STATISTICS IN TRANSITION new series, June 2024
Vol. 25, No. 2, pp. 1–22, https://doi.org/10.59170/stattrans-2024-012
Received – 21.09.2022; accepted – 24.08.2023

Volatility and models based on the extreme value theory for gold returns

Dominik Krężołek[1], Krzysztof Piontek[2]

Abstract

In this study, we use daily gold log-returns to analyse the quality of forecasting expected shortfalls (ES) using volatility and models based on the extreme value theory (EVT). ES forecasts were calculated for conditional APARCH models formed on the entire distribution of returns, as well as for EVT models. The results of ES forecasts for each model were verified using the backtesting procedure proposed by Acerbi and Szekely. The results show that EVT models provide more accurate one-day ahead ES forecasts compared to the other models. Moreover, the asymmetric theoretical distributions for innovations of EVT models allow the improvement of the accuracy of ES forecasting.

Key words: expected shortfall, volatility models, EVT, gold returns, backtesting.

1. Introduction

Economic processes observed in the contemporary world are very complex, varied and unpredictable. This is due to the variety of information that flows into the market. Information derives from data, which represent specific facts generated by the reality. An appropriate understanding of the nature of data is crucial in the decision-making process, for investment decisions. The last decade has also seen an increasing interest in other forms of investment of financial assets than those offered by the classical capital market. This is largely due to the uncertainty and unpredictability of the direction of the global economy. The crisis of 2008–2009 caused those investors to decide to transfer capital into other, alternative assets. One of them is gold. The main reason for seeking new markets is the aim to minimize the risk of undertaken investment activity and to hedge positions against adverse trends in the global economy.

[1] University of Economics in Katowice, Poland. E-mail: dominik.krezolek@ue.katowice.pl. ORCID: https://orcid.org/0000-0002-4333-9405.

[2] Wroclaw University of Economics and Business, Poland. E-mail: krzysztof.piontek@ue.wroc.pl. ORCID: https://orcid.org/0000-0001-9197-561X.

Risk is an integral part of any investment activity. By understanding the sources and factors generating risk, it is possible to manage it efficiently and to minimize its adverse consequences. Risk management is related to the decision-making process and the identification of tools allowing for risk reduction, as well as the construction of strategies enabling its monitoring and reporting. The measurement of risk using an appropriate measure is an important part of the risk management process. The Basel Committee on Banking Supervision based on the Basel III document (the Third Basel Accord or Basel Standards) recommends some risk measures that should be used by financial institutions and investors to ensure adequate capital reserves. These include Value-at-Risk proposed by Risk Metrics, but also Expected Shortfall, which is some extension of VaR.

Value-at-Risk, as a measure of risk, was proposed in 1994 by the US financial institution J.P. Morgan, whose analysts developed a risk management system commonly known as RiskMetrics™. VaR is defined as a statistical measure that assesses (in an unambiguous way) the amount of potential loss in the market value of a financial instrument for which the probability of reaching or exceeding it within a specified time horizon is equal to a tolerance level established by the decision maker [Dowd, 1999; Trzpiot, 2004; Doman, Doman, 2009]. The Expected Shortfall (ES), on the other hand, is a risk measure that expresses the expected value of return at a level exceeding VaR. It is commonly known in the literature as Conditional VaR (CVaR). Its advantage over the discussed VaR consists in the fact that VaR determines the minimum loss from an investment in α possible cases, and thus it unambiguously determines a certain threshold of return. In contrast, the ES focuses also on values exceeding VaR, determining the average level of losses in the conditional sense. Additionally, the ES, unlike VaR, satisfies all conditions of a coherent risk measure: the condition of monotonicity, subadditivity, positive homogeneity and translation invariance [Artzner et al. 1999], so it can be used as an evaluation for risk in the case of complex portfolio structures.

Validation of VaR and ES risk measures aims to verify that the estimated risk measure reliably assesses the actual risk under the given assumptions. The methods of such verification are referred to as backtesting methods. Backtesting is a statistical procedure in which actual profits and losses are compared with the corresponding estimates of the risk measure. The backtesting process verifies the hypothesis that the frequency of occurrence of return over a specified period is consistent with an assumed level of significance. Such tests are referred to as unconditional coverage tests. The implementation and application of these tests are generally straightforward, as they do not consider the point at which the significance level is exceeded. However, from a theoretical point of view, a correct model for risk assessment should not only provide an estimate of the appropriate level of exceedances but also verify that they are uniformly distributed over time, i.e. independent of each other. Clustering of exceedances indicates that the model does not accurately capture market volatility and correlations. Such tests that consider the dynamic aspect of exceedances and their independence are referred to as the conditional coverage tests [Jorion, 2001].

In this article, we attempt to assess the volatility of gold returns. Modeling gold returns can provide valuable insights that help understand the volatility in the market of other financial assets. Investments in gold have many unique characteristics that make them attractive in today's complex financial world. Gold is often seen as a "safe haven" during times of economic and financial uncertainty. When other assets such as stocks or bonds experience significant declines, investors often turn to gold, which usually leads to a price increase. Gold is an asset of particular importance for hedging and diversification of investment portfolios, and therefore it is important to predict future volatility of this asset. Gold is also seen as an effective hedge against inflation. When the value of money decreases, the value of gold often rises, which can help preserve the real value of an investment. Another important fact about gold is related to investment risk. Gold exhibits low correlation with many other asset classes, meaning that its price often behaves differently than other investments. Therefore, adding gold to a portfolio can help diversify risk. Moreover, gold is one of the most universally accepted assets worldwide and can be sold almost anywhere. In addition, gold is a very liquid asset, meaning it can be easily bought and sold [Li et al., 2022].

Given the above, we compare volatility models and EVT models for forecasting extreme risk (ES) and then we apply the approach to backtesting ES proposed by Acerbi and Szekely [Acerbi et al., 2014]. The paper consists of the following sections. Section 2 reviews the literature on VaR and ES risk measures including backtesting. Section 3 describes the methodology used in the study: APARCH model of volatility and some basics of the Extreme Value Theory (EVT). Furthermore, extreme risk measures for volatility and EVT models are defined and the procedure for backtesting ES under the approach of Acerbi and Szekely is described. Section 4 presents the results of the empirical study on the example of gold returns, while in Section 5 all findings are summarized and commented.

2. Literature overview

The problem of risk is a constant object of analysis of researchers around the world. There are many studies that discuss methods of risk measurement using appropriate measures, as well as demonstrate how these methods are used in practice. Risk is related to the issue of volatility, therefore a wide range of methods is based on models describing volatility. In addition, research papers that concern analysis of volatility and risk mainly focus on assets such as stocks, foreign exchange rates, cryptocurrencies, while the application of these methods in alternative investments, such as gold, is less popular.

