Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Degree BSc^2 Econometrics & Economics

Hedging S&P500 with Oil using the Hedging Random Forest

Thijs ter Horst (536830)

Abstract

This paper evaluates the performance of the Hedging Random Forest (HRF) model in hedging the SP 500 index using the Brent Oil index. The HRF model considers transaction costs and estimates time-varying hedge ratios based on monthly return data and macroeconomic variables. The performance of the HRF model is compared to five benchmark models: OLS, DCC-GARCH, ADCC-GARCH, DCC-GJR-GARCH, and GOGARCH. Although the HRF model does not significantly outperform the benchmark models in terms of hedging effectiveness, HRFs with high transaction cost sensitivities perform well in terms of utility gains that incorporate portfolio returns, risks, and transaction costs. The variable importance analysis of the HRF reveals the significance of a diverse set of macroeconomic drivers, including the 10-year Treasury spread, M2 money supply, and JPYUSD exchange rate.

Supervisor:	Anastasija Tetereva
Date final version:	1st July 2023

1 Introduction

Understanding and using the correlation between stock and oil markets is essential for portfolio managers to minimise the risk of adverse price fluctuations. Recent macroeconomic events have illustrated the impact of oil supply uncertainties, and the related oil price fluctuations, on stock performance. The global shift towards renewable energy and the simultaneous industrial expansion and development of lower income countries makes the demand side of the energy industry volatile. Simultaneously, the supply of fossil fuels is decreasing and local supply is impacted by geopolitics and energy policies. The exposure of investors to these uncertainties and complex interactions requires them to offset energy-related risks, a process called hedging. Whereas hedging via the futures market could reduce risks, futures are illiquid and often not a perfect fit for the required hedge. Cross-asset hedging offsets risk by leveraging co-movement information between two different assets. Central in this type of hedging is to determine the risk-minimising hedging ratio, i.e. the ratio of the hedge to the initial position that minimises the portfolio risk. The minimum-variance hedge ratio gives the position (in $ amounts) that portfolio managers should take in oil to offset their oil price exposure per $1 of the stock position. Hence, minimising the downside risk of a position in the S&P500 index using the Brent Oil index by finding the minimum-variance hedge ratio is a relevant topic to study.

In the literature, many studies have focused on estimating the time-varying hedge ratios for oil and equity markets using different estimation techniques (Arouri, Jouini & Nguyen, 2011; Basher & Sadorsky, 2016; Batten, Kinateder, Szilagyi & Wagner, 2017). However, most of these studies ignore the transaction costs present in rebalancing portfolio weights. The Hedging Random Forest (HRF) (Van der Bij, Ter Horst, Palim & Zegwaart, 2023) is a model that seeks to find the optimal time-varying portfolio weights whilst considering portfolio rebalancing costs. It is a local linear random forest, based on the local linear Macroeconomic Random Forest (MRF) (Coulombe, 2020), that uses macroeconomic data to allocate observations to the leaves. Transaction costs are directly considered by the change in hedge ratios between consecutive time periods. This results in a smoother set of optimal portfolio weights throughout time. Van der Bij et al. (2023) have shown substantial in-sample gains when hedging the S&P 500 with the VIX.

Therefore, the central goal of this paper is to assess the performance of the Hedging Random Forest (HRF) in hedging the S&P500 index using the Brent Oil index.

Time-varying minimum-variance hedge ratios are predicted based on monthly return data of the S&P500 and Brent Oil index and monthly data on financial and macroeconomic variables from January 1990 to January 2023. The HRF model is estimated for different sensitivities to transaction costs. The performance of these HRF models is compared to that of several benchmark models, i.e. Ordinary Least Squares (OLS), DCC-GARCH (R. Engle, 2002), ADCC-GARCH (Cappiello, Engle & Sheppard, 2006), DCC-GJR-GARCH (Glosten, Jagannathan & Runkle, 1993), and GOGARCH models (Van der Weide, 2002). The latter four are analysed as they have shown promising results in the context of hedging stocks with oil. For the HRF and benchmark models, one-step-ahead hedge ratios are constructed using a rolling window technique. The estimation window is set at 120 observations (5 years). This produces 208 one-step-ahead one-step-ahead forecasts. All models are refit every 24 observations (2 years)

and monthly rebalancing is assumed. The performance of the proposed models over the rolling windows is evaluated using the hedging effectiveness (HE), i.e. the proportion of market risk from S&P500 that is offset by the hedge when using Brent oil, and a utility measure in terms of risk, returns, and transaction costs.

