a varying tail heaviness in the noise distribution.

Next, a real-life data set was used to estimate quantiles of U.S. wages conditional on some predictors using the ERF model. A case was discussed with ten extra generated noise variables to increase the dimension of the predictor space and a case was discussed without these extra noise variables. There was not a large difference in the results of these two cases and ERF and GBEX outperformed GRF and the Unconditional GPD for higher quantile levels. For the moderately high quantile level 0.9, the performance of ERF and GBEX was only moderately better. The first two sections have shown that ERF is a great method in tackling the difficulties of estimating extreme level quantiles in a high dimensional predictor space.

In another real-life experiment the ERF method was tested on a financial data set consisting of 5 commodity futures returns. To do this, the Autoregressive ERF model was introduced, which predicts quantiles of the inverse returns using lagged inverse returns and lagged predicted quantiles. The predicted quantile of inverse returns was then multiplied by -1 to obtain VaR estimates of the actual returns. Using Christoffersen tests, the performance of Autoregressive ERF and the performance of 2 other models, GARCH(1,1) and CAViaR, was evaluated. The estimates of the CAViaR model have correct conditional coverage for all quantile levels, for all commodities. The estimates of the GARCH model have correct conditional coverage in half of the cases. The estimates for Autoregressive ERF only have correct conditional coverage for the

two lower quantile levels of the oil futures returns.

This paper tried to answer whether the ERF method of Gnecco et al. (2022) is an acceptable method for estimating daily extreme Value at Risk of commodity futures. We labelled a method as acceptable if the estimates had correct conditional coverage which implies the fraction of VaR violations is approximately equal to the nominal probability, while also spread out over the sample. Based on this research, we conclude that the Autoregressive ERF method is not an acceptable method for estimating extreme levels of VaR for commodity futures returns. The only case where the estimates did have correct conditional coverage was for the moderate quantile levels 0.05 and 0.01 for oil future returns.

The predictive power of Autoregressive ERF however, is not entirely negligible. Due to this, it might be interesting to investigate Autoregressive ERF further. The Autoregressive ERF function was created using the ERF package in R. This package makes use of a GRF to predict an intermediate quantile and similarity weights. Using the ERF package, it is only possible to adjust the minimum node size κ and the penalty value λ. For future research, it could be interesting to investigate Autoregressive ERF by tuning other hyperparameters which belong to GRF. In addition to this, other adjustments of the method can be researched. The number of lagged predictors or number of lagged estimated quantiles can be increased or even other predictors such as the absolute lagged returns can be used. Another recommendation for future research is finding a good comparison metric. The conditional coverage and independence tests only test the correct conditional coverage of VaR models but do not make a comparison between different model estimates. The Dynamic Quantile test of Engle & Manganelli (2004) or the metric of Wang & Li (2013) might be interesting to use. These recommendations are all beyond the scope of this paper.

Appendices

A U.S. Wage Analysis

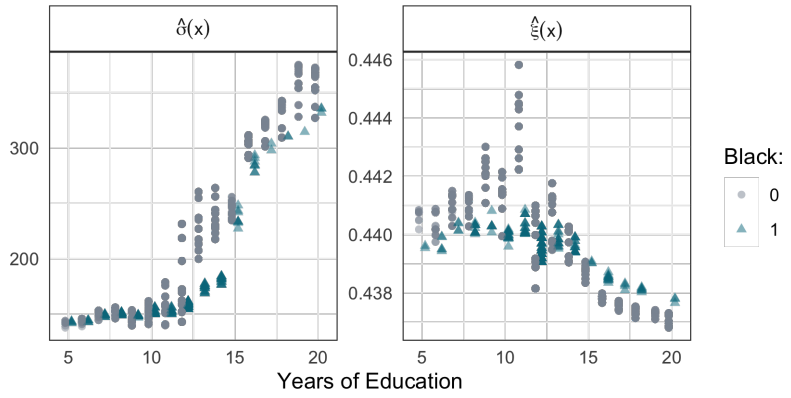

Figure 12: *Estimated GPD parameters as a function of years of education for black (triangles) and white (circles) subgroups without extra generated noise variables.*

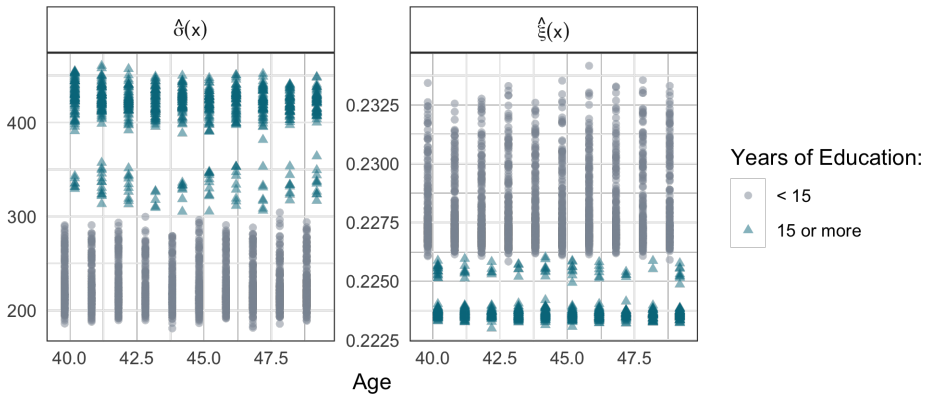

Figure 13: *Estimated GPD parameters as a function of age for black (triangles) and white (circles) subgroups with 10 extra generated noise variables.*

