Central European Economic Journal

sciendo

ISSN: 2543-6821 (online)

Journal homepage: http://ceej.wne.uw.edu.pl

Szymon Lis, Marcin Chlebus

Combining forecasts? Keep it simple

To cite this article

Lis, Sz., Chlebus, M. (2023). Combining forecasts? Keep it simple. Central European Economic Journal, 10(57), 343-370.

DOI: 10.2478/ceej-2023-0020

To link to this article: https://doi.org/10.2478/ceej-2023-0020

Central European Economic Journal

Szymon Lis
University of Warsaw, Faculty of Economic Sciences, 44/50 Długa street, 00-241 Warsaw, Poland
corresponding author: sm.lis@student.uw.edu.pl

Marcin Chlebus
University of Warsaw, Faculty of Economic Sciences, 44/50 Długa street, 00-241 Warsaw, Poland

Combining forecasts? Keep it simple

Abstract
This study contrasts GARCH models with diverse combined forecast techniques for Commodities Value at Risk (VaR) modeling, aiming to enhance accuracy and provide novel insights. Employing daily returns data from 2000 to 2020 for gold, silver, oil, gas, and copper, various combination methods are evaluated using the Model Confidence Set (MCS) procedure.
Results show individual models excel in forecasting VaR at a 0.975 confidence level, while combined methods outperform at 0.99 confidence. Especially during high uncertainty, as during COVID-19, combined forecasts prove more effective. Surprisingly, simple methods such as mean or lowest VaR yield optimal results, highlighting their efficacy.
This study contributes by offering a broad comparison of forecasting methods, covering a substantial period, and dissecting crisis and prosperity phases. This advances understanding in financial forecasting, benefiting both academia and practitioners.

Keywords
Machine learning | GARCH models | combined forecasts | commodities | VaR

JEL Codes
C53, G32, Q01

1. Introduction

The accuracy and reliability of models in predicting Value at Risk (VaR) are significantly contingent upon the quality and nature of the underlying data. While models exhibiting parsimony might demonstrate robust performance during periods of stable economic conditions, their efficacy could diminish when confronted with heightened market volatility (Angabini & Wasiuzzaman, 2011; Bhowmik & Wand, 2020). Conversely, extensively parameterised models could prove fitting during times of elevated volatility, but may not exhibit the same appropriateness during phases of market tranquillity (Laurent et al., 2012; Bhowmik & Wand, 2020). In terms of investigation of the selection of different VaR models for daily energy commodities returns there were a lot of articles using standalone models and some combinations of forecasts. Among others, Laporta et al. (2018) considered GARCH, EGARCH, GJR-GARCH, Generalised Autoregressive Score (GAS), Conditional Autoregressive Value at Risk (CAViaR), and a Dynamic Quantile Regression (DQR) model. They then pooled information from each model using a weighted average approach. The empirical analysis was conducted on seven energy commodities. The results showed that the quantile approach (i.e., the CAViaR and DQR models) outperforms all others for all series considered and that VaR aggregation yields better results. Also, Andreani et al. (2021) used a method that combines different types of data to analyse daily volatility and correlations among energy commodities. Then, they compared their approach to existing models that ignore the pandemic's impact. They find that their method is better in assessing the pandemic's effects on energy market interactions. However, to date, no solitary model or method has emerged as the preeminent choice within the realm of VaR forecasting, given the inherent complexities and multifaceted dynamics at play (Bernardi & Catania,

2016; Žiković et al., 2015; Bayer, 2018; Buczyński & Chlebus, 2018).

A potential avenue to address this challenge could involve the development of more intricate models tailored to closely align with prevailing economic conditions, or alternatively, the amalgamation of multiple forecasts. Substantial empirical evidence underscores the superiority of forecast combinations compared to individual models (Jeon & Taylor, 2013; Bayer, 2018; Taylor, 2020). Various studies highlight that across diverse assets, model typologies, and evaluation periods, fused forecasts yield heightened precision, positioning them within the 'green zone' criteria stipulated by Basel II regulations (Halbleib & Pohlmeier, 2012; Fameliti & Skintzi, 2020). Nevertheless, earlier scholarship challenges the notion of enhanced predictive accuracy through forecast amalgamation (Armstrong, 1989; Terui & Van Dijk, 2002).

Timmermann (2006) presents a compelling rationale for the fusion of forecasts to stabilise and enhance the predictive aptitude of standalone models. Primarily, the amalgamation of forecasts originating from varying model assumptions, specifications, or information sets offers distinct advantages. Additionally, these blended forecasts exhibit robustness in the face of structural disruptions. Lastly, the potential repercussions of model mis-specification are mitigated through the confluence of predictions derived from a multitude of models.

Numerous methodologies for amalgamating forecasts in the context of VaR have been explored (Giacomini & Komunjer, 2005; McAleer et al., 2010; Huang & Lee, 2013; Parot et al., 2019). However, the scholarly landscape lacks a comprehensive juxtaposition of forecast accuracy achieved through diverse combination techniques for VaR. This present study endeavours to address this gap by conducting a comparative analysis of VaR forecasts concerning commodity prices, including gold, silver, copper, oil, and gas. The investigation encompasses both standalone methods, namely GARCH, GARCH-t, GARCH-st, QML-GARCH, and Indirect GARCH (CaViaR), as well as fused forecasts achieved through several approaches.

Ghoddusi et al. (2019) provides a comprehensive review of the literature on the applications of machine learning (ML) in energy economics and finance. The authors critically review more than 130 articles published between 2005 and 2018 and identify applications of ML in areas such as predicting energy prices (e.g., crude oil, natural gas, and power), demand forecasting, risk management, trading strategies, data processing, and analysing macro/energy trends. The authors also discuss the achievements and limitations of existing literature and identify current gaps while offering some suggestions for future research.

Applied combined models also include the application of machine learning techniques that in financial risk management has emerged as a groundbreaking approach, introducing new dimensions of accuracy and adaptability (Aziz & Dowling, 2019; Mashrur et al., 2020). By leveraging advanced algorithms and data-driven methodologies, machine learning offers the capability to process vast amounts of financial data in real time, to uncover intricate patterns, and to generate forecasts with heightened precision (Rundo et al., 2019; Wasserbacher & Spindler, 2022). This fusion not only promises to address the limitations of individual models in varying market conditions but also aligns with the contemporary drive towards more sophisticated and data-driven risk assessment strategies.

The comparative evaluation is conducted across various temporal segments: an entire period spanning from mid-2004 to 2020, phases characterised by market tranquillity (July 2004 to 2006, 2009 to 2013, and 2016 to 2019), periods of market upheaval (2007 to 2008, 2014 to 2015, March 2020 to December 2020), and a distinct interval marked by the emergence of the coronavirus pandemic (March 2020 to December 2020). The primary objective of this study is to assess the accuracy of the aforementioned forecast methods within these delineated timeframes and under differing market conditions.

The choice to focus on commodity indices, such as gold, silver, copper, oil, and gas, warrants clarification. While other indices like the S&P 500 could be equally relevant, commodities hold distinct characteristics that make them valuable to study. Commodity markets often respond differently to economic and geopolitical factors (Stuermer & Valckx, 2021; Xiao et al., 2022), presenting unique challenges and opportunities for risk prediction. Thus, our investigation into VaR forecasting extends to these commodities to provide a comprehensive understanding of their specific dynamics.

According to the aim of the study, the following hypotheses were formulated:

Hypothesis 1. Over the entire period, forecast combining methods will be more accurate than individual methods. This hypothesis is based on the fact that individual methods often fail during three periods of crisis throughout the whole period.

Hypothesis 2. In the period of calm, forecast combining methods will prove to be more accurate than individual methods. This is based on the fact that combining methods can utilise the best features of each model by weighing them, and therefore produce more accurate results.

Hypothesis 3. In times of crisis, forecast combining methods will turn out to be more accurate than individual methods. The justification for this is similar to that of Hypothesis 2.

Hypothesis 4. During the period of data available for the current coronavirus pandemic, forecast combining methods will be more accurate than standalone VaR. As above.

It should be noted that in all hypotheses, the greater accuracy of the forecast combining methods is interpreted as the superiority of at least one of these methods, not all of them. As most of the methods are being used for the first time for combining VaR forecasts, hypotheses regarding the primacy of one of the combination methods were not stated.

The remainder of this paper is organised as follows. Section 2 introduces the methodology and provides details on individual methods, ways of combining forecast, and exploratory data analysis. Section 3 presents the results of the empirical application. Section 4 consists of a conclusion and an outlook on potential future research areas.

2. Methodology

2.1. Data

The data used in this study were futures prices for gold, silver, copper, oil, and gas obtained from Yahoo Finance (COMEX Gold futures [GC]. COMEX Silver futures [SI]. COMEX Copper futures [HG]. NYMEX WTI Crude Oil futures [CL]. NYMEX Gas futures [NG]). The data were collected for the period from 01/09/2000 to 01/12/2020, comprising a total of 6,215 records. Log returns were calculated for each commodity based on the adjusted price.

It is worth noting that the number of missing values for each financial instrument varied, ranging from 1079 to 1138. Most of the missing values are for Sundays, when there is no listing on the stock exchange. Therefore, observations for Sundays were removed. For the remaining missing values, the last available value was used to fill in the gaps. This approach has been shown to be as robust for time series data as other known data gap-filling methods (Caillault et al., 2017).

The study aims to assess VaR over a period of 4157 days from 5 July 2004 to 1 December 2020. The VaR is assessed using a one-day-ahead with 99th and 97.5th alpha levels. The model parameters are updated with each observation using a rolling window approach, where data from the last 1000 observations are used for estimation. This approach is commonly used and has been demonstrated to be robust for time series data (Caillault et al., 2017). The VaR assessment horizon is divided into two sub-periods – the calm period and the crisis period – to evaluate models for situations with different volatilities. The division of the assessment period is shown in Figure 1, which shows the logarithmic returns for each commodity.

The crisis period is marked by three dark grey areas, the first from 1 January 2007 to 31 December 2008, covering the 2007–2009 financial crisis, when a significant increase in commodities prices was observed due to investors' shift towards investing in commodities (Phillips & Yu, 2011). The second period of the crisis, from 1 January 2014 to 31 December 2015, was characterised by frequent drops in prices, mainly due to a surplus of supply in relation to demand, slowdown in the development of the world economy, and a boom on the stock market (Dudziński, 2016). The third period of the crisis, from 1 January 2020 to 1 December 2020, was marked by the outbreak of the COVID-19 pandemic, which significantly disrupted the dynamics of prices of all raw materials (Mensi et al., 2020).

The remaining time is represented by light grey areas, for example, from 5 July 2004 to 31 December 2006, from 1 January 2009 to 31 December 2013, and from 1 January 2016 to 31 December 2019. Unfortunately, for gas throughout the entire VaR testing period, the variability is consistently high, and for copper, it is stably low, which may result in a smaller benefit from the use of the proposed forecast combining methods.

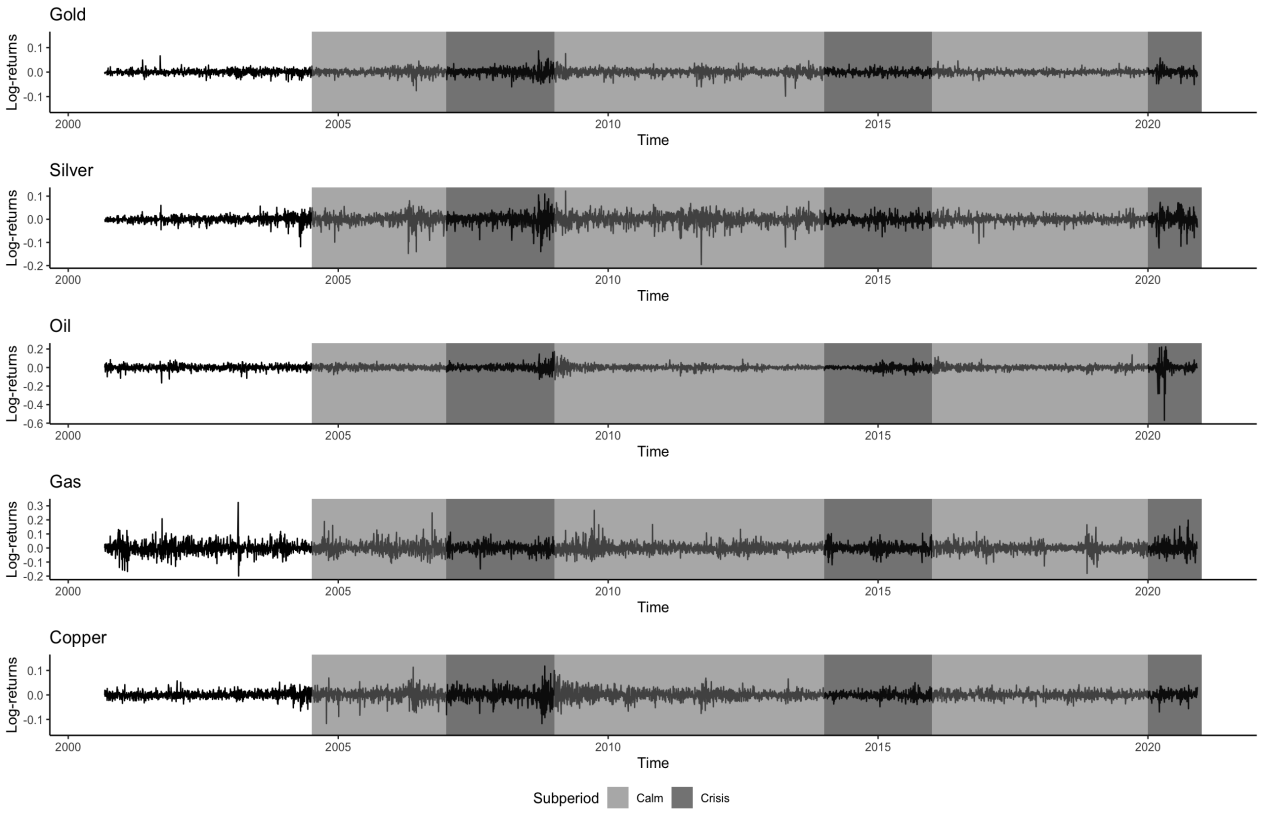


Figure 1. Returns for commodities (light grey for calm period, dark grey for crisis period)

2.2. Standalone models

In this research, we employed a variety of models to estimate the Value-at-Risk (VaR) of commodities. The selection of these models was driven by the need to address the limitations of assuming normality, which is often impractical for real-world financial scenarios, as highlighted by Holthausen & Hughes (1978), Szakmary et al. (2010), and Youssef et al. (2015).