Zijing and Zhang [Zijing et al., 2016] analyzed volatility in gold returns using GARCH-type models (GARCH, EGARCH, and TGARCH) with an error term described by GED distribution, while Włodarczyk [Włodarczyk, 2017] analyzed asymmetry and long memory effects on forecasting conditional volatility and risk in the gold and silver market using linear and nonlinear GARCH models. Naeem, Tiwari, Mubashra, and Shahbaz [Naeem et al., 2019] examined volatility on precious metals using Markov-Switching GARCH (MSGARCH) models. They revealed the existence of regime shifts in GARCH models and confirmed the advantages of regime-switched models over classical one-dimensional GARCH models. Mensi, Vinh Vo and Hoon Kang [Mensi et al., 2022] examined the volatility spillovers between the US stock market (S&P500) and both oil and gold before and during the global health crisis. They applied the FIAPARCH-DCC model to the 15-minute intraday data. They results showed negative (positive) conditional correlations between the S&P500 and gold (oil). Moreover, they indicated that gold offers more diversification gains than oil does during the pandemic. Vidal and Kristjanpoller [Vidal et al., 2020] investigated the forecast of gold volatility by combining two deep learning methodologies: short-term memory networks (LSTM) added to convolutional neural networks. They highlighted that these types of hybrid architectures have not been used in time series prediction. Their results showed a substantial improvement when this hybrid model was compared to the GARCH and LSTM models. Kayal and Maheswaran [Kayla et al., 2021] discovered the persistence of excess volatility in gold spot price data that engenders excessive path dependence, whereas it is not the same with silver. They used the extreme value estimator and the VRatio and observed that the strong mean-reverting characteristic in gold makes it a better investment choice than silver. Morales and Andreosso-O'Callaghan [Morales et al., 2021] showed that the daily returns of silver have a standard deviation which is more than twice that of gold. Elsayed, Gozgor and Yarovaya [Elsayed et al., 2022] examined the dynamic connectedness of return and volatility spillovers among cryptocurrency index (CRIX), gold, and uncertainty measures. Apart from traditional uncertainty measures, they also considered two novel uncertainty measures: Cryptocurrency Policy Uncertainty and Cryptocurrency Price Uncertainty indices. They observed that cryptocurrency policy uncertainty was the main transmitter of the return spillovers to other variables. In addition, gold was a net receiver of both the return and the volatility spillovers.

In the literature, there are many papers on risk measurement, however, the vast majority concern the Value-at-Risk proposed by RiskMetrics™. Daníelsson, Jorgensen, Samorodnitsky, Sarma and de Vries [Daníelsson et al., 2013] studied the properties of

VaR. They showed that the VaR measure is subadditive in the corresponding tail region of the return distribution. They also noted that estimating VaR using the historical simulation method may lead to a violation of the subadditivity assumption. They proposed to estimate the VaR risk measure using semi-parametric extreme value theory (EVT) methods. Alexander and Sarabia [Alexander et al., 2012] proposed to estimate risk associated with a value-at-risk model and adjust VaR estimates with respect to estimation and model specification errors. Huang, Huang, Chikobvu, and Chinhamu [Huang et al., 2015] predicted VaR using extreme value theory (EVT) and generalized Pareto distribution (GPD). Other researchers analyzed the forecast quality of VaR estimation using GARCH models. This approach was adopted by Walid, Shawkat and Khuong [Walid et al., 2014] by using nonlinear FIAPARCH models. Yu, Yang, Wei and Lei [Yu et al., 2018] measured VaR using GARCH-type models, extreme value theory (EVT) and copula models. The results of backtesting showed that GARCH-EVT type models and Copula models were able to improve the accuracy of VaR estimation. In contrast, Cheung and Yuen [Cheung et al., 2020] introduced an uncertainty model for the distribution of returns and investigated the impact of this volatility on VaR using a worst-case scenario approach. They showed that the choice of loss model is significant when an uncertainty model is implemented. Fiszeder, Fałdziński and Molnar [Fiszeder et al., 2019] propose some modification of multivariate DCC model to calculate forecasts of VaR. They show that regardless of whether in-sample fit, covariance forecasts or value-at-risk forecasts are considered, the model they propose outperforms not only the standard DCC model, but also an alternative range-based DCC model.

Cheng and Hung [Cheng et al., 2011] evaluated the asymmetry and kurtosis of returns distribution in the crude oil and metals market using skewed Student's *t* distribution and GARCH-type volatility models. The empirical results showed that the predictions of VaR obtained using skewed distribution were more accurate in comparison with a symmetric distribution. Eling [Eling, 2014] applied skewed Student's t distribution to risk analysis in insurance. They have shown that the skewed Student's t distribution is a notably promising distribution for modeling returns on assets such as stocks, bonds, monetary market instruments, and hedge funds. Fernandez-Perez, Frijns, Fuertes, and Miffre [Fernandez-Perez et al., 2018] analyzed the relationship between the skewness of the distribution of commodity futures and expected returns.

When it comes to backtesting risk measures, most of the research papers focus on VaR. Only few works deal with Expected Shortfall. Jalal and Rockinger [Jalal et al., 2008] use a circular block bootstrap to consider the possible dependency among exceedances. Applying the two-step procedure, they found that ES forecasts captured actual shortfalls satisfactorily. Righi and Ceretta [Righi et al., 2015] evaluate unconditional,

conditional and quantile (expectile) regression-based models for ES predictions under the ES backtest approach proposed by McNeil and Frey [McNeil et al., 2000]. Clift, Costanzino and Curran [Clift et al., 2016] apply three approaches recently proposed in the literature for backtesting ES consider a GARCH volatility specification with normal distribution for ES forecasting. Kratz, Lok and McNeil [Kratz et al. 2019] demonstrate that backtests of the forecasting models used to derive ES can be based on a multinomial test of Value-at-Risk exceptions at several levels, using heavy-tailed distributions and GARCH volatility models. Bu, Liao, Shi, and Peng [Bu et al., 2019] propose a new method to capture the dynamics of ES across time horizons using wavelet analysis. Their results confirm that the different frequency components of stock returns exhibit different persistence. Lazar and Zhang [Lazar et al., 2019] propose to measure the model risk of Expected Shortfall as the optimal correction needed to pass several ES backtests. They also investigate properties of proposed measures of model risk using GARCH models. del Brio, Mora-Valencia and Perote [del Brio et al., 2020] apply backtesting techniques for both Value-at-Risk and Expected Shortfall under parametric and semi-nonparametric approaches for modeling commodity ETFs. They recommend the application of leptokurtic distributions and semi-nonparametric techniques to mitigate regulation concerns about global financial stability of commodity business. Argyropoulos and Panopoulou [Argyropoulos et al., 2019] reviews the major VaR and ES forecast evaluation methods and evaluates their performance under a common simulation and financial application framework. They suggest that focusing on specific individual hypothesis tests provides a more reliable alternative than the corresponding conditional coverage ones. Acereda, Leon and Mora [Acereda et al., 2020] calculate ES risk measure for distributions of cryptocurrencies using various error distributions and GARCH-type models. Their results highlight the importance of estimating ES for cryptocurrencies using a generalized GARCH model and a non-normal error distribution with at least two parameters.