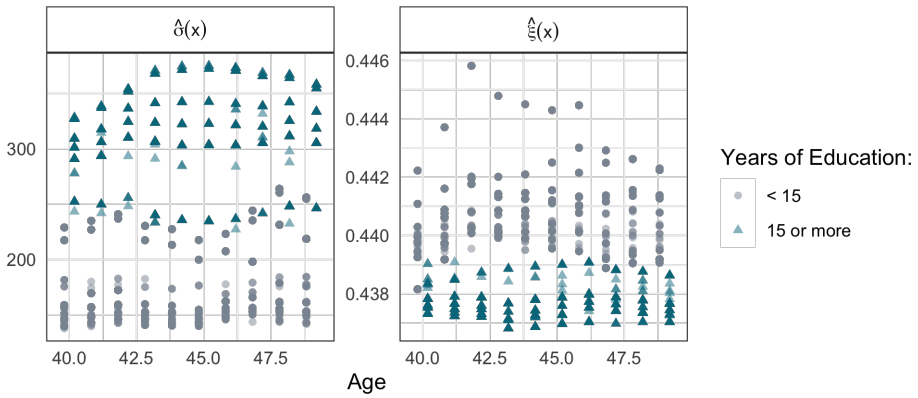

Figure 14: *Estimated GPD parameters as a function of age for black (triangles) and white (circles) subgroups without extra generated noise variables.*

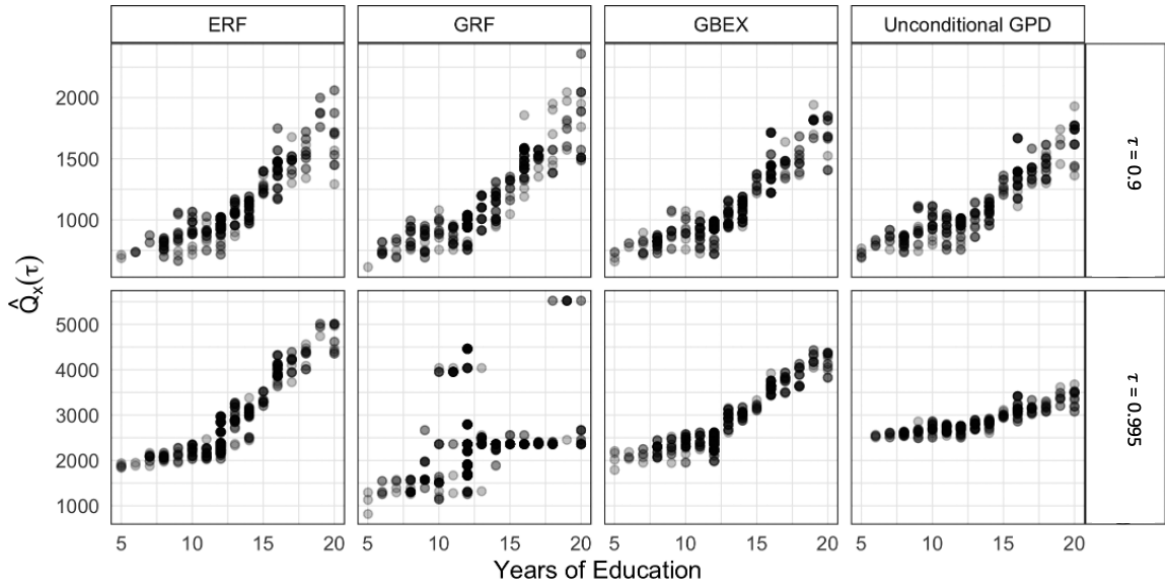

Figure 15: *Predicted quantiles for* $\tau = 0.9, 0.995$ *for ERF, GRF, GBEX and Unconditional GPD without extra generated noise variables*