Firstly, the GARCH process (Bollerslev, 1986) was employed. The GARCH process is rooted in the autoregressive nature of volatility, modeling volatility as a function of past squared returns. This model assumes constant conditional variance and is governed by two key components: the autoregressive parameter (ARCH term) and the moving average parameter (GARCH term). It has the advantage of being simple, but its limitation lies in its inability to capture more complex patterns in volatility, such as asymmetry or fat tails.

Next, we used the GARCH-t model (Bollerslev, 1987) which introduces the assumption that the errors follow a standardised Student t-distribution. This accommodates the presence of outliers and heavy tails in financial data, making it suitable for situations where extreme events are more likely. However, this flexibility comes at the cost of increased model complexity. The GARCH-st model further extends the GARCH-t model by considering a skewed t-Student distribution for the errors. This additional parameter captures skewness in the return distribution, making it well-suited for cases where volatility exhibits asymmetric behaviour. However, the model's complexity increases with the inclusion of the skewness parameter.

To further enrich our analysis, we introduced the Quasi-Maximum Likelihood GARCH (QML-GARCH) model based on the works of Bollerslev & Woolridge (1992) and Engle & Manganelli (2004). The QML-GARCH model estimates conditional variance using the GARCH process and derives VaR by quantiles of standardised residuals. This approach sidesteps the need to explicitly model the distribution of returns, offering simplicity. However, it assumes that standardised residuals are normally distributed, potentially limiting its performance in capturing non-normal features of financial data.

Lastly, we used the Conditional Autoregressive Value-at-Risk (CAViaR) model (Engle & Manganelli, 2004), which directly focuses on modeling quantiles of the distribution rather than the entire distribution itself. This approach is based on the notion that volatility clusters in financial data. It proves beneficial when capturing the inherent autocorrelation in volatility, but it disregards modeling the entire return distribution.

Each model exhibits specific strengths and limitations. The GARCH process might struggle to capture more intricate volatility patterns, while the GARCH-t and GARCH-st models could face challenges with increased complexity. The QML-GARCH model's reliance on normality assumptions could hinder its performance in cases of non-normality. The CAViaR model, while accounting for volatility autocorrelation, doesn't address the entire return distribution. It is crucial to acknowledge these limitations when selecting a model for a given dataset.

Combining the forecasts of various GARCH models holds significant value due to the distinct strengths and limitations of each individual model. By merging models' forecasts, we leverage the diverse insights they provide, leading to enhanced accuracy in VaR estimates. This approach addresses the dynamic and uncertain nature of financial markets while mitigating the shortcomings inherent in any single model. Ultimately, forecast combination empowers more robust risk management by delivering comprehensive and adaptable VaR assessments that align with complex market dynamics.

In order to select the optimal parameters for each model, we used the Akaike Information Criterion (AIC), which is commonly used in selecting models (Tsay, 2005). All models met the following conditions: (1) the sum of the parameters was lower than 1, (2) all of the parameters were statistically significant, (3) the Ljung-Box test on standardised squared residuals indicated that standardised residuals were white noise (ARCH effect removal), and (4) the p-value of the LM ARCH test indicated no ARCH effects among the residuals of the model. If autocorrelation was observed in the residuals, we added the Autoregressive Moving Average (ARMA) process to the GARCH, creating the ARMA-GARCH and eliminating the autocorrelation.

2.3. Combined forecasts

2.3.1. Mean of forecast

In previous studies, researchers have often utilised the simple mean of all forecasts obtained to combine forecasts, due to its simplicity. This approach has performed well in VaR forecasting (Clemen & Winkler, 1986; Timmermann, 2006; Halbleib & Pohlmeier, 2012; Huang & Lee, 2013). However, if the values are correlated, then combining them with the mean captures the same information again, resulting in a mere bias-variance trade-off (Hastie, Tibshirani, & Friedman, 2011, p. 223). Consequently, an increase in the variance increases the expected square error of the prediction. It is, therefore, better to consider the correlations between the forecasts and select the least correlated ones. Thus, the model with the best in-sample backtesting results is combined with the forecast from the model with the lowest correlation. For this method two VaR models are chosen by assessing the mean performance of the best model determined through the MCS procedure during a specified period. Additionally, the model with the least correlation to this best-performing model is selected (for details see section 3.1).

2.3.2. The lowest VaR

This method offers the advantage of always selecting the most conservative forecast, which may be particularly useful during crisis periods. However, a potential disadvantage of this approach is the possibility of consistently overestimating VaR, even during calm periods, which is not desirable. Previous studies by McAleer et al. (2010) and Buczyński & Chlebus (2019) have successfully applied this method. This model forecasts from all models that were considered as input.

2.3.3. The highest VaR

This method is useful when all measures turn out to be overly conservative regardless of time. In such cases, using the most liberal measure could be a good solution. However, it also poses a risk of constant underestimation of VaR. This approach has been previously employed by McAleer et al. (2010) and Buczyński & Chlebus (2019). For this model forecasts from all models were considered as input.

2.3.4. Conditional quantile optimisation method

This method models a conditional p-order quantile using a linear combination of two known quartiles determined using individual methods (Giacomini & Komunjer, 2005):

$$Q_p(r_{T+s}) = \beta_{T,0} + \beta * VaR^1_{T+s} + (1 - \beta_{T,1}) * \\ * VaR^2_{T+s} \equiv VaR_{12,T+s} * \beta_T = 1,2,\dots,S \quad (1)$$

Where: $VaR^1_{2,T+s}=(1,VaR^1_{T+s}, VaR^2_{T+s})$ a vector of VaR predicted from standalone models, $\beta_T=(\beta_{T,0}, \beta_{T,1}),(1-\beta_{T,1}))$ a vector of parameters for VaR predicted from standalone models

The vector of weights λ_T is determined by solving the following minimisation problem:

$$\hat{\lambda}_T = \underset{\lambda_T}{\operatorname{argmin}} \begin{matrix}\{\sum_{r_T \geq VaR_{12,T}*\lambda_T} p * |r_T - VaR_{12,T} * \beta_T| + \\ \sum_{r_T < VaR_{12,T}*\lambda_T}(1-p) * |r_T - VaR_{12,T} * \beta_T|\}\end{matrix} \quad (2)$$

Where r_T – returns for a particular commodity in time T.

The main advantage of the quantile regression approach is that it does not require explicit distribution assumptions for return data. The same forecasts as for the average were used here. For this method two models were chosen, as described in section 2.3.1.

2.3.5. Penalised quantile regression – LASSO

The least shrinkage and selection operator (LASSO) penalty is a popular regularisation method proposed by Tibshirani (1996) for variable selection. The LASSO method retains the advantages of best subset selection by providing a sparse solution, ensuring model stability, and providing objective estimates for large coefficients (Fan & Li, 2001). It encourages sparsity by shrinking some regression coefficients to exactly zero, leading to a simpler and more interpretable model. This characteristic is particularly valuable in situations where there are many predictors, and only a subset of them are truly relevant. LASSO's ability to perform variable selection aids in identifying the most important features for the given task, which can lead to improved model generalisation and performance. Moreover, LASSO's stability and objective estimates for large coefficients contribute to its efficacy in statistical modeling. For this method two models were chosen as described in section 2.3.1.

2.3.6. Penalised quantile regression – elastic net (EN)

The elastic net penalty, proposed by Bayer (2017), is a hybrid approach that combines the LASSO and ridge penalties. This combination strikes a balance between individual variable selection (as in LASSO) and grouping correlated variables together (as in ridge). The EN can be especially effective when dealing with multicollinearity among predictors. By setting some coefficients to zero and shrinking others, the elastic net retains the benefits of both LASSO and ridge, making it a versatile choice for various scenarios. The parameter that controls the trade-off between LASSO and ridge penalties can be tuned to achieve the desired level of sparsity and regularisation. The parameter that balances ridge and LASSO penalties was set to 0.5. For this method two models were chosen as described in section 2.3.1.

2.3.7. Quantile Random Forest (QRF)

Meinshausen & Ridgeway (2006) introduced the QRF, an extension of random forests, to model conditional quantiles. Empirical evidence suggests that its predictive power is competitive (Andreani et al., 2022). In this study, the influence is calculated by permuting the out-of-bag data and logging a forecast mean squared error for each tree. The same process is repeated after permuting each predictor variable, and the difference is averaged over all trees and normalised by the standard deviation of the differences (Grömping, 2009). QRF's ability to model conditional quantiles makes it well-suited for situations where capturing different parts of the response distribution is crucial, as opposed to just estimating the mean. For this model forecasts from all models were considered as input. The following hyperparameters were applied:

- Number of trees (the number of trees to grow) was set to 500.
- Number of randomly sampled predictors was set to the square root of the total number of predictors. This hyperparameter influenced the diversity of trees in the forest.
- Minimum observations for split attempt was 5 observations. Adjusting this parameter allowed for exploration of trade-offs between tree complexity and predictive accuracy.

- Minimum observations in terminal node: The granularity of the resulting tree structure was controlled by setting a minimum of 1 observation in terminal nodes.
- Permutations for permutation importance: Three permutations were used for permutation importance calculations. This technique helps identify predictor variables that contribute significantly to the model's performance.

2.3.8. Generalised Boosted Regression (GBM) Model

GBM proposed by Friedman (2001) is a powerful tool for forecasting quantile distributions. The algorithm adds a new decision tree, referred to as a 'weak learner', at each iteration to best minimise the loss function. Iteration continues until the maximum number of iterations specified by the user is reached. One of the greatest practical advantages of using the GBM model is its flexibility and accuracy in forecasting. However, determining the influence of individual dependent variables on the final result in this method is challenging. To estimate the influence of inputs, we utilised the relative influence method by Friedman (2001). The GBM model can be effectively used in combining forecasts due to its inherent ability to handle ensemble learning and sequential model building. Combining forecasts from different models or sources can often lead to improved predictive accuracy and robustness, and the GBM model is well-suited for this purpose. The following hyperparameters were applied:

- The number of boosting iterations equal to 100. This hyperparameter refers to the number of boosting iterations or the number of decision trees that will be created in the ensemble.
- The maximum tree depth in the ensemble equal to 1. This parameter controls the maximum depth of an individual decision tree within the ensemble.
- The minimum observations in a terminal node equal to 10. It sets the minimum number of observations required in a terminal (leaf) node of a decision tree.
- The learning rate or shrinkage equal to 0.1. Also known as the learning rate, this parameter scales the contribution of each tree in the ensemble.
- The fraction of data for bagging equal to 0.5. This determines the proportion of the training data used for building each individual tree in the ensemble.
- The fraction of data used for training equal to 1. This hyperparameter specifies the fraction of the dataset used for training.
- The number of cross-validation folds (0 for no cross-validation) equal to 0. The number of cross-validation folds used during model training.
- Verbose is a Boolean operator indicating whether to print progress equal to false. This determines whether the model's training progress is printed during training.
- The number of CPU cores to be utilised equal to NULL. The number of CPU cores utilised during model training.

2.3.9. Quantile Regression Neural Network (QRNN)

The QRNN model (Cannon, 2010) has been proposed as a promising alternative to parametric ANN models for modeling extreme events (e.g., Pradeepkumar & Ravi, 2017). Method of calculation is presented in Figure 2.

First, output from the j-th hidden-layer node $g_j(t)$ is given by applying the hyperbolic tangent, a sigmoidal transfer function, to the inner product between $x_i(t)$ and the hidden-layer weights $w_{ij}^{(h)}$ plus the hidden-layer bias $b_j^{(h)}$. An estimate of the conditional quantile is then given by applying sigmoid transfer function to are the output-layer weights, $w_j^{(o)}$, and is the output-layer bias, $b^{(o)}$. One of the main advantages of the model is its ability to estimate the conditional quantiles of the response variable. This makes it valuable in scenarios where understanding extreme outcomes is crucial, such as in risk management. The model employs a combination of hidden layers and transfer functions to capture complex relationships in the data. However, the nonlinearity of the model can make interpreting the results more challenging than with linear models. One example is that James (2000) used daily exchange rates, comparing to GARCH-based quantile estimates. The results suggested that the QRRN offers a useful alternative for GARCH quantile forecasts. For QRNN the following hyperparameters were used:

- The parameter 'tau' represents the predicted quantile level. It determines the quantile of the predicted distribution that the model aims to

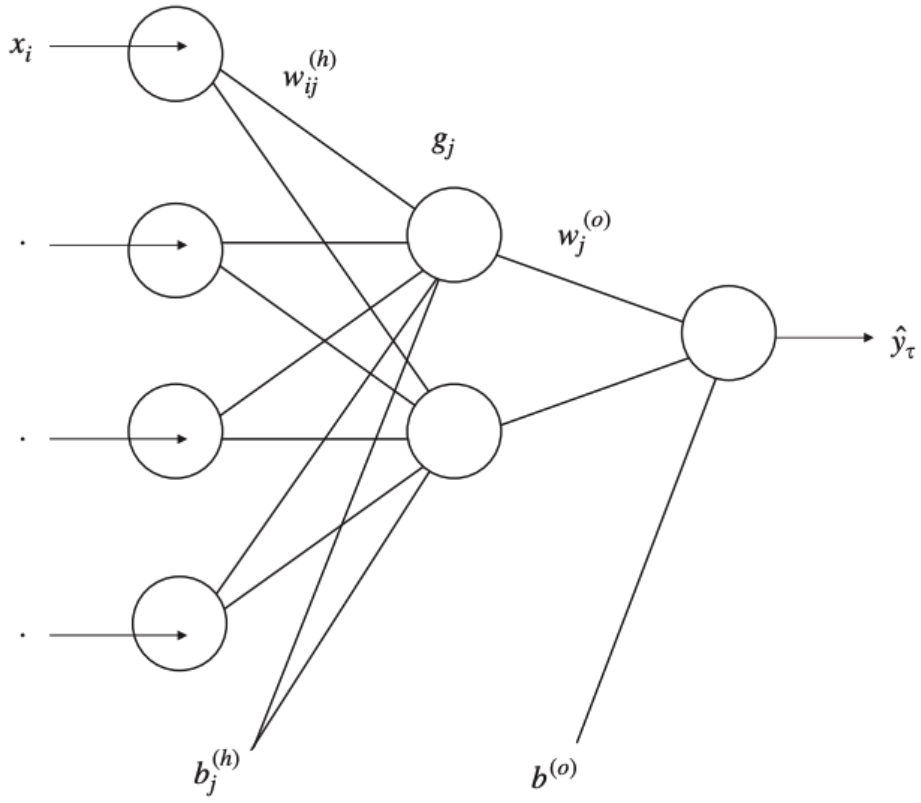


Figure 2. The diagram of QRNN model with four predictors and two hidden nodes
Source: Cannon, 2011

estimate. A chosen value of 0.5 corresponds to the median prediction.