The HRF model is unable to significantly outperform OLS and benchmark models in terms of only the hedging effectiveness. Additionally, lower values of the transaction cost penalty λ result in smoother HE paths compared to higher values of λ. The results suggest that relatively high transaction cost penalties should be used in periods with high volatility, whereas low transaction cost sensitivities should be used in case the hedged asset is less volatile than the hedge. Moreover, HRF models with high transaction cost sensitivities perform best amongst all HRF models. HRF models with high λ significantly outperform the multivariate GARCH models in terms of utility, but fail to provide significantly better results compared to OLS. Finally, an analysis of the variable importance of the HRF shows that a diverse set of variables plays an important role in modelling the equity-oil dynamics, including the M2 money stock, JPYUSD exchange rate and 10-year Treasury spread.

The contribution of this research to the literature on local linear models is threefold. Firstly, the out-of-sample performance of the HRF is evaluated. Secondly, the performance of the HRF is compared to models that have shown promising results in previous literature. Finally, measuring the utilities extends the performance analysis of the HRF beyond the HE.

This paper also extends the literature on estimating the minimum-variance hedge ratio for S&P500 and oil. To the best of the author's knowledge, random forests have not previously been applied to estimating the optimal time-varying hedge ratio. In addition, transaction costs are explicitly considered in estimating the hedge ratios. Moreover, the variable importance of the financial and macroeconomic drivers is analysed. This makes the HRF model interpretable and helps investors identify important macroeconomic drivers.

The remainder of the paper is organized as follows. The relevant literature is discussed in Section 2. Section 3 describes the data collected and used in this study. Section 4 discusses the methodology of the HRF, benchmark models, and used performance measures. Section 5 presents the results of the estimated hedge ratios, portfolio returns, drivers, and hedging effectiveness of the different models. Section 6 concludes the findings, and Section 7 outlines the limitations of this paper and further research opportunities.

2 Literature review

This section gives a short review of relevant papers that focus on the development of local linear models (Section 2.1), and hedging equities with oil (Section 2.2).

2.1 Local Linear Models

Random Forests (RFs) (Breiman, 2001) are an ensemble of multiple individual decision trees, where bagging and bootstrapping (Breiman, 1996; Bühlmann & Yu, 2002) increase the smoothness of the trees compared to single trees. RFs handle interactions and nonlinearities in data, while bypassing the overfitting issues of regular regression decision trees. Typical RFs fit a

constant function to each of the different tree leaves. One of the weaknesses of trees is that they do not exploit the smoothness of the prediction surface, but consider it as step-wise. The same step-wise issue holds for random forests. Several papers have considered random forests as adaptive kernel methods (Hothorn, Lausen, Benner & Radespiel-Tröger, 2004; Athey, Tibshirani & Wager, 2019). Based on this literature, Bloniarz, Talwalkar, Yu and Wu (2016) consider local linear regression with supervised weighting functions, and Friedberg, Tibshirani, Athey and Wager (2020) adapt the tree-splitting procedure. As an extension to the latter, Coulombe (2020) introduced the Macroeconomic Random Forest (MRF). This model expands multiple nonlinear time series models by adapting RFs for macro forecasting, interpreting estimates as generalized time-varying parameters, and introducing a five step Olympic podium kernel for time stabilization. In the context of predicting recessions and the Philips curve, the MRF model achieves substantial empirical gains as compared to other nonlinear models, both when modelling short-run regime-switching behaviour and long-run trends. Van der Bij et al. (2023) introduce the HRF. This model extends the MRF by modelling transaction costs into the changing slope parameter estimate. For hedging the S&P500 with the Volatility Index (VIX), the HRF has shown substantial in-sample improvements hedging effectiveness compared to standard GARCH(1,1), OLS, and QTLS models.

This paper adds to the body of literature of local linear models in three ways. Firstly, the out-of-sample performance of the HRF has not been evaluated in earlier literature. The performance of the HRF has been analysed in-sample, which is useful for understanding model fit, but less relevant for investors that seek future returns. Secondly, the HRF model has previously been compared to OLS, QTLS, and univariate GARCH, all relatively simple benchmark models. This paper compares the HRF performance with several benchmark models that have previously proven to be promising in the context of hedging stocks with oil. Finally, the models are analysed in terms of utilities. This measure more accurately reflects investors' practical tradeoffs between portfolio rebalancing costs, risks, and returns than the HE.