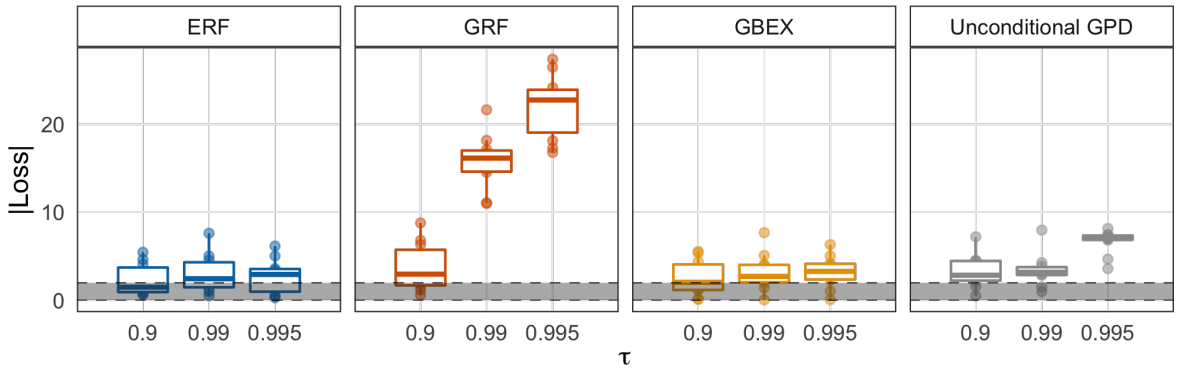

Figure 16: *The absolute value of the loss metric in equation* (19) *for ERF, GRF, GBEX and Unconditional GPD fitted to the original response of U.S. wage data. The gray region shows the 95% confidence interval of a standard normal distribution's absolute value. Without extra generated noise variables*

B Value at Risk plots and Christoffersen test results

B.1 CBOT Wheat Futures

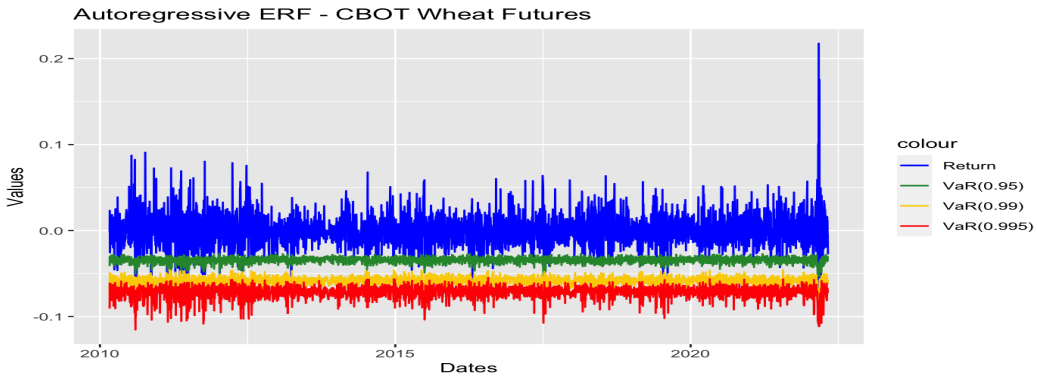

Figure 17: *Historical daily returns of CBOT Wheat Futures for the period 2010-03-02 to 2022-04-29 (3168 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

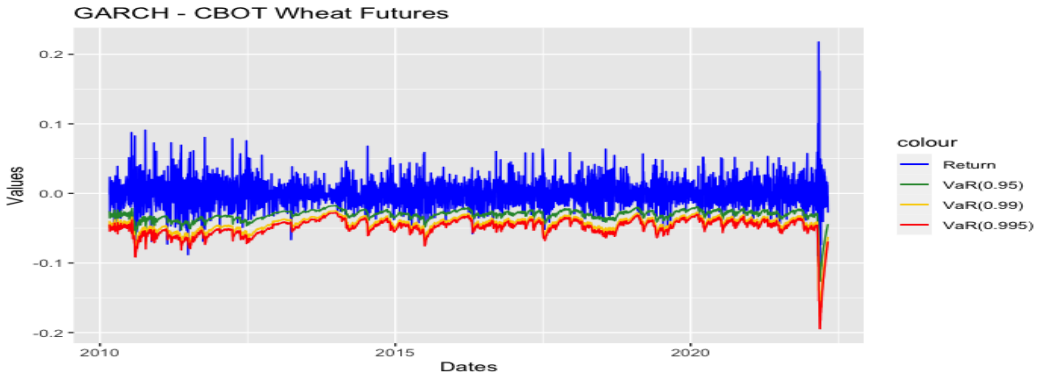

Figure 18: *Historical daily returns of CBOT Wheat Futures for the period 2010-03-02 to 2022-04-29 (3168 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

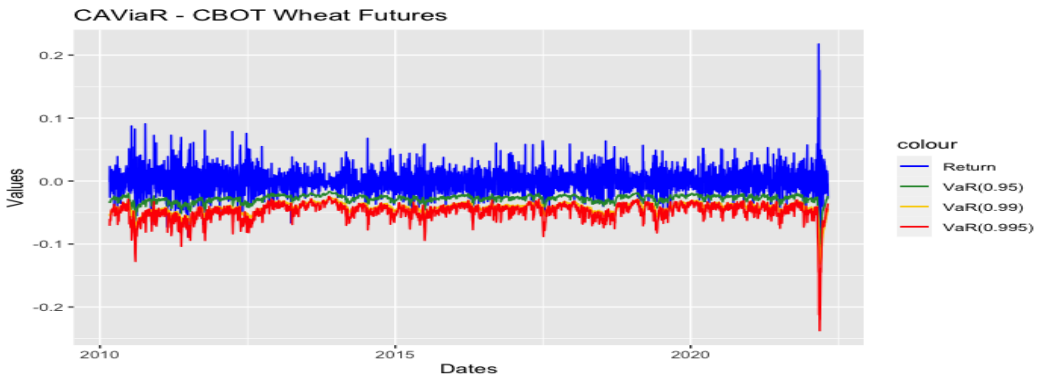

Figure 19: *Historical daily returns of CBOT Wheat Futures for the period 2010-03-02 to 2022-04-29 (3168 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