- The 'iter.max' parameter equal to 5000 defines the maximum number of iterations that the training process of the QRNN model will undergo. It limits the number of optimisation steps taken during training to enhance convergence and prevent overfitting.
- The number of 'trial equal to' runs conducted for model fitting. Multiple trial runs can offer a more comprehensive evaluation of the model's performance and provide a more stable estimate of its capabilities.
- The false 'bag' parameter determines whether bagging (Bootstrap Aggregating) is utilised during training.
- The 'lower' equal-to-infinity parameter sets the lower bound for predictions made by the model. This parameter ensures that the model's predictions do not fall below a certain threshold, which can be crucial when dealing with certain types of data.
- The 'init.range' parameter equal to (-0.5, 0.5, -0.5, 0.5) represents the range within which the model's weights are initialised. Proper weight initialisation is essential to ensure that the model starts with reasonable parameters, aiding faster convergence and better performance.
- The 'monotone' parameter equal to NULL introduces monotonicity constraints on predictors.
- The 'additive' parameter equal to false enforces additive constraints on the model's predictions.
- The 'eps.seq' (equal to a decreasing sequence of numbers. It starts with 2 raised to the power of -8 and then continues with each subsequent number being smaller. The difference between each number is 2 raised to the power of -4.) defines a sequence of epsilon values used for ramp functions. Ramp functions are important in QRNN models to introduce non-linearity. The specified sequence controls the smoothness of the ramp functions.
- The parameters 'Th' (equal to sigmoid function) and 'Th.prime' are the default activation and derivative functions used within the QRNN model. These functions play a pivotal role in shaping the

behaviour of the model's hidden units, determining how information flows through the network.

Described above hyperparameters were not optimised due to the calculation constraint and the willingness to present only the results for further development of this field.

The same as for standalone models for combining models, parameters were updated for each observation using a rolling window approach, where data from the last 1000 observations are used for estimation (Caillault et al., 2017).

The output for all models is the forecast prediction and for some of them (e.g., regressions) also weights, while for other the impact of particular prediction is assessed in another way (see section 3).

In summary, each of these regression models brings unique strengths to the table. LASSO and elastic net aid in variable selection and regularisation, QRF excels in capturing conditional quantiles, GBM offers flexibility and high accuracy, and QRNN is tailored for modeling extreme events and quantiles. The choice of model depends on the specific characteristics of the data and the goals of the analysis.

2.4. Backtesting

Backtesting is a crucial step in assessing the accuracy and reliability of financial models, particularly those used for risk measurement like VaR. It involves comparing the predicted outcomes of a model with the actual outcomes that occurred in the real world during a specific time period. In this context, the goal of backtesting is to evaluate how well the different forecasting models perform in estimating VaR.

2.4.1. Excess Ratio (ER) test

The Excess Ratio test is a backtesting technique that compares the proportion of times the actual loss exceeds the VaR estimate to the expected proportion. If the model is accurate, the actual losses exceeding VaR should be roughly in line with the expected proportion.

2.4.2. Kupiec test (UC)

The Kupiec test, proposed by Kupiec in 1995, is another backtesting method used to evaluate the accuracy of VaR forecasts. It focuses on the number of exceptions – instances where actual losses exceed the VaR estimate. The test assesses whether the number of exceptions matches the expected number based on the chosen confidence level.

2.4.3. Christoffersen test (CC)

The Christoffersen test, introduced by Christoffersen in 1998, is a more comprehensive backtesting approach that considers both the frequency of exceptions and the magnitude of excess losses. It takes into account the entire distribution of forecast errors to assess the model's performance.

2.4.4. Dynamic Quantile test (DQ)

The Dynamic Quantile test, proposed by Engle & Manganelli in 2004, is designed to evaluate the conditional coverage property of VaR models. It assesses whether the VaR estimates are able to capture the changing volatility and risk in different market conditions.

2.4.5. Traffic Light Test (TL)

The Traffic Light test, introduced by the Basel Committee on Banking Supervision (BCBS) in 1996, is a simplified backtesting method that categorises model performance based on whether the actual loss is above or below the VaR estimate. It uses a traffic light system to indicate whether the model's performance is 'green', 'yellow', or 'red'.

2.4.6. Model Confidence Set Procedure (MCS)

The Model Confidence Set Procedure, proposed by Hansen et al. in 2011, is a statistical technique used to compare and rank multiple models' forecasting accuracy. It's particularly useful when you have several competing models, as is the case in your context of combining forecasts for VaR. The procedure constructs a set of models that are likely to have high out-of-sample forecast accuracy. It helps determine which models are more likely to provide the best forecasts in the future, based on their historical performance. This method is often used to select the best models for forecasting (Laporta et al., 2018).

In summary, the backtesting methods and the Model Confidence Set Procedure provide a comprehensive framework to assess the performance of different VaR forecasting models. They allow one to evaluate how well these models capture the actual risk and losses in financial markets and aid in the selection of the most reliable forecasting approach.

3. Results

3.1. Data analysis

To investigate the characteristics of returns for commodities, we computed basic statistics. Table 1 presents the minimum and maximum values, skewness, kurtosis, and quantiles of daily logarithmic rate of return along with Jarque-Bera's test value and its p-value (in parentheses).

As shown in Table 1, none of the commodities follow a normal distribution, and all have leptokurtic distributions (excess kurtosis far above 0). Additionally, the distributions are left-skewed for oil, gold, silver, and copper, and right-skewed for gas. This result is in line with the findings of previous studies (Tse, 2016) on gas. Oil has the highest kurtosis and skewness values, which can be attributed to the consistent rises in oil prices over the last 20 years, with sharp declines during bad economic times, a trend that was also observed for gold and silver.

Using the MCS procedure, the following models were obtained:

- Gold – GARCH(1,1), AR(1)-GARCH-t(1,1), AR(1)-GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Silver – GARCH(1,2), AR(1)-GARCH-t(1,1), AR(1)-GARCH-st(1,1), AR(1)-QML-GARCH(1,1), Indirect GARCH(1,1),
- Oil – GARCH(1,1), GARCH-t(1,1), GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Gas – GARCH(1,1), GARCH-t(1,1), GARCH-st(1,1), QML-GARCH(1,1), Indirect GARCH(1,1),
- Copper – GARCH(1,1), GARCH-t(1,2), GARCH-st(1,1), QM-GARCH(1,1), Indirect GARCH(1,1).

In Figures 3 and 4, the relationship between forecast outcomes from distinct models is illustrated. Nevertheless, relying solely on correlation as the basis for selecting the most optimal forecast can result in suboptimal outcomes. To address this concern, we undertook a comprehensive approach. Specifically, we computed the average performance of the most proficient model utilising the MCS procedure during the in-sample period spanning from September 1, 2000, to July 2, 2004. Subsequently, we determined the VaR forecast that exhibited the least correlation with this model. The preeminent average results for each commodity are as follows:

- For the gold market, the GARCH and CaViaR models yielded the best outcomes for both p-values, attaining a correlation of 0.93 for a p-value of 0.025 and 0.81 for a p-value of 0.01.
- In the silver market, employing the GARCH-st + CaViaR models produced the optimal results, achieving a correlation of 0.87 for a p-value of 0.025 and a correlation of 0.8 for a p-value of 0.01.
- Within the gas market, the GARCH-t + CaViaR models exhibited superior performance for both

Table 1. Statistics of prices' log-returns

Commodity	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	J-B test	Skewness	Ex. Kurtosis
Gold	-0.0982	-0.0049	0.0005	0.0004	0.006	0.0864	6398 (<0.001)	–0.2658	8.7160
Silver	-0.1955	-0.0080	0.0011	0.0003	0.0090	0.1220	13942 (<0.001)	–0.9263	10.8079
Oil	-0.2799	-0.0128	0.0008	0.0001	0.0130	0.3196	52559 (<0.001)	–1.9164	52.6291
Gas	-0.1990	-0.1911	-0.0007	-0.0001	0.0173	0.3238	6833 (<0.001)	0.5643	8.7537
Copper	-0.1169	-0.0082	0.0002	0.0003	0.0089	0.1177	4279 (<0.001)	–0.1731	7.6239

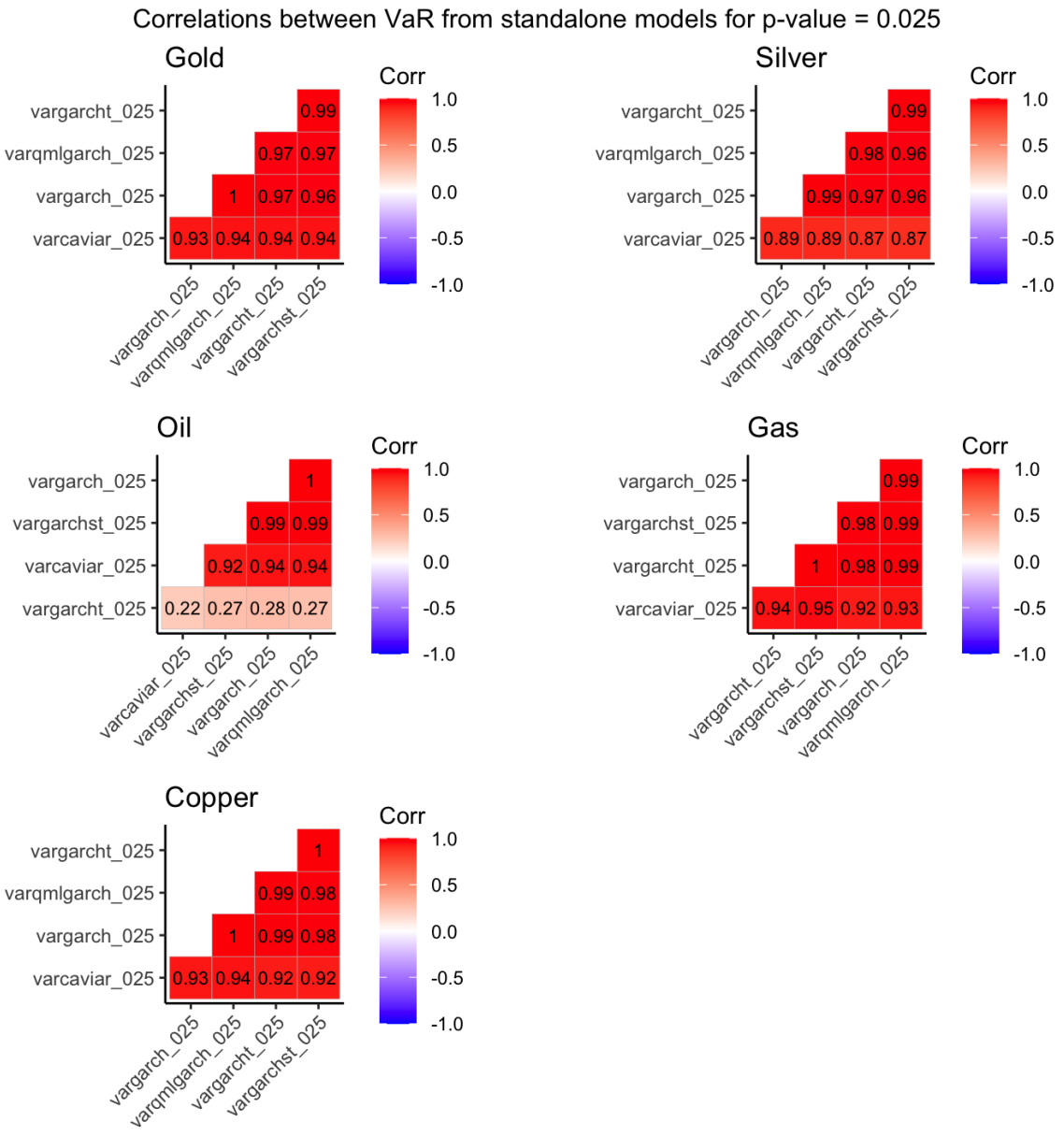


Figure 3. Correlations between VaR forecasts for p-value = 0.025

p-values, yielding a correlation of 0.94 for a p-value of 0.025 and 0.89 for a p-value of 0.01.

- Similarly, for the oil market, the GARCH-t + CaViaR models were most effective, resulting in a correlation of 0.22 for a p-value of 0.025 and 0.27 for a p-value of 0.01.
- Finally, in the copper market, the GARCH-st + CaViaR models demonstrated the highest proficiency for both p-values, yielding a correlation of 0.93 for a p-value of 0.025 and 0.9 for a p-value of 0.01.