The method of modeling volatility using heavy-tailed distributions and Extreme Value Theory, which we present in this paper, is particularly relevant to extreme risk analysis. Many traditional methods of financial analysis assume that returns follow a normal distribution. However, in practice, returns often exhibit "heavy tails" meaning that extremely large gains or losses are more likely than a normal distribution would predict. Heavy-tailed distributions and EVT provide better modeling for these extreme events. On the other hand, analyzing heavy-tailed distributions and EVT can help investors understand the risk associated with potentially large losses. This knowledge can be useful in creating risk management strategies and diversifying portfolios. EVT is specifically designed for modeling and predicting extreme events. This can be especially useful for modeling the risk of a financial crisis or other "black swan" type events. Furthermore, using heavy-tailed distributions and EVT can increase the credibility of financial modeling and forecasting results. Results based on these methods may be more robust to extreme events.

3. Methodology

Considering vast types of processes observed in the world around us it is possible to detect unexpectable events that generate risk at a level very far from the expected one. That type of risk is called the extreme risk and is associated with events occurring with low probability, but if they do occur, they generate extreme losses [Jajuga, 2008]. If extreme risk analysis is of interest, the theory of extreme events plays a special role. Extreme statistics are used to estimate some characteristics which help to define rare events. These statistics are e.g. quantiles of empirical distributions of examined phenomena or parameters defining periods, in which the analyzed processes take some extreme values [Gumbel, 2004]. During last century we can list many examples of events that have caused various types of drastic changes in the market, for example Black Monday (19 Oct 1987), World Trade Center (11 Sept 2001), recent financial crisis (2008–2009), crisis on crude oil market (2014) or pandemic of COVID-19 (March 2020 until present). All these events had a significant impact on the volatility of market processes and thus also on the level of risk.

3.1. Risk measures for conditional volatility models

In this paper we consider the APARCH model for volatility [Ding et al., 1993]. APARCH, which stands for Asymmetric Power ARCH, is an extension of GARCH class models that introduces certain additional features. APARCH(1,1) is less complex than APARCH(p,q) models with higher orders p and q. Less complex models are usually easier to estimate, interpret, and validate. High model complexity can lead to overfitting, which means that the model fits the training data well but performs poorly on test data or new data. The main advantages of the APARCH model over other GARCH-type models are:

- Modeling asymmetry (leverage effect): The APARCH model allows for the modeling of asymmetry, also known as the leverage effect, which is common in financial data. The leverage effect is the phenomenon where negative price changes have a greater impact on volatility than positive changes of the same magnitude. Most standard GARCH models do not account for this effect.
- Flexibility in modeling extreme events: The APARCH model allows for more flexibility in modeling extreme events (known as 'fat tails'), which are often observed in financial data.
- Powering of variance: The APARCH model allows for the powering of the variance (or conversely, the standard deviation) to any non-necessarily integer

power. This feature is useful in situations where we care about modeling the volatility directly (e.g. percentage volatility), rather than the variance.

Taking into account the characteristics above, we estimate volatility according to APARCH(1,1) model defined by the equation:

$$\sigma_t^\delta = \omega + \alpha_1(|\varepsilon_{t-i}| - \gamma_1\varepsilon_{t-i})^\delta + \beta_1\left(\sigma_{t-j}\right)^\delta \tag{1}$$

where $\omega, \alpha_i, \gamma_i, \beta_j, \delta$ are unknown model parameters.

These parameters play an important role in understanding the concept of the APARCH model:

- ω: This is the so-called "intercept" parameter. It influences the average level of the series' conditional variance.
- α_1: This parameter measures the impact of the squared error from the previous period on today's variance. It determines how much a large (or small) error value in the previous period increases (or decreases) the predicted variance today.
- γ_1: This parameter introduces asymmetry into the model. It determines how different the effects on variance of positive and negative errors from the previous period are.
- β_1: This parameter measures the impact of the predicted variance from the previous period on today's variance. It determines how persistent the effects of shocks on variance are.
- δ: This parameter determines the power to which the conditional standard deviation is raised. It allows for the modeling of the variance (or conversely, the standard deviation) to any non-necessarily integer power.

Moreover, the parameters of any APARCH model must satisfy certain conditions for the model to be well-defined. For instance, for the APARCH(1,1) model, the following conditions must be satisfied: $\omega > 0$, $\alpha_1 \geq 0, \beta_1 \geq$, $\alpha_1 + \beta_1 < 1$, and $\delta > 0$. For error term we consider t-Student, skewed t-Student, GED, and skewed GED distributions defined by the following probability distribution functions:

- Student's t distribution:

$$f(z|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{z^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function and v is defined as degrees of freedom ($v > 0$).

- Skewed Student's t distribution:

$$f(z|\xi, v) = \frac{2}{\xi+(\xi)^{-1}} s\{g[\xi(sz + m)|v]\mathrm{I}_{(-\infty,0)}(z + ms^{-1}) + \\ g\left[\left(\frac{sz+m}{\xi}\right)\middle|v\right]\mathrm{I}_{(0,+\infty)}(z + ms^{-1})\} \tag{3}$$

where $g(\cdot\,|v)$ is the density of symmetric Student's t distribution, ξ is the skewness parameter defined as $\xi^2 = \frac{P(z \geq 0|\xi)}{P(z<0|\xi)}$ and v is defined as degrees of freedom $(v > 0)$. Moreover, two additional parameters m (for mean) and s^2 (for variance) must be defined:

$$m = E(\varepsilon|\xi\ ) = M_1(\xi - \xi^{-1}) \tag{4}$$

$$s^2 = Var(\varepsilon|\xi) = (M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2 \tag{5}$$

where $M_r = 2\int_0^{+\infty} s^r g(s)ds$ is the absolute moment generating function.

- GED distribution:

$$f(z|v) = \frac{v}{\left[2^{-\left(\frac{2}{v}\right)}\frac{\Gamma\left(\frac{1}{v}\right)}{\Gamma\left(\frac{3}{v}\right)}\right]^{\frac{1}{2}} 2^{\left(1+v^{-1}\right)}\Gamma\left(\frac{1}{v}\right)} \exp\left(\frac{z}{2v}\right) \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function and v is defined as degrees of freedom $(v > 0)$.

- Skewed GED distribution:

$$f(z|\xi, v) = \frac{2}{\xi + (\xi)^{-1}} s\{g[\xi(sz + m)|v]\mathrm{I}_{(-\infty,0)}(z + ms^{-1}) + \\ g\left[\left(\frac{sz+m}{\xi}\right)\Big|v\right]\mathrm{I}_{(0,+\infty)}(z + ms^{-1})\} \tag{7}$$

where $g(\cdot\,|v)$ is the density of symmetric GED distribution, ξ is the skewness parameter defined as $\xi^2 = \frac{P(z \geq 0|\xi)}{P(z<0|\xi)}$ and v is defined as degrees of freedom $(v > 0)$.

All parameters of APARCH(1,1) model have been estimated using Maximum Likelihood Estimation (MLE).