Table 4: *Christoffersen test results for unconditional and conditional coverage of the VaR of CBOT Wheat Futures returns using Autoregressive ERF, GARCH and CAViaR methods for the period 2010-03-02 to 2022-04-29 (3168 observations).*

Model	τ	EE	T_1	LR_{UC}	P_{UC}	LR_{CC}	P_{CC}
Autoregressive ERF	0.95	158	88	38.984	0.000	42.949	0.000
	0.99	31	11	18.225	0.000	18.301	0.000
	0.995	15	4	12.714	0.000	12.725	0.002
GARCH	0.95	158	118	11.852	0.001	12.423	0.002
	0.99	31	23	2.655	**0.103**	2.992	**0.224**
	0.995	15	15	0.046	**0.831**	0.188	**0.910**
CAViaR	0.95	158	1158	0.001	**0.974**	0.005	**0.998**
	0.99	31	32	0.003	**0.955**	0.656	**0.720**
	0.995	15	17	0.083	**0.773**	0.267	**0.875**

B.2 NYMEX Crude Oil Futures

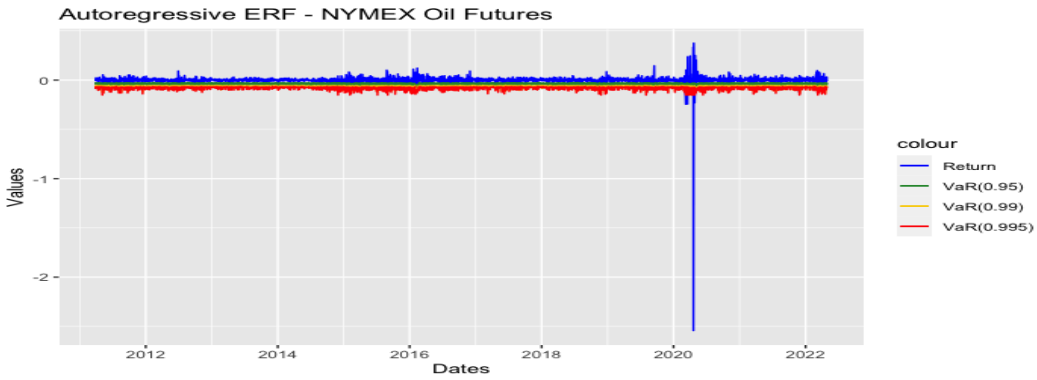

Figure 20: *Historical daily returns of NYMEX Crude Oil Futures for the period 2011-03-29 to 2022-05-01 (2910 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

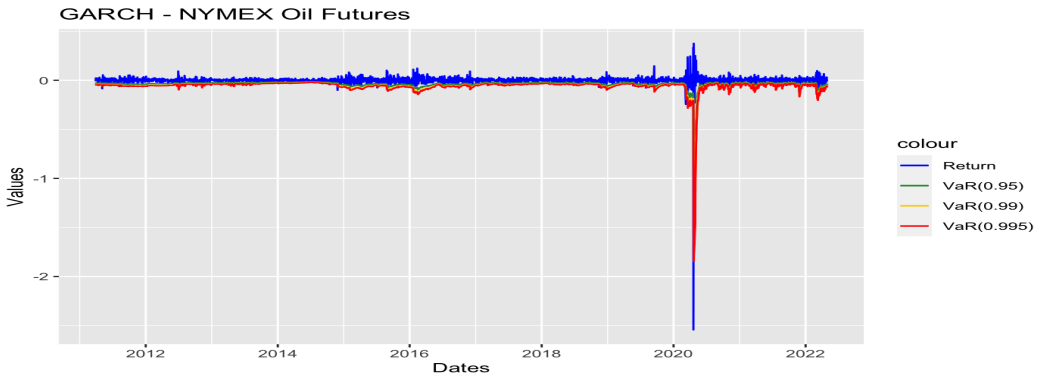

Figure 21: *Historical daily returns of NYMEX Crude Oil Futures for the period 2011-03-29 to 2022-05-01 (2910 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