3.2. Empirical results for individual and combined methods

The analysis of the models began with an evaluation of the visual comparison between the predicted VaR values from different models and the actual observed returns. Figure 5 presents a visual representation of log-returns and out-of-sample Value at Risk (VaR) sequences from specific models. Among these models, the CaViaR model stands out as the most cautious estimator for all considered assets. In situations where other models produce relatively higher VaR values, the CaViaR model consistently generates substantially lower estimates. This discrepancy is particularly

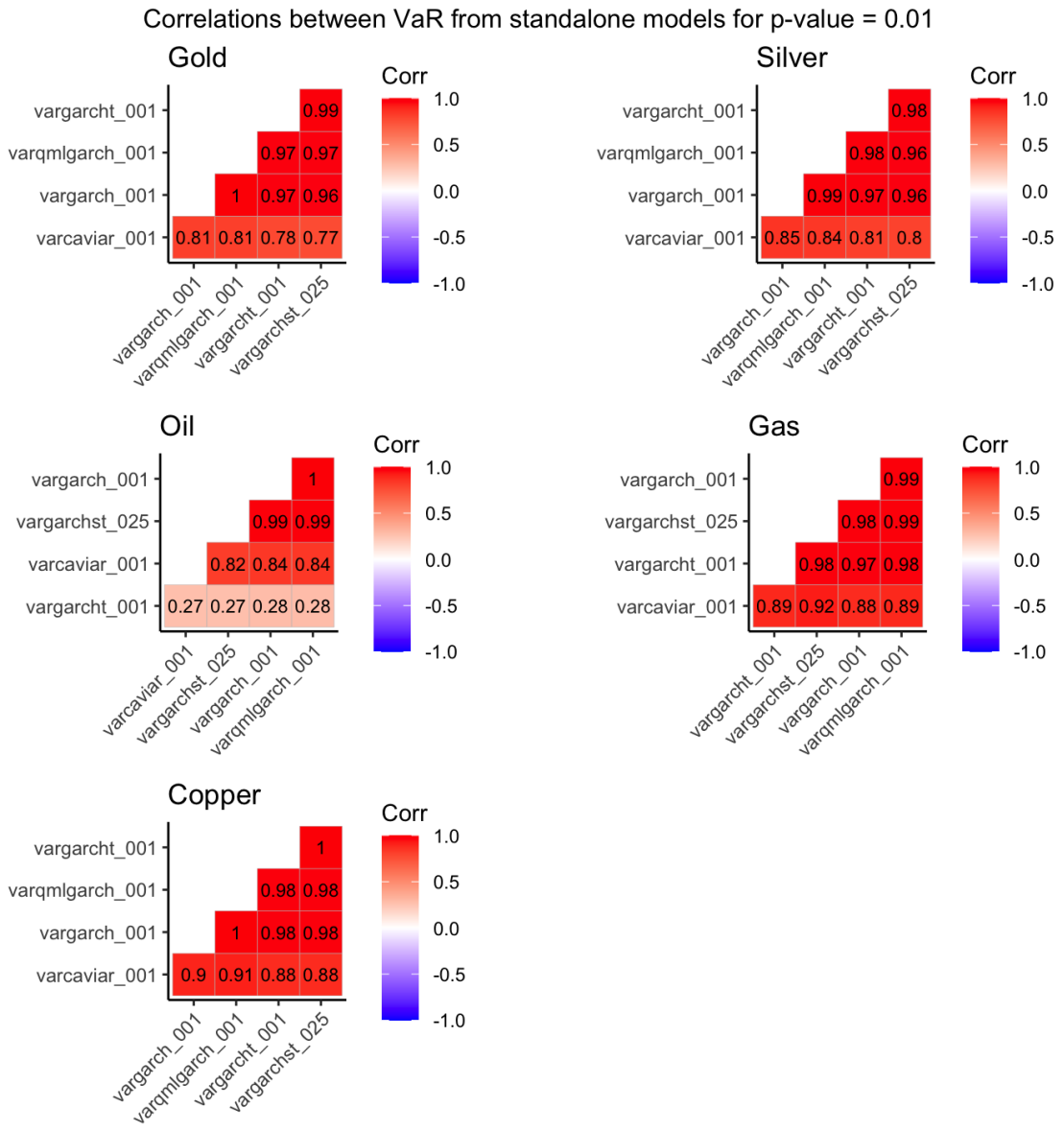


Figure 4. Correlations between VaR forecasts for p-value = 0.01

noticeable in cases such as gold or silver at a 0.99 confidence level.

This tendency can likely be attributed to the fact that, across all models, the mean values fall within the range of CaViaR's estimates. Consequently, the VaR estimates from the CaViaR model consistently position themselves between the CaViaR estimates and those of the alternative models.

Moreover, when applying quantile regression with elastic net regularisation, the resulting VaR forecasts appear flattened across the entire time span. This introduces the possibility of multiple instances where the VaR is exceeded.

Distinctly, the approach of gradient boosting quantile regression exhibits the most lenient VaR predictions (excluding the highest VaR instances). Conversely, the conditional quantile optimisation method assumes an intermediate position. However, an interesting observation emerges for oil and copper assets at a 0.99 confidence level, where this method tends to sustain lower VaR levels for extended durations.

Tables 2 through 11 herein present the outcomes derived from the backtesting process encompassing distinct assets, alongside various methodologies, at two discrete confidence levels, specifically 0.975 and 0.99. The tabulated information encompasses the findings

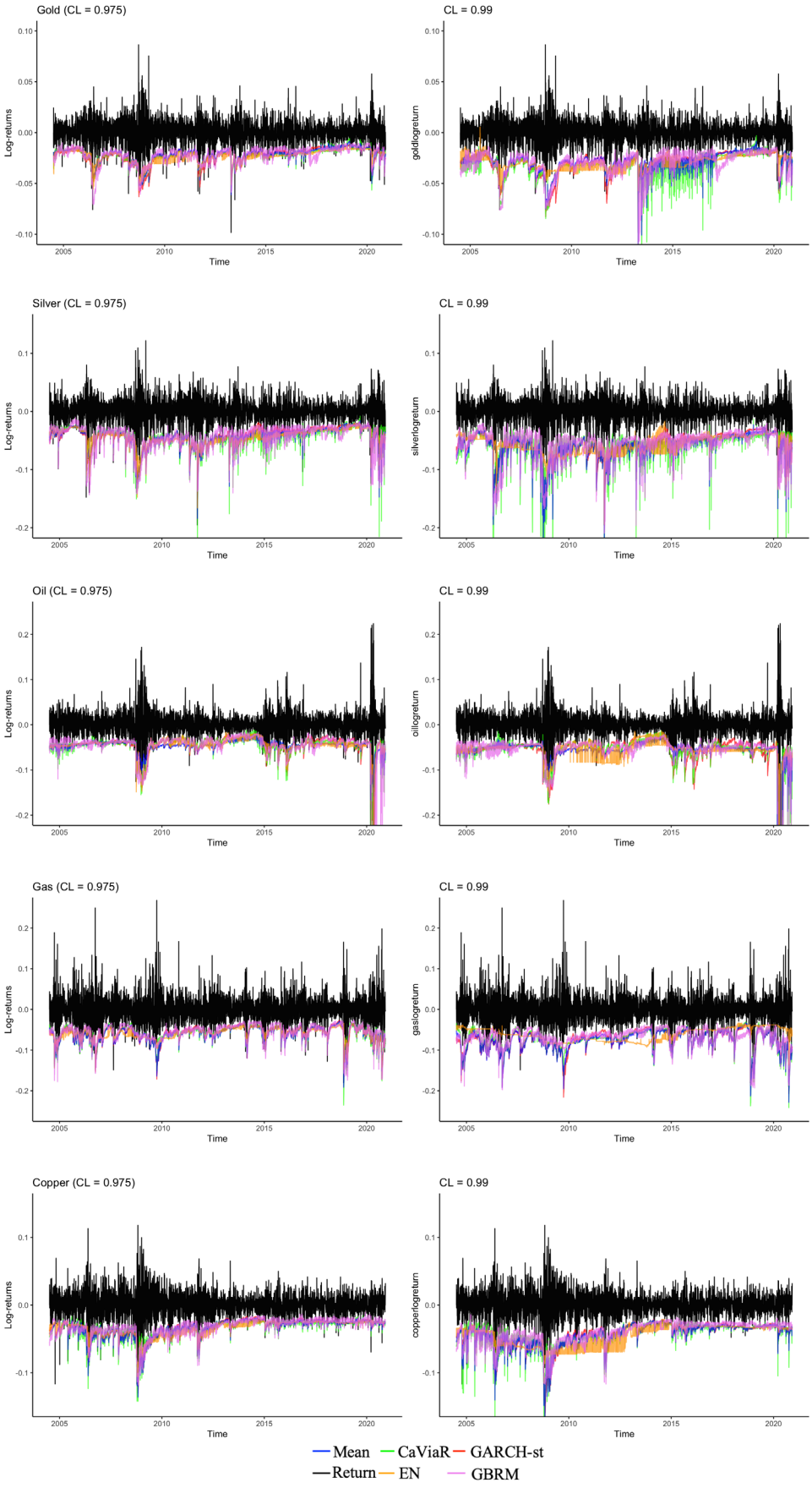


Figure 5. Returns and VaR forecast for confidence levels: 0.975 (on left) and 0.99 (on right)

yielded by the ER, UC, CC, DQ, and TL tests. Notably, the results are organised with a delineation across four distinct temporal segments: the comprehensive evaluation span, the phase subsequent to crises, the crisis-interval proper, and the epoch of the ongoing coronavirus crisis.

Across the entire evaluation timeframe, it is discerned that for all commodities, there exists at minimum one individual methodology that exhibits satisfactory performance. This is indicated by the observation that at least one of the applied tests demonstrates a p-value surpassing 0.05, concomitant with a favourable outcome in the TL test. Particularly noteworthy is the consistent adherence to these conditions by the GARCH-st model, which consistently manifests compliance across all assets throughout the entirety of the assessment period.

Nevertheless, a singular instance of non-conformity is observed with the CaViaR model. This model, while predominantly meeting the stipulated criteria, exhibits deviation in a solitary scenario – specifically, in the case of oil under the 0.99 confidence level. Furthermore, the GARCH-t model garners commendable results for the gold, gas, and copper assets. The standard GARCH model, conversely, aligns well solely with the oil asset at a confidence level of 0.975, and with the gas asset under both confidence levels. It is also noteworthy that the QML-GARCH model exclusively demonstrates proficiency in the context of the gas asset at the prescribed confidence levels.

The outcomes derived from the amalgamated models exhibit a certain degree of complexity. In relation to gold, at a confidence level of 0.975, the majority of forecast aggregation techniques exhibit encouraging outcomes. It is worth noting that the highest Value at Risk (VaR), random forests, and neural networks present shortcomings in this context. Conversely, at a confidence level of 0.99, solely the lowest VaR conforms to the criteria outlined in the preceding paragraph. When considering silver, both at the aforementioned confidence levels, the simple mean fulfils all stipulated conditions. Furthermore, at a confidence level of 0.975, the lowest VaR, CQOM, and two regression techniques (elastic net and lasso) yield favourable results.

For the commodity oil, accurate forecasts are provided by the mean at a confidence level of 0.975, while no forecast aggregation approach proves promising at the 0.99 confidence level. In the case of gas, at a confidence level of 0.975, precise outcomes are offered by the mean, highest VaR, lowest VaR, elastic net, and lasso. Conversely, at a confidence level of 0.99, solely the mean and lowest VaR demonstrate accurate forecasting. Across both confidence levels, robust results are furnished by the average, lowest VaR, and elastic net. Moreover, at a confidence level of 0.975, lasso also performs effectively.

Summarising the evaluation conducted over the entire assessment period, the GARCH-st or CaViaR model emerges as the most suitable among the individual models. Meanwhile, the mean and regression-based combined models consistently exhibit superior performance in the role of forecast aggregators, albeit with variations contingent upon the specific commodity under consideration. During periods characterised by stability, numerous models – both individual and combined – display accuracy in their forecasts. Notably, GARCH falters for silver at both confidence levels and for gold at a confidence level of 0.99, while GARCH-t falls short for silver at a confidence level of 0.975.

Among the forecast combining models, random forests and neural networks universally fall short in delivering accurate results for all commodities, regardless of the confidence level. Similarly, the highest VaR and CQOM also fail to meet expectations. Optimal results are consistently obtained from the mean, lowest VaR, and two regression methodologies, which consistently deliver accurate forecasts across confidence levels and commodities. Once more, the mean and regression-based aggregation models stand out for their promising performance.

In periods of crisis, it becomes evident that standalone methods are ill-equipped to accurately forecast VaR. This deficiency is observed across various assets including gold, silver, oil, gas, and copper. For example, for gold, GARCH-st and CaViaR demonstrate accurate VaR forecasts at a confidence level of 0.975, while GARCH-t excels at a confidence level of 0.99. In contrast, no individual models yield satisfactory outcomes for silver, as indicated by excess ratios exceeding 3.3% at a confidence level of 0.975 and 1.53% at 0.99. GARCH-st remains the sole effective method for VaR forecasting in the case of oil, specifically at a confidence level of 0.99.

For gas, models such as GARCH, GARCH-t, and QML-GARCH exhibit accuracy at a confidence level of 0.975, while GARCH-t alone proves effective at the 0.99 confidence level. Similarly, with regard to copper,

no models yield satisfactory results at a confidence level of 0.975, and CaViaR stands out as the most effective technique at a confidence level of 0.99.

During the COVID-19 pandemic crisis, no individual model demonstrates effectiveness in accurately forecasting VaR for gold, silver, and oil, irrespective of the confidence level. Excess ratios approximate 5% at a confidence level of 0.975 and 2.5% at 0.99 for these assets. However, for gas, all individual methods effectively forecast VaR at the 0.975 confidence level, while GARCH-t, GARCH-st, and CaViaR prove effective at the 0.99 confidence level. For copper, GARCH-t effectively forecasts VaR at a confidence level of 0.975, whereas no individual model performs adequately at a confidence level of 0.99, with excess ratios exceeding 2.16%.

Among the forecast aggregation methods, employing the lowest VaR consistently proves effective, except for gold at a confidence level of 0.99, silver at both confidence levels, and copper at a confidence level of 0.99. Furthermore, the Combined Quantile Opinion Mining (CQOM) approach demonstrates effectiveness for gold and gas at a confidence level of 0.975. Other methodologies such as the mean, elastic net, and lasso prove effective solely for gas VaR forecasts at a confidence level of 0.975, while the mean remains effective at the 0.99 confidence level.