For conditional volatility model we compute VaR and ES risk measures using formulas:

$$VaR_t^{\alpha} = \mu_t + \sigma_t F^{-1}(\alpha) \tag{8}$$

$$ES_t^{\alpha} = \mu_t + \sigma_t \left(\frac{1}{\alpha}\int_0^{\alpha} F^{-1}(s)ds\right) \tag{9}$$

3.2. Risk measures for Extreme Value Theory models (EVT)

Another way to calculate measures of extreme risk is to estimate the conditional quantile using Extreme Value Theory (EVT). Extreme Value Theory is a branch of statistics dealing with statistical methods and the probabilistic and statistical theory related to extreme events, which are often significant in areas such as meteorology, hydrology, finance, insurance, or even structural engineering. One of the key concepts in EVT is the Generalized Extreme Value (GEV) distribution. It combines three types of extreme value distributions: the Gumbel, Frechet, and Weibull families of distributions. In ETV we deal with some extremes. There are two main types of extremes [Gumbel, 2004]:

- Block maxima: This is the maximum value of a block of data. It is like considering the maximum rainfall recorded every month, for instance.
- Peak over threshold: This considers all values over a certain high threshold, not just the maxima.

In Extreme Value Theory certain theorems play a special role. The first one – the Fisher-Tippett-Gnedenko theorem – is a critical theoretical foundation of EVT, which states that with proper normalization, the maxima of a sequence of random variables converge in distribution to one of the three types of extreme value distributions mentioned above. The Pickands–Balkema–de Haan theorem is another important theoretical foundation of EVT for the peaks-over-threshold approach. This theorem states that above a sufficiently high threshold, the excess distribution over that threshold can be approximated by a Generalized Pareto distribution [Fałdziński, 2014].

EVT provides a rigorous way to make statistical inferences about rare events (extreme events). Because of its ability to predict such events, EVT is increasingly being applied in various fields such as finance, insurance, and environmental science.

The extreme value distribution (EVD) can be described using the following density function [Gumbel, 2004]:

$$EVD_{\gamma}(x) = \begin{cases} exp\left[-(1+\gamma x)^{-1/\gamma}\right],\ 1+\gamma x \geq 0 & for\ \ \gamma \neq 0 \\ exp[-exp(-x)],\ x \in \mathbb{R} & for\ \ \gamma = 0 \end{cases} \tag{10}$$

where γ defines the extreme value index (EVI). It is the most important parameter in this distribution, which measures the thickness of its tail (and thus the probability of an extreme event occurring). The heavier the tail, the higher the EVI value. In the case of a generalized EVD, the tail index is the shape parameter, which is invariant of standardizing the distribution [Németh et al., 2018]. The EVI can be estimated using Hill [Hill, 1975] or Picands estimator [Picands, 1975]. In this paper we use the Peaks-Over-Threshold (POT) approach based on EVT and on the generalized Pareto distribution (GPD), which is the limiting tail distribution for a wide variety of continuous probability distributions.

The POT method is a classical approach used in EVT estimation. It consists of fitting the GPD distribution to the innovations obtained from filtering returns using the model of conditional volatility. For i.i.d. random variable, consider the distribution function of excesses $Y = u - Z$ for a given threshold u, $F_u(y) = P(Y = u - Z \leq y | Z < u) = \frac{[F(u) - F(u-y)]}{[F(u)]}, y \geq 0$. The excesses over threshold u follow GPD distribution, $Y = u - Z \sim GPD(\xi, \beta)$:

$$F_u(y) \approx GPD_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\beta}\right), & \xi = 0 \end{cases} \tag{11}$$

$GPD_{\xi,\beta}(y)$ provides $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\frac{\beta}{\xi}$ if $\xi < 0$, where $\beta > 0$ is the scale parameter and ξ is the shape parameter of the tail of the distribution. Consider the following equation for points $z < u$ in the left tail of F as:

$$F(z) = F(u) - F_u(u - z)F(u) = F(u)(1 - F_u(u - z)) \tag{12}$$

Using the proportion of a tailed data $\frac{T_u}{T}$, the tail estimator of GPD is of the form:

$$\hat{F}(z) = \frac{T_u}{T}\left(1 + \hat{\xi}\,\frac{u - z}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}} \tag{13}$$

All parameters of EVT model have been estimated using Maximum Likelihood Estimation (MLE).

According to the EVT approach, formulas for calculating VaR and ES are presented by:

$$VaR_t^{\alpha} = \mu_t + \sigma_t F_Z^{-1}(\alpha) = \mu_t + \sigma_t\left(u + \frac{\beta}{\xi}\left[1 - \left(\frac{\alpha}{\frac{T_u}{T}}\right)^{-\xi}\right]\right) \tag{14}$$

$$ES_t^{\alpha} = \mu_t + \sigma_t\left(\frac{1}{\alpha}\int_0^{\alpha} F_Z^{-1}(s)ds\right) = \mu_t + \sigma_t\left[\frac{VaR_t^{\alpha}}{1-\xi} - \left(\frac{\beta + \xi u}{1-\xi}\right)\right] \tag{15}$$

3.3. Backtest for Expected Shortfall

In this paper we focus on Expected Shortfall backtesting procedure proposed by Acerbi and Szekely [Acerbi et al., 2014]. This is one of many possible approaches to testing this risk measure proposed in the literature. They present two non-parametric tests, both free from assumptions about the probability distribution of the returns. The advantage of this approach is that it does not require any particular form of theoretical distribution, only the continuity of the distribution function along with independence of observation in the sample. To estimate p-values an algorithm based on Monte Carlo simulations is used. The first test statistics, for testing ES after VaR, is of the form:

$$Z_1 = \frac{1}{N_T}\sum_{t=1}^{T}\frac{I_t r_t}{ES_t^{\alpha}} - 1 \tag{16}$$

where $N_T = \sum_{t=1}^{T} I_t > 0$ with $I_t = \mathbb{I}_{\{r_t < VaR_t^\alpha\}}$ is the indicator function of VaR violations and T is the length of the out-of-sample period.

The null hypothesis says that $P_t^\alpha = F_t^\alpha$ for all t, where F_t^α is the tail of cumulative distribution of forecasts at time t when $r_t < VaR_t^{\alpha,F}$ and P_t^α is the tail of the unknown distribution from which the realized returns r_t are drawn. The risk measures VaR and ES under the theoretical and empirical distributions are denoted by $VaR_t^{\alpha,P}$, $ES_t^{\alpha,P}$, $VaR_t^{\alpha,F}$, $ES_t^{\alpha,F}$. The alternative hypothesis says that $ES_t^{\alpha,P} \leq ES_t^{\alpha,F}$ for all t and $ES_t^{\alpha,P} < ES_t^{\alpha,F}$ for some t, together with $VaR_t^{\alpha,P} = VaR_t^{\alpha,F}$ for all t. We can find that the predicted value of VaR^α is still correct for alternative hypothesis, according to the idea that this test is subordinate to the initial test of VaR. This test is in fact completely insensitive to an excessive number of exceptions, since it is the average of the exceptions taken over themselves. Assuming these conditions $E_{H_0}[Z_1 | N_T > 0] = 0$ and $E_{H_1}[Z_1 | N_T > 0] > 0$.