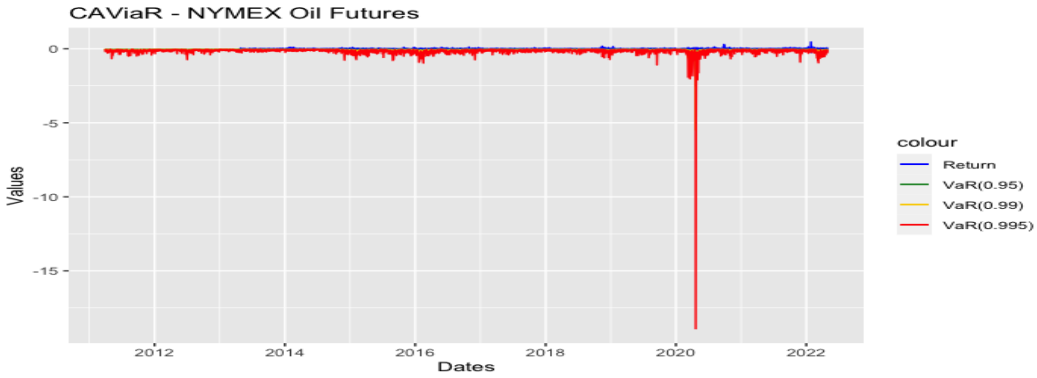

Figure 22: *Historical daily returns of NYMEX Crude Oil Futures for the period 2011-03-29 to 2022-05-01 (2910 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

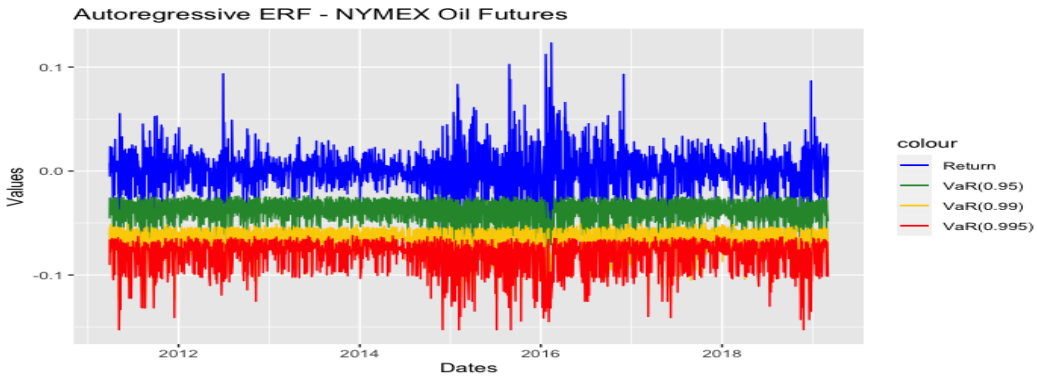

Figure 23: *Historical daily returns of NYMEX Crude Oil Futures for the period 2011-03-29 to 2019-03-04 (2000 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

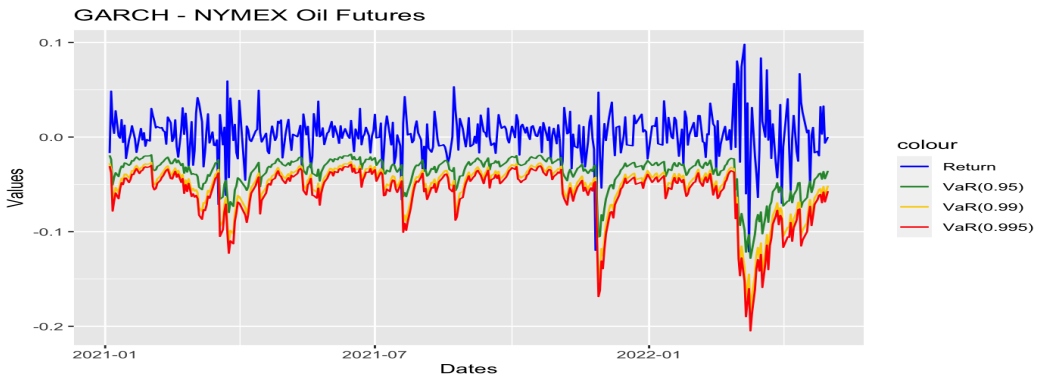

Figure 24: *Historical daily returns of NYMEX Crude Oil Futures for the period 2021-01-04 to 2022-05-01 (400 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

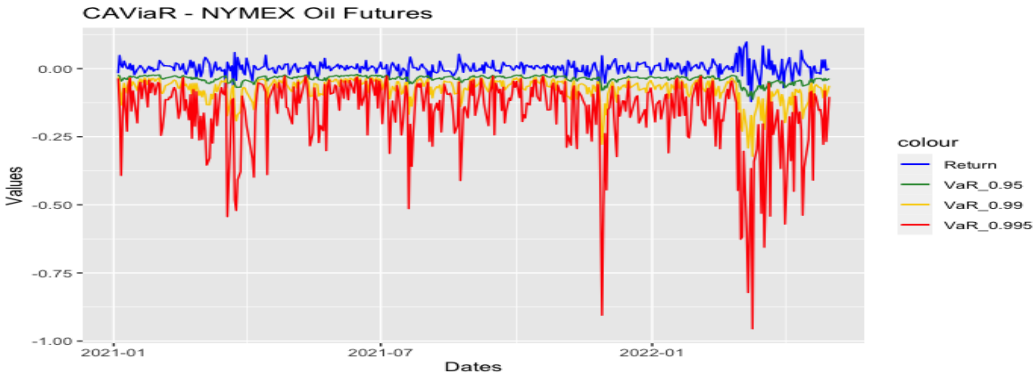

Figure 25: *Historical daily returns of NYMEX Crude Oil Futures for the period 2021-01-04 to 2022-05-01 (400 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