Due to the absence of a clearly predominant model, both in the context of individual models and combined methodologies, a decision was made to implement the Model Selection Criterion (MSC) procedure. The outcomes of this procedure are detailed in Table

Table 2. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gold for confidence level equal to 0.975

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	3.10%	5.77	6.02	21.07	Y	2.85%	1.38	1.55	9.22	G	3.70%	6.45	6.50	16.92	Y	5.63%	6.88	6.98	26.48	Y
GARCH-t	2.81%	1.62	1.77	18.08	G	2.57%	0.06	0.07	2.48	G	3.38%	3.57	3.80	28.92	Y	4.76%	3.85	4.23	21.16	Y
GARCH-st	2.38%	0.24	0.41	16.80	G	2.20%	1.16	1.29	5.88	G	2.82%	0.50	1.36	21.53	G	4.33%	2.61	3.22	22.78	Y
QML-GARCH	3.17%	5.91	6.02	22.59	Y	2.88%	1.66	1.80	9.14	G	3.62%	5.66	5.74	19.26	Y	5.59%	6.80	6.98	26.36	Y
CaViaR	2.65%	0.36	0.74	20.85	G	2.47%	0.01	0.04	15.69	G	3.06%	1.49	2.02	14.36	G	5.19%	5.28	7.53	31.81	Y
Mean	2.81%	1.62	1.77	14.83	G	2.64%	0.24	0.24	9.74	G	3.22%	2.43	2.79	11.04	Y	4.76%	3.85	4.23	15.45	Y
Highest VaR	3.46%	14.18	14.38	41.56	R	3.19%	5.25	5.25	15.70	Y	4.11%	11.05	11.42	33.81	Y	6.49%	10.57	11.54	41.72	Y
Lowest VaR	2.07%	3.37	3.39	10.49	G	1.92%	4.35	4.36	8.59	G	2.42%	0.04	0.14	7.87	G	3.46%	0.79	2.04	7.83	G
CQOM	2.69%	0.63	9.01	56.89	G	2.47%	0.01	4.21	39.29	G	3.25%	2.48	6.46	44.63	Y	3.90%	1.58	2.48	11.99	G
Elastic Net	2.50%	0.00	1.85	24.77	G	2.20%	1.16	1.39	7.99	G	3.27%	2.49	4.23	30.53	Y	6.49%	10.57	11.54	81.74	Y
LASSO	2.62%	0.25	6.95	30.04	G	2.26%	0.69	6.14	14.90	G	3.46%	4.22	5.51	44.51	Y	7.36%	14.83	15.27	91.46	R
QRF	5.15%	92.01	92.12	246.98	R	4.87%	52.90	52.90	158.42	R	5.80%	40.61	41.02	93.76	R	8.66%	22.16	22.59	57.90	R
GBRM	2.89%	2.43	2.51	16.97	G	2.57%	0.06	0.07	5.19	G	3.62%	5.66	5.74	19.85	Y	4.76%	3.85	4.23	17.65	Y
QRNN	4.71%	66.66	70.04	223.39	R	4.49%	38.60	39.35	105.05	R	5.23%	29.10	32.44	212.53	R	9.09%	24.82	29.35	176.38	R

Note: Gray fields indicate p-values greater than 5%. GARCH stands for GARCH(1,1), GARCH-t - AR-GARCH-t(1,1), GARCH-st - AR-GARCH(1,1), QML-GARCH - QML-GARCH(1,1), CaViaR - Indirect GARCH(1,1), Mean stands for simple average from GARCH and CaViaR, Highest VaR means the maximum from GARCH, GARCH-t, GARCH-st, QML-GARCH, and CaViaR, Lowest VaR stands for the minimum from individual models, CQOM stands for Conditional Quantile Optimisation Method applied for GARCH and CaViaR (described in section 2.3.4), Elastic Net stands for forecast combined using quantile regression with elastic net regularisation (described in section 2.3.6), LASSO stands for forecast combined using quantile regression with LASSO regularisation (described in section 2.3.5), QRF stands for forecast combined using Quantile Regression Forests (described in section 2.3.7), GBRM stands for forecast combined using Gradient Boosting Regression Model (described in section 2.3.8), QRNN stands for forecast combined using Quantile Regression Neural Network (described in section 2.3.9). In TL: 1) G stands for green, 2) Y stands for yellow, 3) R stands for red.

Table 3. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gold for confidence level equal to 0.99

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.85%	24.38	24.59	47.78	R	1.65%	10.30	10.36	20.23	Y	2.33%	16.25	16.39	40.37	R	3.46%	8.64	9.89	31.80	Y
GARCH-t	1.13%	0.69	6.40	18.19	G	1.06%	0.12	1.02	3.17	G	1.29%	0.96	6.88	34.25	G	2.19%	2.39	5.38	30.03	Y
GARCH-st	0.91%	0.32	4.16	9.17	G	0.89%	0.36	1.79	3.81	G	0.97%	0.01	2.69	9.69	G	2.16%	2.37	5.38	30.07	Y
QML-GARCH	1.90%	24.66	24.59	47.38	R	1.67%	10.35	10.36	20.02	Y	2.37%	16.43	16.39	40.04	R	3.55%	8.87	9.89	31.55	Y
CaViaR	1.03%	0.05	3.07	15.54	G	1.10%	0.27	1.09	6.80	G	0.89%	0.17	3.18	13.09	G	2.60%	4.13	6.42	26.36	Y
Mean	1.30%	3.43	5.09	18.31	Y	1.30%	2.48	2.89	8.34	G	1.29%	0.96	2.61	18.09	G	3.46%	8.64	9.89	31.23	Y
Highest VaR	1.88%	25.64	26.93	52.17	R	1.65%	10.30	10.36	19.95	Y	2.42%	18.01	19.63	49.68	R	3.46%	8.64	9.89	31.27	Y
Lowest VaR	0.84%	1.11	5.53	12.07	G	0.86%	0.63	2.19	3.78	G	0.81%	0.51	3.89	13.52	G	2.16%	2.37	5.38	30.17	Y
CQOM	1.54%	10.50	13.28	51.94	Y	1.44%	5.04	7.12	34.06	Y	1.77%	6.07	6.77	35.15	Y	2.60%	4.13	6.42	52.00	Y
Elastic Net	1.32%	3.98	11.47	62.74	Y	1.10%	0.27	8.30	39.08	G	1.85%	7.28	7.86	40.96	Y	3.46%	8.64	9.89	43.82	Y
LASSO	1.27%	2.92	4.68	34.99	Y	1.06%	0.12	1.02	7.58	G	1.77%	6.07	6.77	42.25	Y	3.90%	11.30	12.19	48.62	Y
QRF	3.61%	171.00	171.03	448.16	R	3.64%	122.05	122.39	323.10	R	3.54%	48.97	52.20	148.36	R	4.76%	17.29	18.39	73.79	R
GBRM	1.35%	4.56	11.81	80.37	Y	0.96%	0.05	4.83	17.66	G	2.25%	14.56	16.57	99.99	R	2.60%	4.13	6.42	47.14	Y
QRNN	3.42%	150.49	160.82	541.10	R	3.29%	96.70	102.40	349.83	R	3.70%	54.22	58.90	206.00	R	6.06%	27.68	27.71	80.57	R

Note: The same as for the pervious table.

12. At a confidence level of 0.95, individual models exhibit dominance. For instance, in the case of silver, it has been ascertained that utilising the mean is the optimal approach for forecasting VaR across all crises, including the ongoing coronavirus crisis.

Nonetheless, it is pertinent to note that, specifically for this asset, none of the models successfully passed the regulatory traffic light assessment, and the aggregate forecasts, based on the mean, did not satisfy any of the evaluation tests. Conversely, for copper, the findings underscore the effectiveness of combining forecast methodologies. Over the entire assessment period, the mean emerges as the most effective model. This holds true for all crisis situations, including the present coronavirus crisis, where the optimal strategy involves fusing quantile regression forecasts with elastic net regularisation.

When the confidence level is set at 0.99, forecast aggregation methods take precedence. Remarkably, the most effective model across all assets and the entire assessment period is the lowest Value at Risk (VaR). During periods of stability, this holds true for gold, silver, and copper, while for oil, the mean demonstrates superiority, and for gas, the GARCH-t model is favored.

In the context of crisis periods, including the current pandemic crisis, the lowest VaR consistently proves to be the superior model, with the exception of gold. Here, during crises, including the ongoing coronavirus crisis, the forecasting superiority lies with the GARCH-st model. Additionally, for gas, the CQOM method outperforms other forecasting strategies during pandemic crises. Similarly, for copper, the GARCH-st model excels in VaR prediction during pandemics.

This comprehensive analysis underscores a notable preference for forecast aggregation techniques, particularly at the 0.99 confidence level. Intriguingly, simplistic methodologies emerge as the most effective approach.

To elucidate the origins of the prevailing forecast amalgamation techniques, it becomes imperative to delve into the allocation of weights to each discrete

Table 4. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for silver for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	3.58%	17.72	21.25	41.94	R	3.46%	9.96	11.56	22.90	Y	3.86%	8.16	10.22	22.05	Y	5.63%	3.85	6.69	28.62	Y
GARCH-t	3.42%	12.86	21.07	38.19	Y	3.16%	4.76	9.27	19.21	Y	4.03%	10.04	13.55	24.26	Y	4.76%	2.85	6.14	22.35	Y
GARCH-st	2.91%	2.74	13.82	34.46	Y	2.68%	0.36	8.13	24.97	G	3.46%	4.22	7.42	16.69	Y	4.33%	2.74	6.14	23.15	Y
QML-GARCH	3.44%	13.51	21.50	43.07	R	3.40%	8.65	11.93	24.52	Y	3.54%	4.91	10.27	22.85	Y	5.69%	2.90	6.14	23.09	Y
CaViaR	2.84%	1.87	3.74	25.47	G	2.64%	0.24	0.24	9.71	G	3.30%	2.98	6.86	29.91	Y	5.19%	10.57	11.72	38.85	Y
Mean	2.77%	1.17	4.91	16.11	G	2.47%	0.01	0.75	5.15	G	3.46%	4.22	7.55	18.32	Y	4.76%	6.88	8.88	31.12	Y
Highest VaR	4.14%	38.32	44.18	74.44	R	3.81%	17.68	20.61	38.17	R	4.91%	23.23	25.99	42.50	R	6.49%	12.63	13.45	34.59	R
Lowest VaR	2.21%	1.46	6.86	19.48	G	2.09%	2.10	2.47	10.64	G	2.50%	0.00	7.41	22.49	G	3.46%	2.61	6.14	30.82	Y
CQOM	2.38%	0.24	18.45	51.46	G	2.16%	1.44	14.15	34.88	G	2.90%	0.77	6.12	20.39	G	3.90%	2.69	6.14	27.15	Y
Elastic Net	2.48%	0.01	8.00	30.60	G	2.30%	0.50	3.40	15.17	G	2.91%	0.78	6.12	22.45	G	6.49%	3.85	6.69	34.39	Y
LASSO	2.72%	0.79	17.15	72.79	G	2.26%	0.69	12.27	39.12	G	3.78%	7.28	11.65	55.77	Y	7.36%	17.15	17.42	73.08	R
QRF	4.81%	72.00	73.95	217.41	R	4.39%	35.03	35.37	107.96	R	5.80%	40.61	42.44	126.89	R	8.66%	10.57	10.57	55.56	Y
GBRM	3.32%	10.41	19.53	59.65	Y	3.05%	3.42	6.62	22.87	Y	3.95%	9.08	15.27	47.33	Y	4.76%	6.88	8.63	29.33	Y
QRNN	5.00%	83.19	102.20	395.17	R	4.70%	46.17	55.58	253.17	R	5.72%	38.88	48.45	157.84	R	9.09%	39.66	40.11	147.69	R

Note: The same as for the table 2, but here GARCH stands for GARCH(1,2), QML-GARCH stands for AR-QML-GARCH(1,1), Mean stands for simple average from GARCH-st and CaViaR.

Table 5. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for silver for confidence level equal to 0.99

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.36%	56.00	62.62	114.64	R	2.30%	36.32	39.22	67.05	R	2.50%	19.83	23.85	56.09	R	3.46%	12.30	15.63	76.45	Y
GARCH-t	1.30%	3.43	19.68	58.92	Y	1.06%	0.12	8.50	25.70	G	1.85%	7.28	14.45	46.95	Y	2.23%	11.52	15.63	73.12	Y
GARCH-st	1.03%	0.05	21.68	77.52	G	0.75%	1.94	14.40	42.41	G	1.69%	4.96	13.18	47.30	Y	2.16%	11.30	15.63	75.03	Y
QML-GARCH	2.31%	52.56	62.26	120.78	R	2.20%	31.39	34.78	61.32	R	2.58%	21.72	28.68	71.24	R	3.51%	12.73	15.63	74.50	Y
CaViaR	1.11%	0.46	3.05	19.34	G	0.93%	0.16	1.48	4.38	G	1.53%	3.03	4.14	39.88	Y	2.60%	20.58	20.79	88.35	R
Mean	1.01%	0.00	11.44	36.79	G	0.82%	0.98	12.38	38.30	G	1.45%	2.22	3.49	11.08	Y	3.46%	11.90	12.19	48.59	Y
Highest VaR	2.65%	78.37	84.86	157.60	R	2.44%	43.32	45.64	76.26	R	3.14%	36.67	41.01	101.95	R	3.46%	31.46	32.44	130.63	R
Lowest VaR	0.70%	4.29	10.09	20.68	G	0.62%	4.99	7.69	9.93	G	0.89%	0.17	3.18	21.35	G	2.16%	4.13	6.42	55.81	Y
CQOM	1.64%	14.24	29.29	108.75	R	1.51%	6.61	19.03	76.54	Y	1.93%	8.57	11.55	52.32	Y	2.60%	11.60	12.19	54.94	Y
Elastic Net	1.30%	3.43	19.68	87.01	Y	1.10%	0.27	8.30	44.48	G	1.77%	6.07	13.76	55.34	Y	3.46%	14.19	17.72	82.18	R
LASSO	1.20%	1.62	24.93	104.71	G	1.06%	0.12	13.66	64.24	G	1.53%	3.03	12.44	50.30	Y	3.90%	11.93	15.63	65.06	Y
QRF	2.89%	99.07	99.71	290.41	R	2.64%	54.69	56.27	136.88	R	3.46%	46.41	46.60	187.82	R	4.76%	13.24	12.03	98.63	Y
GBRM	1.56%	11.38	23.52	68.29	Y	1.41%	4.32	13.57	42.35	Y	1.93%	8.57	11.55	40.85	Y	2.60%	2.37	11.96	92.74	Y
QRNN	3.61%	171.00	177.55	776.88	R	3.46%	109.12	115.76	459.11	R	3.95%	62.45	63.00	339.63	R	6.06%	47.94	48.22	305.35	R

Note: The same as for the previous table.