The second test statistics, for testing ES directly, is of the form:

$$Z_2 = \frac{1}{T\alpha} \sum_{t=1}^{T} \frac{I_t r_t}{ES_t^\alpha} - 1 \tag{17}$$

provided that $N_T = \sum_{t=1}^{T} I_t > 0$ with $I_t = \mathbb{I}_{\{r_t < VaR_t^\alpha\}}$ is the indicator function of VaR violations and T is the length of the out-of-sample period.

The null hypothesis says that $P_t^\alpha = F_t^\alpha$ for all t, where F_t^α is the tail of cumulative distribution of forecasts at time t when $r_t < VaR_t^{\alpha,F}$ and P_t^α is the tail of the unknown distribution from which the realized returns r_t are drawn. The alternative hypothesis says that $ES_t^{\alpha,P} \leq ES_t^{\alpha,F}$ for all t and $ES_t^{\alpha,P} < ES_t^{\alpha,F}$ for some t, together with $VaR_t^{\alpha,P} \leq VaR_t^{\alpha,F}$ for all t. As $E_{H_0}[N_T] = T\alpha$ we have $E_{H_0}[Z_2] = 0$ and $E_{H_1}[Z_2] = 0$.

Unlike the Z_1 statistics, the sum of VaR violation event returns is now divided by the expected value. Statistics Z_2 will tend to reject the large number of small VaR violation events. This leads to a difference in alternative hypothesis between the two statistics. Rejecting null hypothesis in Z_2 means rejecting VaR as being correctly defined. The advantage of the proposed approach is its relative computational simplicity. The disadvantage is the requirement of using Monte Carlo simulations to obtain critical values and p-values for the test statistic.

4. Empirical study

The empirical study is based on daily log-returns of gold quoted from January 2015 to December 2021 on the London Metal Exchange. The research period was divided into two sub-periods:

- sub-period of the models' parameter estimation: 2015–2017,
- sub-period of the ES forecast: 2018–2021.

In this research we investigate risk using Expected Shortfall (ES). Expected Shortfall (ES) and Value at Risk (VaR) are both measures used in financial risk management to quantify the level of financial risk within a firm or investment portfolio over a specific time frame. However, they differ in how they approach this risk assessment. VaR measures the maximum loss that will not be exceeded with a certain confidence level. In contrast, ES, also known as Conditional VaR (CVaR), estimates the expected loss given that a loss is greater than the VaR. It provides a more comprehensive view of risk by not only considering the worst-case scenarios but also their potential severity. VaR has been criticized for not adequately capturing tail risk, which refers to extreme events that have low probability but high impact. ES is designed to overcome this shortfall by focusing on the tail of the loss distribution, providing a more accurate measure of potential losses in extreme events. ES is a coherent risk measure, meaning it satisfies properties such as subadditivity. Subadditivity implies that diversifying a portfolio reduces risk, a property that VaR does not have. This can be particularly important for risk management, as it encourages appropriate risk diversification. Following the 2008 financial crisis, many financial regulators have favored ES over VaR as a risk measure. For example, the Basel Committee on Banking Supervision recommended the use of ES for determining regulatory capital requirements due to its ability to better capture tail risk.

The methodology described in Section 3 was used to construct the models. All risk measures were estimated for quantiles of 0.01 and 0.05. Table 1 presents the descriptive statistics of the gold return.

Table 1. Descriptive statistics for gold returns

Statistics	Sub-period 2015–2017	Sub-period 2018–2021
Mean	0.00012	0.00033
Standard error	0.00031	0.00027
Median	0.00007	0.00072
Standard deviation	0.00852	0.00871
Coefficient of variation [%]	7019.87	2647.64
Kurtosis	2.53925	4.80494
Skewness	0.19121	-0.60820
Range	0.08087	0.10212
Minimum	-0.03409	-0.05849
Maximum	0.04678	0.04363
N	755	1032
Kolmogorov-Smirnov test for normality	0.07214	0.06243
p-value	<0.001*	<0.001*

* statistical significance at 0.05.

The average values of the gold return in both sub-periods are similar, but the distributions differ. In the first sub-period (2015–2017), the distribution of gold return is skewed to the right, while in the second sub-period it is skewed to the left. The distributions are characterized by a high level of kurtosis. Moreover, the hypothesis that empirical distributions follow a normal distribution was rejected. Time series and empirical distributions of gold return within both sub-periods and along with the fitted normal distribution are shown in Figure 1–2.

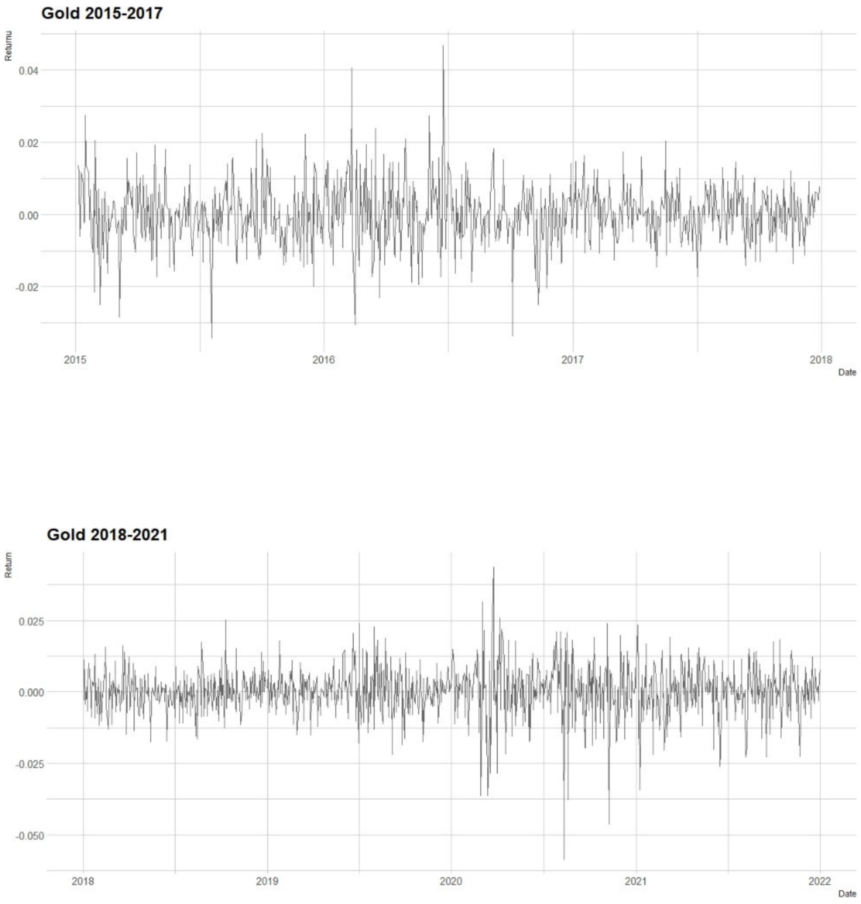
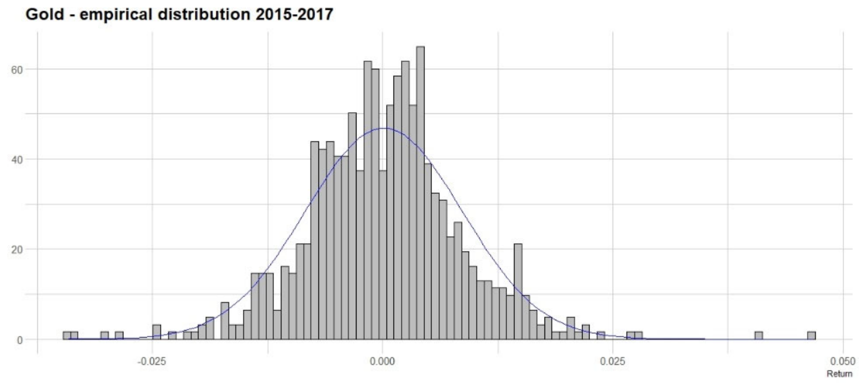