Table 5: *Christoffersen test results for unconditional and conditional coverage of the VaR of NYMEX Crude Oil Futures returns using Autoregressive ERF, GARCH and CAViaR methods for the period 2011-03-29 to 2022-05-01 (2910 observations).*

Model	τ	EE	T_1	LR_{UC}	P_{UC}	LR_{CC}	P_{CC}
Autoregressive ERF	0.95	145	141	0.148	**0.700**	3.099	**0.212**
	0.99	29	32	0.283	**0.595**	4.111	**0.128**
	0.995	14	18	0.764	**0.382**	8.954	0.011
GARCH	0.95	145	153	0.400	**0.527**	0.558	**0.756**
	0.99	29	54	17.186	0.000	17.186	0.000
	0.995	14	39	28.213	0.000	28.567	0.000
CAViaR	0.95	145	145	0.002	**0.966**	3.445	**0.179**
	0.99	29	29	0.000	**0.985**	0.584	**0.747**
	0.995	14	15	0.014	**0.906**	0.169	**0.919**

B.3 NYMEX Natural Gas Futures

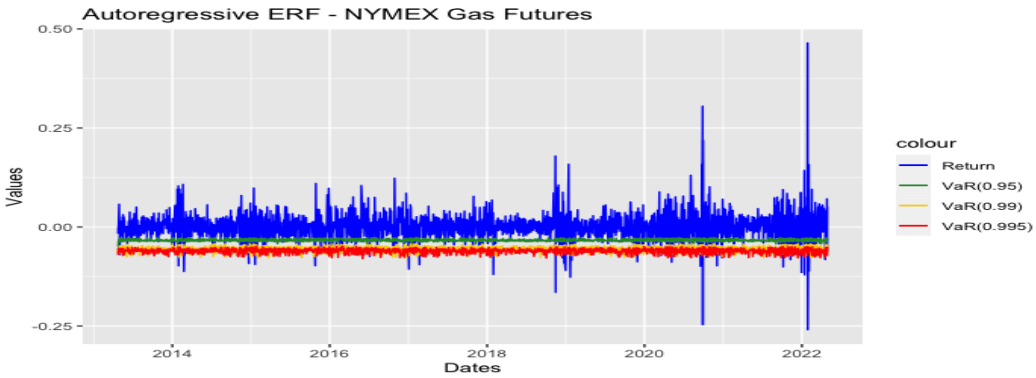

Figure 26: *Historical daily returns of NYMEX Natural Gas Futures for the period 2013-04-24 to 2022-04-29 (2388 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

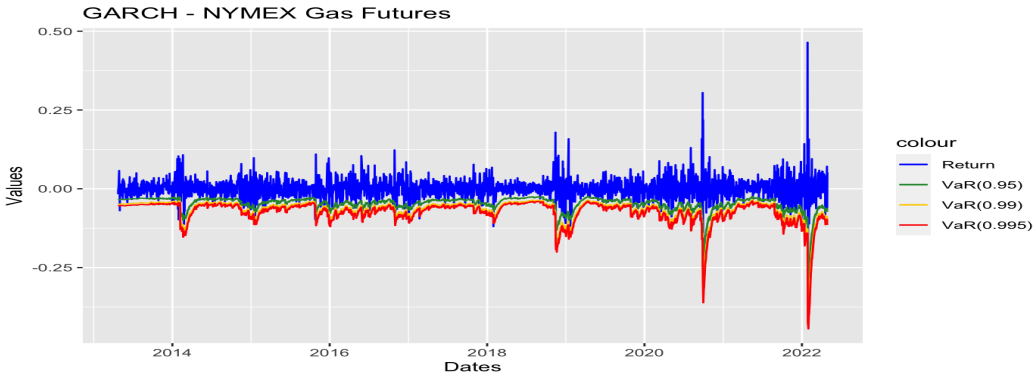

Figure 27: *Historical daily returns of NYMEX Natural Gas Futures for the period 2013-04-24 to 2022-04-29 (2388 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

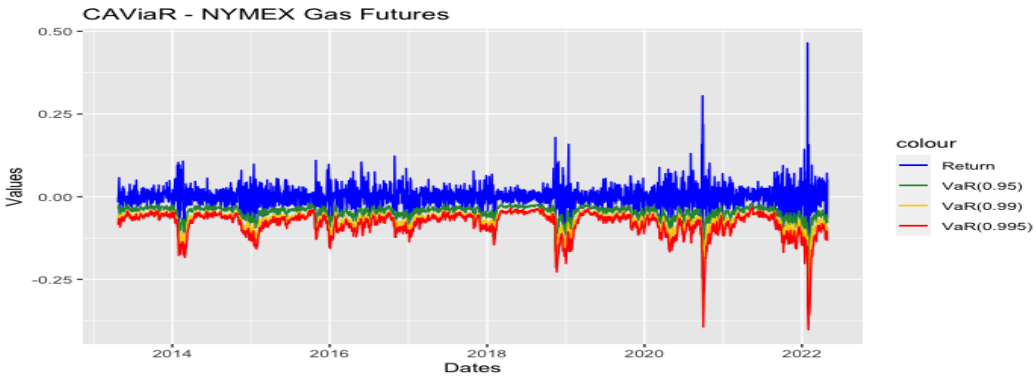

Figure 28: *Historical daily returns of NYMEX Natural Gas Futures for the period 2013-04-24 to 2022-04-29 (2388 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