Table 6. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for oil for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.96%	3.39	3.43	10.66	Y	2.37%	0.22	3.56	7.97	G	4.35%	14.30	15.37	19.82	R	5.63%	6.88	8.63	20.99	Y
GARCH-t	3.54%	16.26	37.79	313.36	R	1.75%	7.51	8.57	55.08	G	7.73%	90.37	99.57	311.51	R	4.76%	22.16	24.91	157.82	R
GARCH-st	2.43%	0.09	0.95	5.04	G	1.96%	3.83	6.10	8.43	G	3.54%	4.91	7.86	9.45	Y	4.33%	2.61	6.14	21.69	Y
QML-GARCH	3.04%	3.53	3.43	10.71	Y	2.39%	0.24	3.56	8.09	G	4.35%	14.30	15.37	19.84	R	5.77%	6.97	8.63	20.99	Y
CaViaR	2.57%	0.09	1.66	14.89	G	1.96%	3.83	3.84	7.94	G	4.03%	10.04	11.73	24.06	Y	5.19%	3.85	6.69	21.18	Y
Mean	2.50%	0.00	0.68	18.03	G	1.37%	18.14	19.25	26.72	G	5.15%	27.59	27.74	37.72	R	4.76%	5.28	5.49	11.16	Y
Highest VaR	4.86%	74.74	83.96	264.82	R	3.16%	4.76	4.76	29.49	Y	8.86%	125.64	131.16	328.47	R	6.49%	39.66	41.25	172.61	R
Lowest VaR	1.59%	16.27	18.80	24.20	G	0.99%	34.98	35.56	29.30	G	2.98%	1.10	3.52	7.61	G	3.46%	0.25	1.96	6.54	G
CQOM	3.01%	4.12	22.26	157.69	Y	1.78%	6.80	6.81	45.91	G	5.88%	42.38	56.14	181.95	R	3.90%	27.60	34.10	133.16	R
Elastic Net	2.98%	3.75	5.04	22.36	Y	1.96%	3.83	3.84	9.29	G	5.39%	32.24	32.77	59.60	R	6.49%	14.83	15.27	63.54	R
LASSO	3.03%	4.51	10.03	69.61	Y	1.78%	6.80	7.78	12.84	G	5.96%	44.17	45.66	108.55	R	7.36%	17.15	18.83	85.82	R
QRF	5.82%	137.72	139.48	253.06	R	5.28%	70.54	70.63	217.80	R	7.09%	72.17	74.44	124.92	R	8.66%	19.60	25.83	74.65	R
GBRM	3.39%	12.23	12.24	24.87	Y	2.81%	1.13	2.11	15.93	G	4.75%	20.50	21.00	32.50	R	4.76%	3.85	4.23	13.35	Y
QRNN	5.65%	125.62	126.08	208.16	R	4.73%	47.48	50.07	297.87	R	7.81%	92.75	92.77	156.26	R	9.09%	53.11	53.12	134.19	R

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,1), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-t and CaViaR.

Table 7. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for oil for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	
GARCH	1.54%	10.50	13.28	25.45	Y	1.27%	1.97	2.92	11.38	G	2.17%	12.95	18.36	24.49	R	3.43%	8.41	13.93	72.02	Y
GARCH-t	2.41%	59.53	80.88	490.49	R	0.96%	0.05	1.25	106.88	G	5.80%	136.83	145.95	536.93	R	2.16%	39.45	44.28	346.76	R
GARCH-st	1.01%	0.00	0.58	3.13	G	0.89%	0.36	0.82	10.06	G	1.29%	0.96	2.61	6.28	G	2.16%	4.13	6.42	24.85	Y
QML-GARCH	1.52%	9.64	12.55	24.38	Y	1.29%	2.01	2.92	11.47	G	2.09%	11.41	17.23	23.12	Y	3.45%	8.55	13.93	72.03	Y
CaViaR	1.49%	8.81	8.82	14.83	Y	1.10%	0.27	0.98	5.46	G	2.42%	18.01	18.10	27.26	R	2.60%	6.24	7.95	36.61	Y
Mean	1.44%	7.26	7.28	20.93	Y	0.96%	0.05	0.59	21.48	G	2.58%	21.72	21.76	30.45	R	3.41%	8.24	7.95	37.61	Y
Highest VaR	3.49%	158.07	169.88	460.74	R	1.99%	22.40	22.42	79.80	R	7.00%	194.17	201.30	574.54	R	3.46%	56.90	61.42	306.55	R
Lowest VaR	0.46%	15.51	18.66	17.14	G	0.27%	21.77	21.81	15.72	G	0.89%	0.17	3.18	8.15	G	2.16%	1.02	4.98	40.55	G
CQOM	2.36%	56.00	94.71	707.99	R	1.51%	6.61	11.24	151.64	Y	4.35%	76.99	104.99	671.01	R	2.60%	39.45	44.28	297.04	R
Elastic Net	1.64%	14.24	14.83	27.95	R	0.99%	0.00	0.58	15.67	G	3.14%	36.67	37.11	69.57	R	3.46%	14.19	14.79	46.58	R
LASSO	1.59%	12.31	14.83	120.13	Y	0.79%	1.41	1.78	41.45	G	3.46%	46.41	47.70	137.00	R	3.90%	17.29	17.67	112.70	R
QRF	3.68%	178.14	179.09	390.91	R	3.26%	94.28	94.76	279.50	R	4.67%	89.32	92.69	240.24	R	4.76%	20.58	26.33	125.31	R
GBRM	1.80%	21.93	22.21	34.37	R	1.48%	5.80	7.09	38.64	Y	2.58%	21.72	23.01	40.31	R	2.60%	6.24	7.95	25.84	Y
QRNN	4.43%	267.55	269.31	1197.81	R	3.40%	104.09	104.84	533.39	R	6.84%	186.18	186.44	816.55	R	6.06%	66.27	69.41	542.78	R

Note: The same as for the previous table.

Table 8. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gas for confidence level equal to 0.975.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.19%	1.72	1.72	5.62	G	1.78%	6.80	6.81	8.64	G	3.14%	1.93	1.98	7.75	G	5.63%	0.79	1.36	10.87	G
GARCH-t	2.09%	2.99	3.01	6.79	G	1.75%	7.51	7.52	8.66	G	2.90%	0.77	0.77	5.75	G	4.76%	1.58	2.32	10.97	G
GARCH-st	2.45%	0.04	0.13	2.93	G	2.09%	2.10	2.47	2.96	G	3.30%	2.98	3.08	10.15	Y	4.33%	1.49	2.32	11.11	G
QML-GARCH	2.24%	1.22	1.22	4.27	G	1.85%	5.50	5.50	6.19	G	3.14%	1.93	1.98	7.59	G	5.63%	0.79	1.36	11.01	G
CaViaR	2.53%	0.01	0.06	12.09	G	1.99%	3.34	3.88	9.26	G	3.78%	7.28	7.71	20.93	Y	5.19%	1.57	2.32	13.96	G
Mean	2.24%	1.22	1.60	9.07	G	1.82%	6.13	7.02	11.42	G	3.22%	2.43	2.50	10.27	Y	4.76%	0.79	1.36	10.88	G
Highest VaR	2.86%	2.14	2.20	8.51	G	2.37%	0.22	0.30	1.96	G	4.03%	10.04	10.71	24.08	Y	6.49%	2.61	3.52	14.67	Y
Lowest VaR	1.83%	8.48	8.72	12.70	G	1.41%	16.94	17.20	17.60	G	2.82%	0.50	0.50	5.85	G	3.46%	0.79	1.36	10.63	G
CQOM	4.52%	56.49	67.33	232.88	R	4.22%	29.41	43.33	169.76	R	5.23%	29.10	29.21	76.81	R	3.90%	0.63	0.77	10.55	G
Elastic Net	2.45%	0.04	0.84	8.38	G	2.02%	2.90	7.23	11.76	G	3.46%	4.22	7.31	17.85	Y	6.49%	1.79	2.32	10.66	G
LASSO	2.45%	0.04	2.09	13.69	G	2.09%	2.10	6.04	14.24	G	3.30%	2.98	3.08	17.62	Y	7.36%	1.86	2.32	9.07	G
QRF	5.53%	117.24	117.29	357.06	R	5.25%	68.99	69.14	221.09	R	6.20%	49.73	49.74	141.98	R	8.66%	6.88	8.44	18.56	Y
GBRM	3.30%	9.83	12.20	37.10	Y	2.85%	1.38	2.39	12.48	G	4.35%	14.30	15.37	39.06	R	4.76%	2.61	3.22	12.55	Y
QRNN	5.27%	99.63	105.52	453.83	R	5.21%	67.45	70.62	349.00	R	5.39%	32.24	35.09	119.62	R	9.09%	6.88	6.98	64.95	Y

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,1), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-t and CaViaR.

Table 9. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for gas for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.03%	0.05	0.95	3.12	G	0.79%	1.41	1.78	5.43	G	1.61%	3.94	4.60	9.29	Y	3.46%	2.37	2.60	26.03	Y
GARCH-t	0.67%	5.05	5.43	8.05	G	0.48%	9.84	9.98	10.05	G	1.13%	0.20	0.51	2.29	G	2.16%	0.19	0.27	2.44	G
GARCH-st	0.99%	0.01	0.82	9.81	G	0.75%	1.94	2.27	6.23	G	1.53%	3.03	3.62	13.25	Y	2.09%	1.02	1.17	27.33	G
QML-GARCH	1.08%	0.28	1.26	3.39	G	0.81%	1.48	1.78	5.38	G	1.77%	6.07	6.87	12.93	Y	3.54%	2.45	2.60	26.19	Y
CaViaR	1.05%	0.07	0.95	5.60	G	0.82%	0.98	1.38	5.06	G	1.53%	3.03	3.62	10.03	Y	2.60%	1.02	1.17	9.73	G
Mean	0.84%	1.11	1.70	4.82	G	0.65%	4.07	4.32	8.78	G	1.29%	0.96	1.37	4.27	G	3.46%	0.29	0.27	2.80	G
Highest VaR	1.25%	2.45	3.77	10.26	Y	0.99%	0.00	0.58	2.86	G	1.85%	7.28	8.14	19.35	Y	3.46%	2.37	2.60	26.07	Y
Lowest VaR	0.63%	6.80	7.12	9.37	G	0.41%	13.10	13.20	12.22	G	1.13%	0.20	0.51	2.41	G	2.16%	0.19	0.27	2.70	G
CQOM	4.40%	264.51	268.51	748.76	R	4.67%	209.24	212.38	615.08	R	3.78%	56.92	57.68	156.26	R	2.60%	11.30	12.03	29.88	Y
Elastic Net	3.49%	158.07	167.67	609.97	R	3.43%	106.60	113.48	430.37	R	3.62%	51.57	54.28	217.59	R	3.46%	61.53	61.96	212.30	R
LASSO	1.35%	4.56	11.81	67.51	Y	1.10%	0.27	8.30	4.77	G	1.93%	8.57	9.06	39.17	Y	3.90%	6.24	6.68	26.86	Y
QRF	3.32%	140.58	140.62	422.57	R	2.78%	62.80	62.83	203.90	R	4.59%	86.18	86.24	262.53	R	4.76%	17.29	18.39	63.33	R
GBRM	1.54%	10.50	10.50	39.50	Y	1.23%	1.51	2.04	16.13	G	2.25%	14.56	15.85	44.46	R	2.60%	2.37	2.60	25.28	Y
QRNN	3.78%	Inf	Inf	1181.52	R	4.12%	160.80	160.80	1024.17	R	2.98%	32.11	34.39	194.18	R	6.06%	4.13	4.46	71.51	Y

Note: The same as for the previous table.

Table 10. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for copper for confidence level equal to 0.975

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	2.96%	3.39	4.77	19.79	Y	2.61%	0.14	0.14	4.71	G	3.78%	7.28	9.54	30.64	Y	5.63%	3.85	6.69	29.71	Y
GARCH-t	2.67%	0.48	3.05	18.85	G	2.33%	0.34	0.45	2.93	G	3.46%	4.22	7.42	36.25	Y	4.76%	1.58	5.92	35.76	G
GARCH-st	2.65%	0.36	3.04	20.04	G	2.37%	0.22	0.30	3.05	G	3.30%	2.98	6.71	39.01	Y	4.33%	2.61	6.14	31.14	Y
QML-GARCH	2.99%	3.44	4.77	19.46	Y	2.65%	0.16	0.14	4.21	G	3.78%	7.28	9.54	30.72	Y	5.63%	3.85	6.69	42.44	Y
CaViaR	2.57%	0.09	0.62	10.85	G	2.26%	0.69	3.75	9.95	G	3.30%	2.98	6.71	16.07	Y	5.19%	2.71	6.14	51.69	Y
Mean	2.43%	0.09	0.95	10.86	G	2.23%	0.91	3.87	4.65	G	2.90%	0.77	6.12	22.07	G	4.65%	2.55	6.14	40.23	Y
Highest VaR	3.27%	9.27	10.61	29.07	Y	2.88%	1.66	1.74	12.79	G	4.19%	12.09	15.09	45.05	Y	6.49%	3.85	6.69	30.17	Y
Lowest VaR	2.12%	2.64	4.58	13.18	G	1.92%	4.35	6.54	7.42	G	2.58%	0.03	6.99	23.06	G	3.46%	1.58	5.92	40.33	G
CQOM	4.43%	51.67	Inf	960.75	R	2.98%	2.65	13.39	230.53	G	7.81%	92.75	121.91	806.55	R	3.90%	218.89	219.00	703.95	R
Elastic Net	2.62%	0.25	1.65	17.45	G	2.37%	0.22	0.51	9.27	G	3.22%	2.43	6.46	38.25	Y	6.49%	3.85	6.69	25.60	Y
LASSO	2.62%	0.25	1.65	43.06	G	2.37%	0.22	0.51	36.09	G	3.22%	2.43	6.46	30.80	Y	7.36%	5.28	7.53	31.78	Y
QRF	4.91%	77.52	78.43	231.96	R	4.39%	35.03	35.06	121.60	R	6.12%	47.85	49.03	128.29	R	8.66%	10.57	13.88	49.29	Y
GBRM	3.08%	5.33	7.41	33.89	Y	2.95%	2.29	3.56	21.25	G	3.38%	3.57	12.79	40.09	Y	4.76%	3.85	6.69	23.83	Y
QRNN	4.52%	56.49	61.01	421.32	R	4.49%	38.60	39.35	325.86	R	4.59%	17.91	23.66	136.29	R	9.09%	10.57	11.54	78.96	Y

Note: The same as for the table 2, but here GARCH-t stands for GARCH-t(1,2), GARCH-st stands for GARCH-st(1,1), Mean stands for simple average from GARCH-st and CaViaR.