Figure 1: Time series for gold returns (subperiod 2015–2017 — top, subperiod 2018–2021 — bottom).



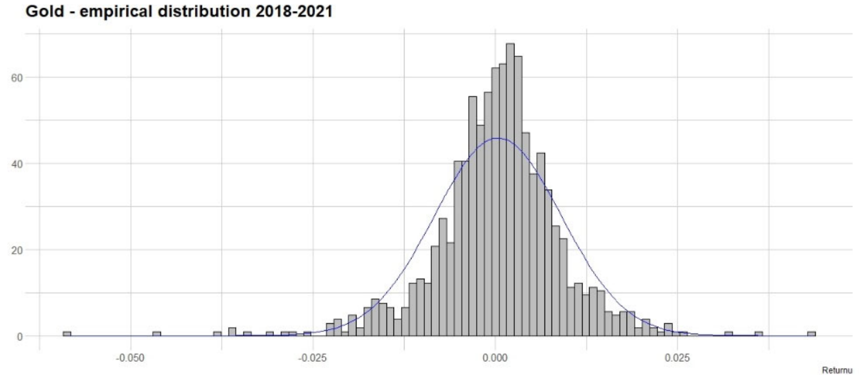


Figure 2: Empirical distributions with fitted normal densities for gold returns (subperiod 2015–2017 — top, subperiod 2018–2021 — bottom).

In the next step of the analysis, the parameters of the conditional volatility and EVT models were estimated. The original study used a variety of volatility models such as GARCH, APARCH, EGARCH, TGARCH, FIGARCH, GARCH-GJR for errors term described by normal, t-Student, skewed t-Student, GED and skewed GED distributions. However, only for APARCH models the largest number of statistically significant parameters and the smallest values of information criterion AIC were obtained. Therefore, only the results for APARCH models are presented in this paper. The results are presented in Table 2.

Table 2: Estimates of model parameters

Model	Parameter	Error~ST		
		Estimates	Standard Error	p-value
APARCH(1,1)	ω	0.0000	0.0000	0.499
	α_1	0.0261	0.0092	0.004*
	γ_1	-0.7735	1.1915	0.516
	β_1	0.9495	0.0445	<0.001*
	δ	1.3537	0.4944	0.006*
EVT	ν	5.6466	1.2022	<0.001*
	ξ	-	-	-
	ξ_{GPD}	0.3431	0.1217	0.001*
	β_{GPD}	0.5543	0.0862	<0.001*
Model	Parameter	Error~ST$_{\text{skewed}}$		
		Estimates	Standard Error	p-value
APARCH(1,1)	ω	0.0000	0.0000	0.495
	α_1	0.0260	0.0089	0.001*
	γ_1	-0.7716	1.1706	0.510
	β_1	0.9498	0.0439	<0.001*
	δ	1.3538	0.4898	0.006*
EVT	ν	5.6524	1.2025	<0.001*
	ξ	0.0095	0.0527	0.856
	ξ_{GPD}	0.3592	0.0621	<0.001*
	β_{GPD}	0.5733	0.0723	<0.001*
Model	Parameter	Error~GED		
		Estimates	Standard Error	p-value
APARCH(1,1)	ω	0.0000	0.0000	0.535
	α_1	0.0185	0.0100	0.006*
	γ_1	-0.4033	0.4030	0.3169
	β_1	0.9776	0.0149	<0.001*
	δ	1.3598	0.3554	<0.001*
EVT	ν	1.3511	0.1000	<0.001*
	ξ	-	-	-
	ξ_{GPD}	0.3318	0.0045	0.001*
	β_{GPD}	0.5982	0.0788	<0.001*
Model	Parameter	Error~GED$_{\text{skewed}}$		
		Estimates	Standard Error	p-value
APARCH(1,1)	ω	0.0000	0.0000	0.538
	α_1	0.0181	0.0098	0.007*
	γ_1	-0.4070	0.4012	0.310
	β_1	0.9781	0.0148	<0.001*
	δ	1.3705	0.3489	<0.001*
EVT	ν	1.3512	0.0995	<0.001*
	ξ	0.0121	0.0527	0.818
	ξ_{GPD}	0.3644	0.1240	<0.001*
	β_{GPD}	0.5417	0.0059	<0.001*

* Statistical significance at 0.05.

Most parameters for the estimated conditional volatility models and all parameters in the EVT models turned out to be statistically significant. A strong long memory effect was observed in the APARCH models (statistically significant, positive values of β_1). Considering the Akaike Information Criterion, the best model is the one in which innovations are described by a t-Student distribution.

Then, the estimated models were used to determine one-day ES forecasts within the period of 2018–2021. The average ES value was computed, and the rate of its violations was calculated for the fixed quantile level. The convergence of the theoretical estimates with the real averaged ES value in the 2018–2021 sub-period was assessed using the RMSE. The p-values for both Z_1 and Z_2 statistics in Acerbi and Szekely approach were also estimated. The results are presented in Table 3.

Table 1: 1-day Ahead average forecasts of ES and p-values for Z_1 and Z_1 statistics

Quantile	Model	$\overline{ES}$	% of violations	p-value Z_1	p-value Z_2	RMSE
0.01	Empirical	-0.0438	0.0100	-	-	-
	APARCH-St	-0.0310	0.0140	0.001*	0.000*	0.0128
	APARCH-St$_{\text{skewed}}$	-0.0316	0.0140	0.003*	0.002*	0.0122
	APARCH-GED	-0.0294	0.0160	0.004*	0.000*	0.0144
	APARCH-GED$_{\text{skewed}}$	-0.0300	0.0160	0.002*	0.001*	0.0138
	EVT-St	-0.0465	0.0090	**0.895**	**0.970**	0.0027
	EVT-St$_{\text{skewed}}$	-0.0474	0.0100	**0.910**	**0.922**	0.0036
	EVT-GED	-0.0441	0.0090	**0.956**	**0.946**	0.0004
	EVT-GED$_{\text{skewed}}$	-0.0450	0.0100	**0.944**	**0.970**	0.0012
0.05	Empirical	-0.0275	0.0500	-	-	-
	APARCH-St	-0.0188	0.0410	0.116	0.417	0.0087
	APARCH-St$_{\text{skewed}}$	-0.0192	0.0430	0.110	0.398	0.0083
	APARCH-GED	-0.0179	0.0440	0.084	0.786	0.0096
	APARCH-GED$_{\text{skewed}}$	-0.0182	0.0440	0.104	0.747	0.0093
	EVT-St	-0.0282	0.0490	**0.851**	**0.922**	0.0007
	EVT-St$_{\text{skewed}}$	-0.0288	0.0480	**0.864**	**0.876**	0.0013
	EVT-GED	-0.0268	0.0500	**0.909**	**0.898**	0.0007
	EVT-GED$_{\text{skewed}}$	-0.0273	0.0510	**0.896**	**0.922**	0.0001

* Statistical significance at 0.05.