Table 6: *Christoffersen test results for unconditional and conditional coverage of the VaR of NYMEX Natural Gas Futures returns using Autoregressive ERF, GARCH and CAViaR methods for the period 2013-04-24 to 2022-04-29 (2388 observations).*

Model	τ	EE	T_1	LR_{UC}	P_{UC}	LR_{CC}	P_{CC}
Autoregressive ERF	0.95	119	180	40.076	0.000	46.897	0.000
	0.99	23	64	46.632	0.000	60.026	0.000
	0.995	11	46	56.455	0.000	66.309	0.000
GARCH	0.95	119	110	0.799	**0.371**	0.977	**0.613**
	0.99	23	29	1.038	**0.308**	1.751	**0.417**
	0.995	11	16	1.253	0.263	1.469	0.480
CAViaR	0.95	119	119	0.001	**0.970**	0.204	**0.903**
	0.99	23	25	0.052	**0.819**	0.581	**0.748**
	0.995	11	12	0.000	**0.986**	0.122	**0.941**

B.4 COMEX Gold Futures

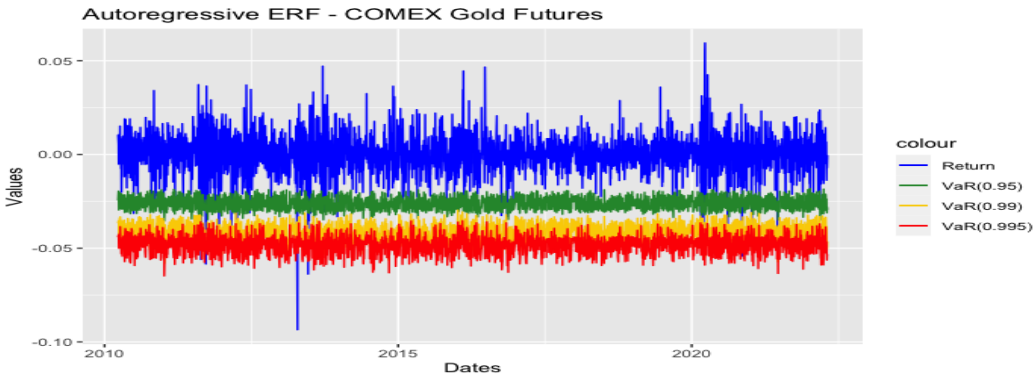

Figure 29: *Historical daily returns of COMEX Gold Futures for the period 2010-03-26 to 2022-04-24 (3162 observations) with estimated VaR using the Autoregressive ERF model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

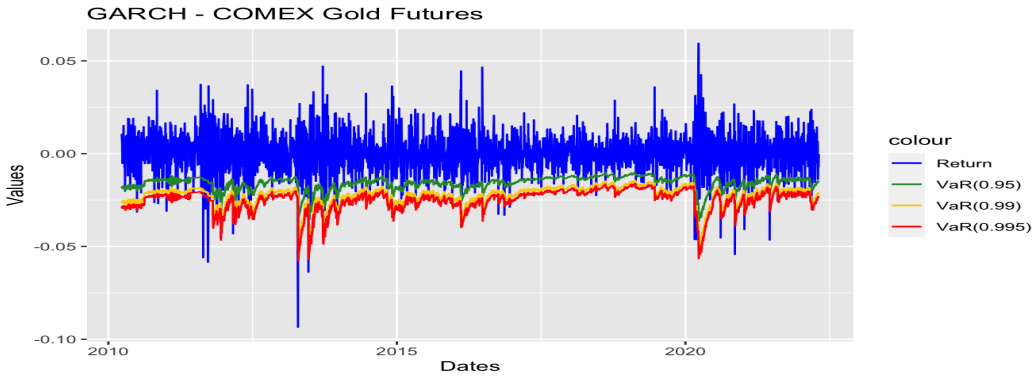

Figure 30: *Historical daily returns of COMEX Gold Futures for the period 2010-03-26 to 2022-04-24 (3162 observations) with estimated VaR using the GARCH(1,1) model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