Table 11. Test results: Excess Ratio (ER), Kupiec (UC), Christoffersen (CC), Dynamic Quantile (DQ) and Traffic Light (TL) divided into the analysed models and periods for copper for confidence level equal to 0.99.

Model	Period I (Whole period)					Period II (All calm periods)					Period I (All crisis periods)					Period I (COVID period)				
	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL	ER	UC	CC	DQ	TL
GARCH	1.71%	17.36	21.56	53.94	R	1.44%	5.04	6.26	9.79	G	2.33%	16.25	24.63	98.67	R	3.46%	11.30	15.63	110.23	Y
GARCH-t	1.18%	1.27	10.40	38.94	G	0.93%	0.16	0.67	8.74	G	1.77%	6.07	18.76	79.58	Y	2.16%	4.13	11.96	99.07	Y
GARCH-st	1.08%	0.28	10.67	40.15	G	0.89%	0.36	0.82	8.16	G	1.53%	3.03	18.17	92.27	Y	2.12%	4.11	11.96	99.36	Y
QML-GARCH	1.78%	20.75	24.47	55.48	R	1.54%	7.47	8.88	12.95	G	2.33%	16.25	24.63	98.80	R	3.49%	11.41	15.63	131.62	Y
CaViaR	1.20%	1.62	3.71	27.18	G	1.13%	0.49	1.25	16.74	G	1.37%	1.53	6.99	30.44	G	2.60%	2.37	5.38	82.34	Y
Mean	1.01%	0.00	11.44	45.52	G	0.89%	0.36	0.82	6.08	G	1.29%	0.96	19.09	118.12	G	3.46%	4.13	11.96	107.81	Y
Highest VaR	1.95%	29.58	32.33	73.24	R	1.72%	12.41	14.15	27.03	G	2.50%	19.83	27.24	95.92	R	3.46%	11.30	15.63	110.69	Y
Lowest VaR	0.87%	0.79	5.01	21.97	G	0.75%	1.94	2.27	8.58	G	1.13%	0.20	7.18	38.09	G	2.16%	2.37	5.38	59.22	Y
CQOM	1.85%	24.38	33.04	376.94	R	1.34%	3.04	3.40	84.60	G	3.06%	34.37	42.11	389.48	R	2.60%	31.46	32.44	159.44	R
Elastic Net	1.27%	2.92	7.43	78.15	Y	1.13%	0.49	1.25	53.05	G	1.61%	3.94	12.74	66.78	Y	3.46%	2.37	5.38	26.06	Y
LASSO	1.23%	2.02	2.21	136.46	G	1.06%	0.12	0.78	55.47	G	1.61%	3.94	4.90	112.99	Y	3.90%	2.42	5.38	27.19	Y
QRF	3.63%	173.63	178.33	542.73	R	3.36%	101.61	101.76	317.23	Y	4.27%	74.00	81.26	275.12	R	4.76%	13.24	12.19	81.67	Y
GBRM	1.73%	18.46	20.31	64.03	R	1.48%	5.80	7.09	29.37	G	2.33%	16.25	20.93	56.39	R	2.60%	9.54	12.19	49.56	Y
QRNN	3.56%	165.79	168.16	374.11	R	4.01%	152.20	153.27	351.61	R	2.50%	19.83	21.28	76.59	R	6.06%	17.29	17.67	104.84	R

Note: The same as for the previous table.

Table 12. The best model for each commodity (rows) and for all periods (columns) for confidence level of 0.975 (upper part), and 0.99 (lower part) achieved using MCS procedure

Model	Period I (Whole period)	Period II (All calm periods)	Period III (All crisis periods)	Period IV (COVID period)
Confidence level = 0.025				
Gold	GARCH-t	GARCH-t	GARCH-t	CQOM
Silver	Mean	GARCH-st	Mean	Mean
Oil	GARCH	GARCH	GARCH-st	GARCH-st
Gas	GARCH-st	Highest VaR	GARCH-st	LASSO
Copper	Mean	GARCH	Elastic Net	Elastic Net
Confidence level = 0.01				
Gold	Lowest VaR	Lowest VaR	GARCH-st	GARCH-st
Silver	Lowest VaR	Lowest VaR	Lowest VaR	Lowest VaR
Oil	Lowest VaR	Mean	Lowest VaR	Lowest VaR
Gas	Lowest VaR	GARCH-t	Lowest VaR	CQOM
Copper	Lowest VaR	Lowest VaR	Lowest VaR	GARCH-st

model, culminating in the formulation of a specific amalgamated projection. In this context, Figure 6 serves as a visual representation, delineating the weights attributed to individual models for various forecast amalgamation methods. A cursory inspection of the charts reveals that during certain intervals, notably from 2007 to 2009, both the regression methodologies and the CQOM accorded near-zero or null weights to specific models, resulting in an intercept-based forecast.

Turning attention to the regularisation techniques of elastic net and LASSO as applied to the gold commodity, definitive conclusions remain elusive. Nonetheless, discernible trends emerge, particularly in crisis periods, where the GARCH-st model garnered heightened prominence, evident across confidence levels of 0.975 and 0.99. Similarly, in the case of silver, the CQOM approach, at a confidence level of 0.975, allocated greater weight to the GARCH model during crisis periods spanning 2014 to 2016 and 2020. Notably, during the subprime crisis from 2007 to 2009, the GARCH-t model took precedence. In contrast, for both elastic net and LASSO methods, crisis epochs prompted a shift in focus towards the GARCH model, whereas tranquil periods saw a preference for the QML-GARCH model. This trend was consistent regardless of the chosen confidence level.

For oil, a similar pattern emerges whereby both elastic net and LASSO methods exhibit a predilection for the CaViaR model during serene intervals. Intriguingly, the crisis that unfolded post-2007 posed significant modeling challenges, as evidenced by the conspicuous absence of weight allocation to forecasts. Instead, emphasis centred on intercept regulation, indicative of the exceptional difficulty posed by this crisis from the perspective of individual model formulation.

In the context of natural gas, for the confidence level of 0.975, both elastic net and LASSO techniques manifested a proclivity for the GARCH-st model during periods of calm, barring the period spanning 2017 to 2020. Notably, this relationship did not hold for the elastic net approach at the 0.99 confidence level.

Evaluating the outcomes for copper, a distinct lack of a singular optimal model emerges, as suitability varies even during periods of tranquillity. Specifically, the GARCH model found favour from 2004 to 2007, succeeded by the CaViaR model from 2009 to 2014, and the GARCH-t model post-2016. However, this trend does not persist at the 0.99 confidence level.

Analysing the Quantile Boosting Regression Model, a consistent pattern emerges wherein the CaViaR model dominates forecasts for all commodities during tranquil intervals, with the exception of oil. For the latter, the preferred model shifts across time spans: CaViaR from 2004 to 2008, GARCH from 2009 to 2014, and GARCH-t from 2016 to 2020. Despite the clarity during tranquil periods, crisis-related dynamics are less apparent. Notably,

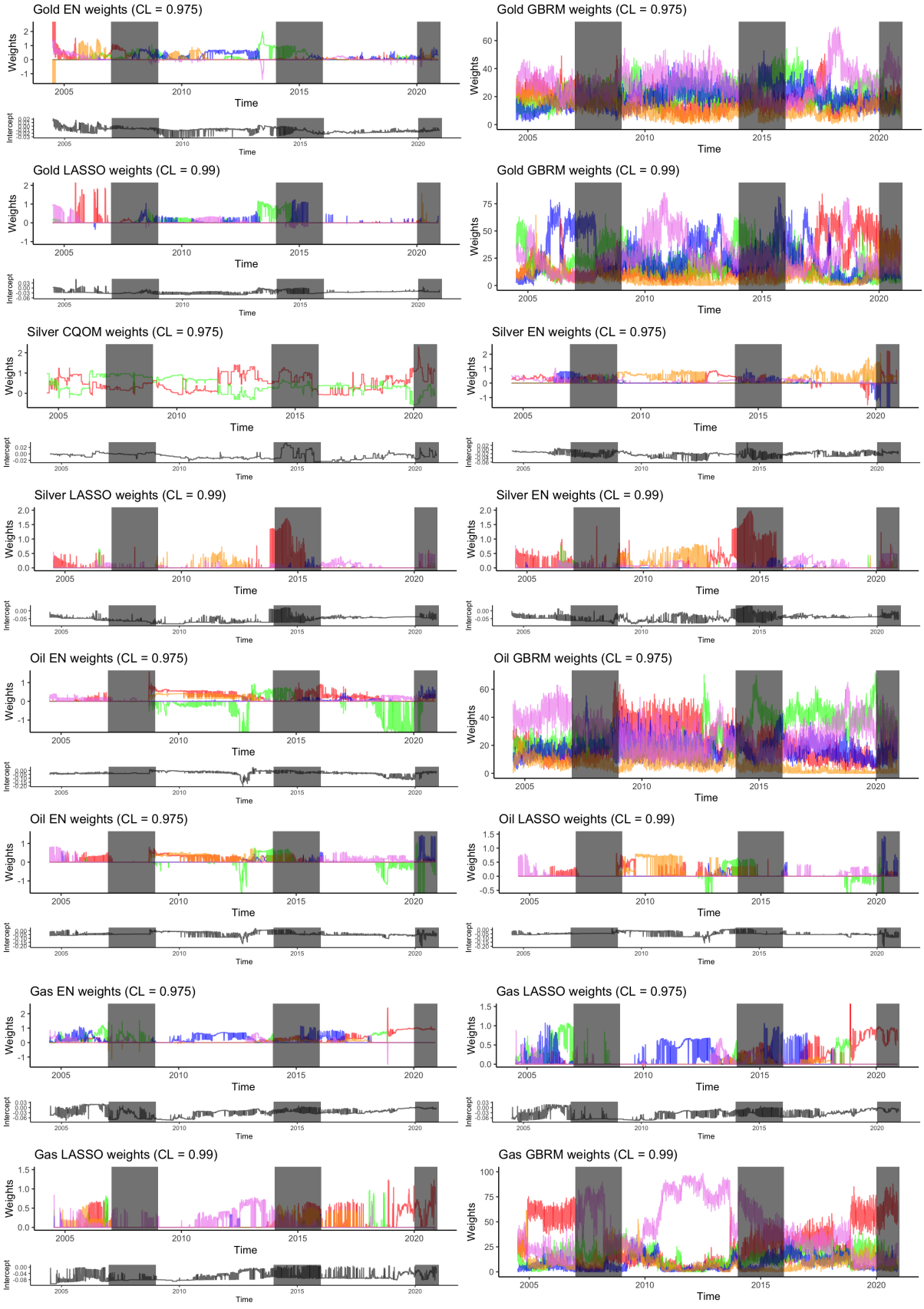
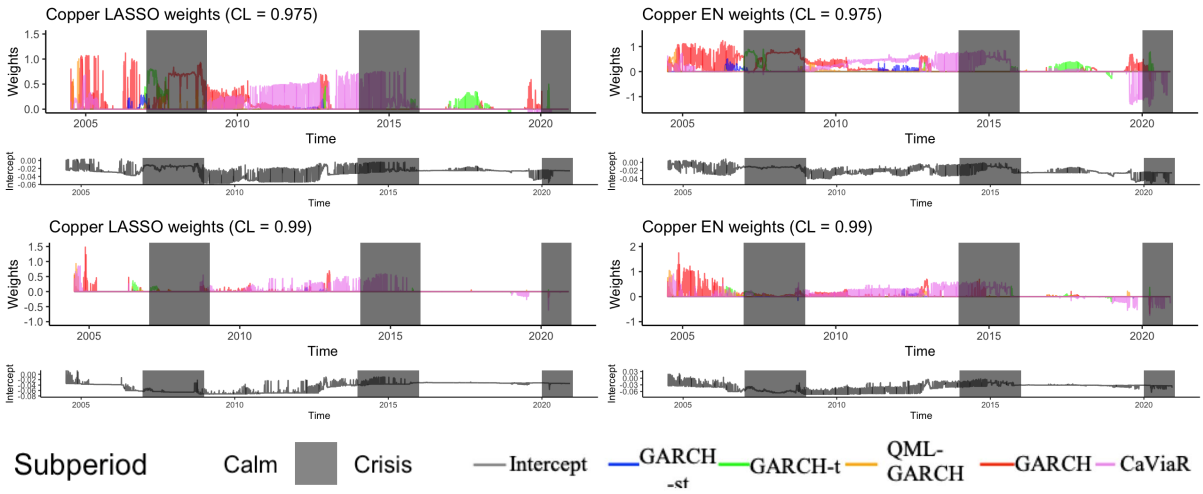


Figure 6. Combing weight for the most promising methods of combining VaR forecasts for all assets for both confidence levels (CL)—0.975 and 0.99



Continued **Figure 6.** Combing weight for the most promising methods of combining VaR forecasts for all assets for both confidence levels (CL)—0.975 and 0.99

the CaViaR model maintained supremacy during oil and gas crises from 2007 onwards. In the case of gold, the CaViaR model only asserted dominance during the coronavirus-induced crisis in 2020, at a 0.975 confidence level. In other scenarios, discerning a leading model proves challenging. For oil, the crisis from 2014 to 2016 witnessed GARCH-t's ascendancy at the 0.975 confidence level, whereas the 2020 crisis saw CaViaR's resurgence. In the realm of gas, the 0.99 confidence level favored CaViaR for the 2014-2020 crisis, while GARCH prevailed for the 2020 crisis.