The results show that regardless of the quantile level there is no reason to reject the null hypothesis in Z_1 and Z_2 tests for models based on EVT. The conclusion seems to be logical, because EVT models focus only on values in the tail of the distribution, while conditional volatility models are built using all values of the return. Thereby, the high quality of ES forecasts for EVT models may be confirmed. The averaged ES values for each model calculated in the second sub-period are presented in Figures 3–4.

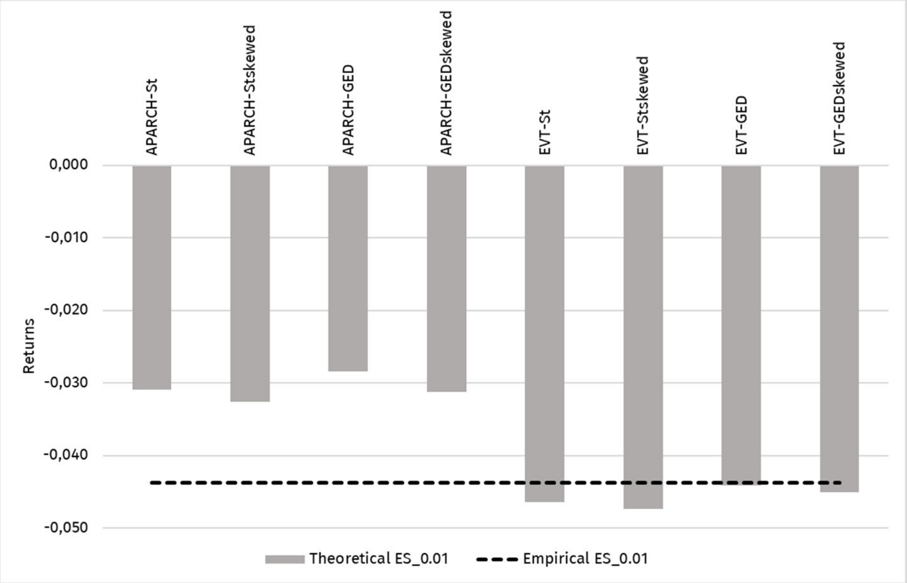


Figure 3: Empirical vs. theoretical averaged ES forecasts for quantile 0.01.

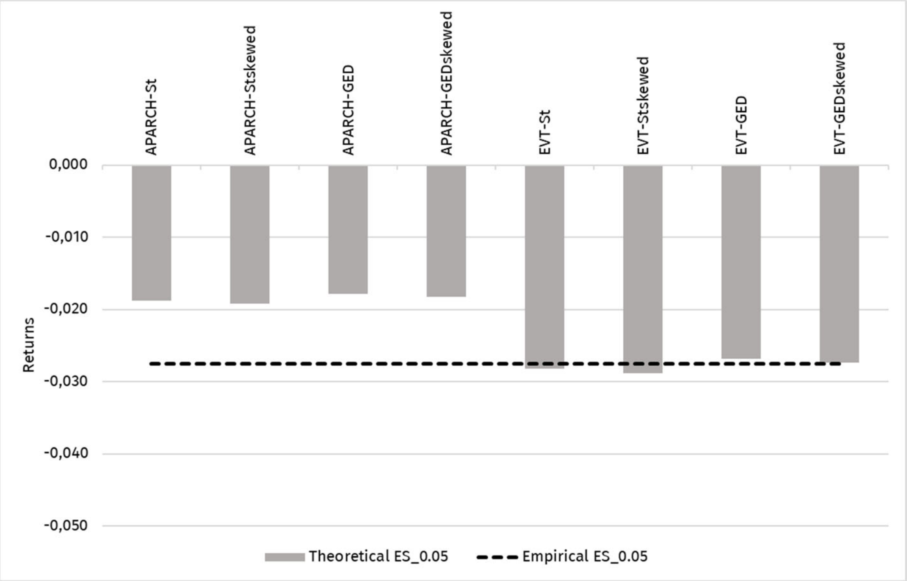


Figure 4: Empirical vs. theoretical averaged ES forecasts for quantile 0.05.

Both figures show that conditional volatility models underestimate ES values, regardless of the quantile level. It is observed that the RMSE values for the EVT models are quite low.

3. Conclusions

In this paper we examined risk using volatility and Extreme Value Theory models. We focus on estimating extreme changes in gold returns for which risk is estimated using ES. There are several reasons why investing in gold is important. First, gold is often seen as a "safe haven" during times of economic and financial uncertainty. When other assets such as stocks or bonds experience significant declines, investors often turn to gold, which usually leads to a price increase. Second, gold is also seen as an effective hedge against inflation. In terms of risk diversification, gold has a low correlation with many other asset classes, which means that its price often behaves differently than other investments. Therefore, adding gold to a portfolio can help diversify risk. The analysis of gold prices is important looking at the global demand. Gold is a unique asset that has both investment and consumption value (e.g. in jewelry). The rise of the middle class in developing countries like China and India may lead to increased demand for gold. The amount of gold in the world is limited, and the extraction process is difficult and costly. This limited supply helps maintain gold's value. Moreover, gold is one of the most universally accepted assets worldwide and can be sold almost anywhere and can be easily bought and sold.

ES is one of the most popular risk measures. As mentioned in Section 4, there are several factors in favour of using this measure instead of VaR. ES is a coherent risk measure and provides a more comprehensive view of risk by not only considering the worst-case scenarios but also their potential severity. Moreover, ES focuses on the tail of the loss distribution, providing a more accurate measure of potential losses in extreme events.

Backtesting provides the means of determining the accuracy of risk forecasts and the corresponding risk model. In this paper we used APARCH conditional volatility models and Extreme Value Theory models to evaluate ES forecasts of gold returns. In addition, we assume different heavy-tailed error distributions. Our study shows that the process of gold returns is characterised by significant, unpredictable, and heterogeneous volatility (e.g. the apparent effect of the COVID-19 pandemic). Moreover, empirical distributions of gold returns were found to be characterized by clustering of variance, leptokurtosis, asymmetry, and fat tails (compared to a normal distribution). A strong long memory effect (statistically significant positive parameter β_1) was observed in the conditional volatility models. We found out that risk estimates based on EVT are closer to real values of ES for EVT models than for APARCH models. The best results were obtained for EVT-GED and EVT-SkewedGED models. ES forecasts obtained for conditional models have p-values close to zero - the null hypothesis is rejected due underestimation of risk. Significant differences in p-value levels were observed between APARCH conditional models and EVT conditional models. The p-value for the Z_1 statistics in the EVT model means that the average of realised ES exceedances is lower than predicted, while for Z_2 additionally that also the percentage of exceedances is lower. Thus, the validity of conditional EVT models was demonstrated. Nevertheless, these are general conclusions, requiring further in-depth research.