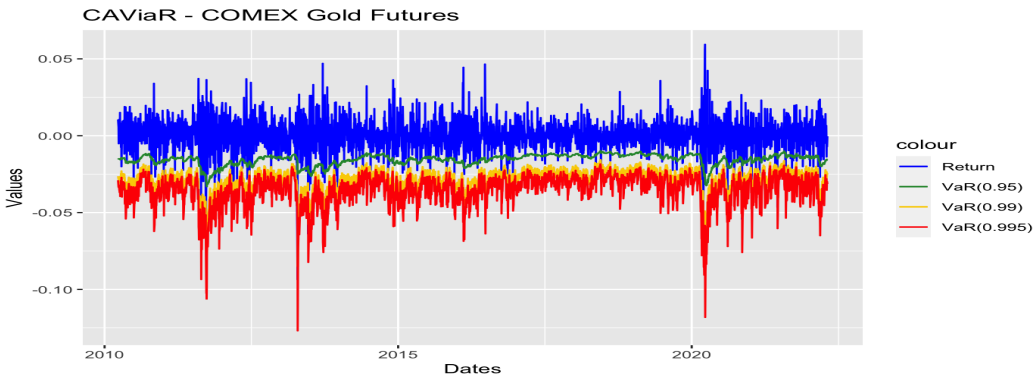

Figure 31: *Historical daily returns of COMEX Gold Futures for the period 2010-03-26 to 2022-04-24 (3162 observations) with estimated VaR using the CAViaR model for quantile levels* $\tau = 0.005, 0.01, 0.05$.

Table 7: *Christoffersen test results for unconditional and conditional coverage of the VaR of COMEX Gold Futures returns using Autoregressive ERF, GARCH and CAViaR methods for the period 2010-03-26 to 2022-04-24 (3162 observations).*

Model	τ	EE	T_1	LR_{UC}	P_{UC}	LR_{CC}	P_{CC}
Autoregressive ERF	0.95	158	42	125.283	0.000	130.786	0.000
	0.99	31	9	22.785	0.000	22.836	0.000
	0.995	15	5	10.145	0.001	10.160	0.006
GARCH	0.95	158	157	0.008	**0.928**	0.204	**0.902**
	0.99	31	68	31.802	0.000	40.304	0.000
	0.995	15	46	38.167	0.000	50.475	0.000
CAViaR	0.95	158	159	0.005	**0.942**	0.005	**0.997**
	0.99	31	32	0.005	**0.946**	0.936	**0.626**
	0.995	15	14	0.217	**0.642**	4.031	**0.133**

References

Angrist, J., Chernozhukov, V., & Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, *74*(2), 539–563.

Angrist, J. D., Chernozhukov, V., & Fernández-Val, I. (2009). *Replication data for: Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure.* Harvard Dataverse. Retrieved from `https://doi.org/10.7910/DVN/JNEOLQ` doi: 10.7910/DVN/JNEOLQ

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Beirlant, J., Wet, T. D., & Goegebeur, Y. (2004). Nonparametric estimation of extreme conditional quantiles. *Journal of statistical computation and simulation*, *74*(8), 567–580.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Chavez-Demoulin, V., & Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(1), 207–222.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, *33*(2), 806–839.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, 841–862.

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: an introduction* (Vol. 21). Springer.

Duffie, D., & Singleton, K. J. (2012). *Credit risk: pricing, measurement, and management.* Princeton university press.

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

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

Gencay, R., & Selçuk, F. (2004). Extreme value theory and value-at-risk: Relative performance in emerging markets. *International Journal of forecasting*, *20*(2), 287–303.

Gnecco, N., Terefe, E. M., & Engelke, S. (2022). Extremal random forests. *arXiv preprint arXiv:2201.12865*.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, *7*(12), 701–702.

Hill, C., & McCullough, B. (2019). On the accuracy of garch estimation in r packages. *Econometric Research in Finance*, *4*(2), 133–156.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Lamoureux, C. G., & Lastrapes, W. D. (1990). Persistence in variance, structural change, and the garch model. *Journal of Business & Economic Statistics*, *8*(2), 225–234.

Martins-Filho, C., Yao, F., & Torero, M. (2018). Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory*, *34*(1), 23–67.

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(6).

Marketwatch. (2021). *Oil prices went negative a year ago: Here's what traders have learned since.* (Retrieved from: `https://www.marketwatch.com/story/oil-prices-went-negative-a-year-ago-heres-what-traders-have-learned-since-11618863839`)

The Observatory of Economic Complexity. (2020). *Wheat exporters and importers.* (Retrieved from: `https://oec.world/en/profile/hs/wheat`)

Odening, M., & Hinrichs, J. (2002). Using extreme value theory to estimate value-at-risk. *Agricultural finance review*.

Ratner, M., et al. (2014). Gold and systemic risk. *The Journal of Alternative Investments*, *17*(1), 79.

Taillardat, M., Fougères, A.-L., Naveau, P., & Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, *34*(3), 617–634.

Taylor, J. W. (2008). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of financial Econometrics*, *6*(3), 382–406.

Velthoen, J., Dombry, C., Cai, J.-J., & Engelke, S. (2021). Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*.

Wang, H. J., & Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *Journal of the American Statistical Association*, *108*(503), 1062–1074.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for us wind gusts. *Journal of the American Statistical Association*, *114*(528), 1865–1879.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 331–350.