Regrettably, a parallel analysis for silver and copper was precluded due to unpromising results yielded by this methodology. The collective outcomes underscore the feasibility of distinguishing dominant models within specific forecast amalgamation methodologies, accentuating the prominence of select individual models over their counterparts.

The first hypothesis positing a heightened degree of forecast accuracy through the amalgamation of methodologies across the entirety of the assessment period has been unequivocally validated at a confidence level of 0.99. This validation extends to a partial extent at a confidence level of 0.975, specifically observed in the context of silver and copper markets.

Subsequently, the second hypothesis, which underscores an enhanced precision in forecast outcomes derived from the confluence of models during periods characterised by tranquillity, finds affirmation solely in the case of the gas market. This affirmation is notable at the confidence level of 0.975, and nearly complete at the 0.99 confidence level, delineating gas as the sole asset wherein the individual model surpassed the composite approach.

In relation to the third hypothesis, which advances the notion of improved forecast accuracy achieved through the integration of methodologies during periods of crisis, partial confirmation is evident. This is evident at the 0.975 confidence level for silver and copper markets, and at the 0.99 confidence level for all assets with the exception of gold.

The fourth hypothesis, postulating the supremacy of forecast combining methodologies over individual models, stands largely substantiated at the 0.975 confidence level. Notably, this is exceptive in the case of oil. At the 0.99 confidence level, the hypothesis obtains partial validation, with gold and copper markets serving as exceptions.

The foremost forecast combining models, which have demonstrated superior performance, are notably the lowest Value at Risk (VaR) and the arithmetic mean.

4. Conclusions

The provided study aimed to assess the effectiveness of different Value at Risk (VaR) forecasting models and forecast combining methods in predicting risk levels for commodities at two distinct confidence levels (0.975 and 0.99). The study not only compared

the accuracy of individual VaR models with combined forecasts but also delved into the impact of individual forecasts on the combined forecast. The research found that, at the 0.975 confidence level, the results were somewhat aligned with expectations, while at the 0.99 confidence level, the results were largely as anticipated.

The present study's findings serve to reinforce previous research, specifically highlighting the efficacy of employing simple combining techniques. Notably, Huang & Lee's (2013) investigation revealed that the mean and median computed across all individual forecasts exhibited superior performance. Their study involved the amalgamation of Value at Risk (VaR) predictions derived from models utilising high-frequency information. Similarly, Bayer's (2018) work demonstrated that the straightforward mean calculation across all forecasts yielded commendable results, often comparable to more intricate methodologies.

Consequently, our comparative analysis deviates from two established conclusions within the mean forecasting literature, as outlined by Timmermann (2006), which advocate for the utilisation of trimming and averaging based on ranks to enhance simpler variants. Additionally, the strategy of selecting a solitary model on a day-to-day basis exhibited inferior performance compared to averaging techniques and even trailed behind several standalone models. This observation aligns with the findings presented in Aiolfi & Timmermann's (2006) research.

Furthermore, Taylor (2020) underscored the benefits of combining methodologies, often resulting in enhanced forecasting accuracy for the mean. This viewpoint is corroborated by the work of Lyocsa et al. (2021). Notably, our analysis reveals that the predictive capacity of the Expected Shortfall (ES) model in the given context produced well-specified predictions devoid of systematic biases.

Despite the utilisation of more complex combining methods, the basic techniques—specifically the lowest VaR and the average—proved to be the most effective in many instances. Still, regression methods exhibited promising outcomes, implying potential for further refining these methods through various parameter adjustments.

The study yielded an intriguing observation: during times of crises, particularly the COVID-19 pandemic, individual models outperformed combining methods for forecasting gold prices. This phenomenon could be attributed to gold's unique role as a safe-haven asset during crises, prompting individuals, institutions, and even countries to seek shelter in gold to safeguard their wealth. Nonetheless, the study acknowledged the necessity for more comprehensive investigation to comprehensively understand this trend.

In essence, the study's primary takeaway is that forecast combining methods, particularly those emphasising simplicity, offer value. This central finding holds notable implications for both the realm of scientific research and practical application within the financial and risk management sectors. The study builds on prior research by confirming the dominance of the average method in forecast combination. This finding corroborates and strengthens the existing knowledge base, offering researchers a more comprehensive picture of which methods are consistently effective across different scenarios. The study's exploration of regression methods and their promising results opens up avenues for further investigation and development in this area. Researchers can delve deeper into refining regression-based techniques, exploring adjustments to parameter estimation windows, tuning parameters, and regularisation strategies. This has the potential to contribute to the advancement of quantitative modeling methodologies. For practitioners in the financial and risk management sectors, the study's findings hold practical implications that can inform decision-making processes and risk management strategies.

However, the study's scope was confined to commodities, necessitating caution when extending its findings to other markets. The study also recognised limitations, such as the use of incomplete data for backtesting during the COVID-19 crisis and default options in employing machine learning models. To expand upon this research, future studies could validate hypotheses in different markets, experiment with diverse loss functions, include forecasts from a broader array of individual models, evaluate hypotheses with complete data during pandemic crises, and explore enhancements for both combining methods.

In summary, this study contributes to the understanding of VaR forecasting and forecast combining methods in the context of commodities. It highlights the significance of simplicity in combining forecasts and recognises potential avenues for further research and improvement. However, generalising the findings demands caution, given the study's specific focus and acknowledged limitations.

References

Andreani, M., Candila, V., & Petrella, L. (2022). Quantile Regression Forest for Value-at-Risk Forecasting Via Mixed-Frequency Data. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance: MAF 2022* (pp. 13-18). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-030-99638-3

Angabini, A., Wasiuzzaman, S. (2011). GARCH Models and the Financial Crisis: A Study of the Malaysian. *The International Journal of Applied Economics and Finance, 5*(3), 226-236. https://doi.org/10.3923/ijaef.2011.226.236

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting*, *5*(4), 585-588. https://doi.org/10.1016/0169-2070(89)90013-7

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. *Disrupting finance: FinTech and strategy in the 21st century*, 33-50.

Basel Committee. (1996). Overview of the Amendment to the Capital Accord to Incorporate Market Risks. *Discussion Paper, Basel Committee on Banking Supervision.*

Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and statistics*, *8*, 56-77. https://doi.org/10.1016/j.ecosta.2017.08.001

BCBS (1996). Supervisory Framework for the Use of 'Backtesting' in Conjunction with the Internal Models Approach to Market Risk Capital Requirements.

BCBS (2010). The Basel III Capital Framework: A Decisive Breakthrough. Speech by Hervé Hannoun at BoJ-BIS High Level Seminar on Financial Regulatory Reform: Implications for Asia and the Pacific, Hong Kong SAR.

Bernardi, M., Catania, L. (2016). Comparison of Value-at-Risk models using the MCS approach. *Computational Statistics*, *31*(2), 579-608. https://doi.org/10.1007/s00180-016-0646-6

Bhowmik, R., & Wang, S. (2020). Stock market volatility and return analysis: A systematic literature review. *Entropy*, *22*(5), 522. https://doi.org/10.3390/e22050522

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics. 31*(3), 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bollerslev, T. (1987). Conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics, 69*(3), 542-547. https://doi.org/10.2307/1925546

Bollerslev, T., Woolridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances *Econometric Reviews 11.* https://doi.org/10.1080/07474939208800229

Buczyński, M., Chlebus, M. (2018). Comparison of semi-parametric and benchmark value-at-risk models in several time periods with different volatility levels. *e-Finanse: Financial Internet Quarterly*, *14*(2), 67-82. https://doi.org/10.2478/fiqf-2018-0013

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

Caillault, É. P., Lefebvre, A., and Bigand, A. (2017). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2017.08.019

Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes: An International Journal*, *24*(6), 673-685. https://doi.org/10.1002/hyp.7506

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences*, *37*(9), 1277-1284. https://doi.org/10.1016/j.cageo.2010.07.005

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862. https://doi.org/10.2307/2527341

Clemen, R. T., Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39-46. https://doi.org/10.2307/1391385

Danielsson, J. (2013). The new market-risk regulations. VoxEU.

Danielsson, J., Morimoto, Y. (2000). *Forecasting extreme financial risk: A critical analysis of practical*

methods for the Japanese market. Institute for Monetary and Economic Studies, Bank of Japan.

Dudziński, J. (2016). Ceny w handlu międzynarodowym w drugiej dekadzie XXI wieku. Kierunki zmian i ich czynniki. *International Business and Global Economy*, *35*(2), 249-260. https://doi.org/10.4467/23539496IB.16.061.5642

Duffie, D., Pan, J. (1997). An overview of value at risk. *Journal of Derivatives*, *4*(3), 7-49. http://doi.org/10.3905/jod.1997.407971

Engle, R. F., Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics, 22*(4), 367-381. http://doi.org/10.1198/073500104000000370

Fameliti, S. P., & Skintzi, V. D. (2020). Predictive ability and economic gains from volatility forecast combinations. *Journal of Forecasting*, *39*(2), 200-219. http://doi.org/10.1002/for.2622

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232. http://dx.doi.org/10.1214/aos/1013203451

Gençay, R., Selçuk, F., Ulugülyağci, A. (2003). High volatility, thick tails and extreme value theory in value-at-risk estimation. *Insurance: Mathematics and Economics*, *33*(2), 337-356. http://dx.doi.org/10.1016/j.insmatheco.2003.07.004

Giacomini, R., Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, *23*(4), 416-431. http://doi.org/10.1198/073500105000000018

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308-319. https://doi.org/10.1198/tast.2009.08199

Halbleib, R., Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212-1228. https://doi.org/10.1016/j.jedc.2011.10.005

Hansen, P. R., Lunde, A., Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453-497. https://doi.org/10.3982/ECTA5771

Holthausen, D. M., Hughes, J. S. (1978). Commodity returns and capital asset pricing. *Financial Management*, 37-44. https://doi.org/10.1177/0972262912460186

Huang, H., Lee, T. H. (2013). Forecasting value-at-risk using high-frequency information. *Econometrics*, *1*(1), 127-140. https://doi.org/10.3390/econometrics1010127

Ichev, R., Marinč, M. (2018). Stock prices and geographic proximity of information: Evidence from the Ebola outbreak. *International Review of Financial Analysis*, *56*, 153-166. https://doi.org/10.1016/j.irfa.2017.12.004

Jeon, J., Taylor, J. W. (2013). Using CAViaR models with implied volatility for Value-at-Risk estimation. *Journal of Forecasting*, *32*(1), 62-74. http://dx.doi.org/10.1002/for.1251

Kupiec, P. (1995). Techniques for verifying the accuracy of risk management models. *Journal of Derivatives, 3*(2), 73-84. https://doi.org/10.3905/jod.1995.407942

Laporta, A. G., Merlo, L., & Petrella, L. (2018). Selection of value at risk models for energy commodities. *Energy Economics 74,* 628–643.

Laurent, S., Rombouts, J. V., & Violante, F. (2012). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*, *27*(6), 934-955. https://doi.org/10.1002/jae.1248

Lyócsa, Š., Todorova, N., & Výrost, T. (2021). Predicting risk in energy markets: low-frequency data still matter. *Applied Energy*, *282*, 116–146.

Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *IEEE Access*, *8*, 203203–203223.

McAleer, M., Jimenez-Martin, J. A., Perez Amaral, T. (2010). Has the Basel II Accord encouraged risk management during the 2008-09 financial crisis? *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.1397239

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

Mensi, W., Sensoy, A., Vo, X. V., Kang, S. H. (2020). Impact of COVID-19 outbreak on asymmetric multifractality of gold and oil prices. *Resources Policy*, *69*, 101829. https://doi.org/10.1016%2Fj.resourpol.2020.101829

Phillips, P. C., Yu, J. (2011). Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, *2*(3), 455–491. http://dx.doi.org/10.3982/QE82

Parot, A., Michell, K., & Kristjanpoller, W. D. (2019). Using Artificial Neural Networks to forecast Exchange Rate, including VAR-VECM residual analysis and prediction linear combination. *Intelligent Systems in Accounting, Finance and Management*, *26*(1), 3-15. https://doi.org/10.1002/isaf.1440

Pradeepkumar, D., & Ravi, V. (2017). Forecasting financial time series volatility using particle swarm optimisation trained quantile regression neural network. *Applied Soft Computing*, *58*, 35-52. https://doi.org/10.1016/j.asoc.2017.04.014

Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, *9*(24), 5574.

Stuermer, M., & Valckx, N. (2021). Four Factors Behind the Metals Price Rally. IMF.

Szakmary, A. C., Shen, Q., Sharma, S. C. (2010). Trend-following trading strategies in commodity futures: A re-examination. *Journal of Banking & Finance*, *34*(2), 409–426. http://dx.doi.org/10.1016/j.jbankfin.2009.08.004

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, *36*(2), 428–441. https://doi.org/10.1016/j.ijforecast.2019.05.014

Terui, N., Van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, *18*(3), 421–438. https://doi.org/10.1016/S0169-2070(01)00120-0

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, *1*, 135–196. https://doi.org/10.1016/S1574-0706(05)01004-9

Tsay, R. S. (2005). *Analysis of Financial Time Series* (Vol. 543). John Wiley & Sons.

Tse, Y. (2016). Asymmetric volatility, skewness, and downside risk in different asset classes: Evidence from futures markets. *Financial Review*, *51*(1), 83–111. https://doi.org/10.1111/fire.12095

Wasserbacher, H., & Spindler, M. (2022). Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. *Digital Finance*, *4*(1), 63–88.

Xiao, D., Su, J., & Ayub, B. (2022). Economic policy uncertainty and commodity market volatility: implications for economic recovery. *Environmental Science and Pollution Research*, *29*(40), 60662–60673.

Youssef, M., Belkacem, L., Mokni, K., 2015. Value-at-Risk estimation of energy commodities: A long-memory GARCH–EVT approach. *Energy Economics*, *51*, 99–110. https://doi.org/10.1016/j.eneco.2015.06.010