References

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

Alexander, C., Sarabia, J. M., (2012). Quantile uncertainty and Value-at-Risk model risk. *Risk Anal. Int. J.*, 32 (8), pp. 1293–1308.

Argyropoulos, Ch., Panopoulou, E., (2019). Backtesting VaR and ES under the magnifying glass. *International Review of Financial Analysis*, 64, pp. 22–37.

Artzner, P., Delbaen, F., Eber, J. M., Heath, D., (1999). Coherent Measures of Risk, *Mathematical Finance*, Vol. 9(3), pp. 203–228.

Bu, D., Liao, Y., Shi, J., Peng, H., (2019). Dynamic expected shortfall: A spectral decomposition of tail risk across time horizons. *Journal of Economic Dynamics & Control*, 108, 103753.

Cheng, W-H., Hung, J-C., (2011). Skewness and leptokurtosis in GARCH-typed VaR estimation of petroleum and metal asset returns. *Journal of Empirical Finance*, 18, pp. 160–173.

Cheung, K. C., Yuen F. L., (2020). On the uncertainty of VaR of individual risk. *Journal of Computational and Applied Mathematics*, 367, 112468.

Clift, S. S., Costanzino, N., Curran, M., (2016). Empirical Performance of Backtesting Methods for Expected Shortfall, http://dx.doi.org/10.2139/ssrn.2618345.

Daníelsson, J., Jorgensen, B. N., Samorodnitsky, G., Sarma, M., de Vries, C. G., (2013). Fat tails, VaR and subadditivity. *Journal of Econometrics*, 172, pp. 283–291.

Del Brio, E. B., Mora-Valencia, A., Perote, J., (2020). Risk quantification for commodity ETFs: Backtesting Value-at-Risk and Expected Shortfall. *International Review of Financial Analysis*, 70, 101163.

Ding, Z., Granger, C. W. J., Engle, R. F., (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1, pp. 83–106.

Doman, M., Doman, R., (2009). Modelowanie zmienności i ryzyka. Metody ekonometrii finansowej, Wolters Kluwer Polska, Kraków.

Dowd, K., (1999). Beyond Value at Risk: The New Science of Risk Management, John Wiley & Sons, Chichester.

Eling, M., (2014). Fitting asset returns to skewed distributions: Are the skew-normal and skew-student good models?. *Insurance: Mathematics and Economics*, 59, pp. 45–56.

Elsayed, A. H., Gozgor, G., Yarovaya, L., (2022). Volatility and return connectedness of cryptocurrency, gold, and uncertainty: Evidence from the cryptocurrency uncertainty indices. *Finance Research Letters*, 47, 102732.

Fałdziński, M., (2014). Teoría wartości ekstremalnych w ekonometrii finansowej. *Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika w Toruniu*, Toruń.

Fernandez-Perez, A., Frijns, B., Fuertes, A-M., Miffre J., (2018). The skewness of commodity futures returns. *Journal of Banking and Finance*, 86, pp. 143–158.

Fiszeder, P., Fałdziński, M., Molnár, P., (2019). Range-Based DCC Models for Covariance and Value-at-Risk Forecasting. *Journal of Empirical Finance*, 54, pp. 58–76

Gumbel, E. J., (2004). Statistics of Extremes. *Dover Publications*, Inc, Mineola, New York.

Hill, B. M., (1975). A simple general approach to inference about the tail of the distribution. *The Annals of Statistics*, 3(5), pp. 1163–1174.

Huang, C. K., Huang, C. S., Chikobvu, D., Chinhamu, K., (2015). Extreme risk, Value-at-Risk and Expected Shortfall in the gold market. *Int. Bus. Econ. Res. J.*, 14(1), pp. 91–107.

Jajuga, K., (2008). Zarządzanie ryzykiem. *Polskie Wydawnictwo Naukowe PWN*, Warszawa.

Jalal, A., Rockinger, M., (2018). Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. *Journal of Empirical Finance*, 15, pp. 868–877.

Jorion, P., (2001). Value-at-Risk, The New Benchmark for Managing Financial Risk, 2nd Edition, McGraw-Hill, New York.

Li, X., Guo Q., Liang Ch., Umar M., (2022). Forecasting gold volatility with geopolitical risk indices. *Research in International Business and Finance*, 64, 101857.

Kayla, P., Maheswaran, S., (2021). A study of excess volatility of gold and silver. *IIMB Management Review*, 33, pp. 133–145.

Kratz, M., Lok, Y. H., McNeil, A. J., (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking and Finance*, 88, pp. 393–407.

Lazar, E., Zhang, N., (2019). Model risk of expected shortfall. *Journal of Banking and Finance*, 105, pp. 74–93.

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

Mensi, V., Vinh, Vo X., Hoon Kang, S., (2022). COVID-19 pandemic's impact on intraday volatility spillover between oil, gold, and stock markets. *Economic Analysis and Policy*, 74, pp. 702–715.

Morales, L., Andreosso-O'Callaghan, B., (2011). Comparative analysis on the effects of the Asian and global financial crises on precious metal markets. *Research in International Business and Finance*, 25(2), pp. 203–227.

Naeem, M., Tiwari, A. K., Mubashra, S., Shahbaz, M., (2019). Modeling volatility of precious metals markets by using regime-switching GARCH models. *Resources Policy*, 64, 101497.

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

Trzpiot, G., (2004). O wybranych własnościach miar ryzyka. *Badania Operacyjne i Decyzje*, no. 3-4, pp. 91–98.

Walid, C., Shawkat, H., Khuong, N. D., (2014). Volatility forecasting and risk management for commodity markets in the presence of asymmetry and long memory. *Energy Econ.*, 41, pp. 1–18.

Włodarczyk, B., (2017). Prognozowanie zmienności stóp zwrotu na rynkach złota i srebra z uwzględnieniem efektu asymetrii i długiej pamięci. *Studia i Prace WNEiZ US*, no. 50, T. 1, pp. 231–247.

Vidal, A., Kristjanpoller, W., (2020). Gold volatility prediction using a CNN-LSTM approach. *Expert Systems with Applications*, 157, 113481.

Yu, W., Yang, K., Wei, Y., Lei, L., (2018). Measuring Value-at-Risk and Expected Shortfall of crude oil portfolio using extreme value theory and vine copula. *Physica A*, 490, pp. 1423–1433.

Zijing, Z., Zhang, H. K., (2016). The dynamics of precious metal markets VaR: A GARCH-EVT approach. *Journal of Commodity Markets*, 4, pp. 14–27.