2.2 Hedging equities with oil

Due to the central role of energy price exposure in equity markets, several studies have focused on estimating the time-varying hedge ratios between oil and stocks.

Arouri et al. (2011) estimate four bivariate GARCH models (BEKK-GARCH, VAR-GARCH, CCC-GARCH and DCC-GARCH models) using weekly data from 1998 to 2009 to investigate volatility spill-overs between oil and stock market sectors in the US and Europe. For Europe, they find a spillover effect from oil to equity, and for the US, they find a bidirectional spillover effect between oil and the S&P500. The BEKK-GARCH and DCC-GARCH model perform best. Chang, McAleer and Tansuchat (2011) research the BEKK-GARCH, VARMA-GARCH, CCC-GARCH, and DCC-GARCH model for hedging BRENT and WTI crude oil spot with their corresponding crude oil futures. Hedges calculated from DCC-GARCH have the highest hedging effectiveness, while the BEKK-GARCH results in the worst hedges. Based on these results, the DCC-GARCH is included as one of the benchmark models.

Basher and Sadorsky (2016) model volatility dynamics using the DCC-GARCH, ADCC-GARCH and GOGARCH models to estimate the daily hedge ratios between emerging market

stock prices, oil prices, VIX, gold prices and bond prices between. The GARCH models are refit every 20 months and analysed for a series of rolling window, one-step-ahead forecasts. The authors conclude that oil is the best hedge for emerging markets. Additionally, hedging ratios from the ADCC-GARCH and GOGARCH are most effective for hedging emerging market stock prices. Based on these findings, the ADCC-GARCH and GOGARCH model are included as benchmark models.

However, the mentioned literature does not consider the transaction costs that result from the frequent rebalancing of the hedged portfolio positions. Chen and Sutcliffe (2012) show that these hedged portfolios can result in expensive trading due to high transaction costs. The effects of portfolio rebalancing and transaction costs, e.g. bid-ask spreads, on dynamic hedging strategies have been explored in the literature (Coakley, Dollery & Kellard, 2008; Kroner & Sultan, 1993). Batten, Kinateder, Szilagyi and Wagner (2021) emphasise the practical and economic significance of different hedging strategies to investors by comparing the expected utility gains from hedge positions between several equity indices (S&P500 and MSCI indices) and oil indices (Brent and WTI crude oil). Specifically, their approach considers the returns, transaction costs and risks that result from different estimates of the model returns. The utility approach will be used in this paper. Time-varying utilities are also analysed to accurately reflect an investor's tradeoff between portfolio transaction costs, returns and risks throughout time. This allows for more holistic conclusions on the performance of the HRF and benchmark models as compared to the earlier literature, which only measures performance in terms of the hedging effectiveness.

Whereas the incorporation of transaction costs is useful in the performance evluation, it considers transaction costs only after the optimal hedge ratios have been determined. The strength of the HRF model is that it ensures smoother estimates of the time-varying hedge ratio by explicitly considering the presence of rebalancing costs when the values of the hedging ratios are optimised.

Random forests have been previously applied to selecting the optimal portfolio (Tan, Yan & Zhu, 2019), but to the best of the author's knowledge, not to estimating the optimal time-varying hedge ratio. An issue that arises with machine learning models is the lack of interpretability. Another contribution of this paper is that the macroeconomic and financial drivers behind the time-varying hedge ratios are identified via the variable importance of the HRF model. Batten et al. (2021) identified the the implied volatility index (VIX), gold price and term spread as important drivers of stock-oil portfolios based on six macroeconomic variables. This paper contributes by identifying the main drivers behind the time-varying hedge ratios from a set of 450 variables. This improves the interpretability of the random forest model and provides a clearer understanding of the benefits of stock-oil hedging for investors.

3 Data

This section introduces and analyses the data that has been used. Section 3.1 discusses the data, and associated preliminary analyses for the S&P 500 and Brent Oil return data. Section 3.2 introduces the macroeconomic data as input for the HRF.

3.1 S&P 500 and Brent Oil returns

Daily closing prices (in USD) of the S&P 500 were collected from Yahoo Finance for a period from December 1989 to February 2023 (Yahoo Finance, 2023). The S&P 500 index is chosen as proxy for the US equity market as it is the most studied equity index in the literature. Daily energy closing prices (in USD) of the Global price of Brent Crude Oil were collected from the Federal Reserve Bank of St. Louis (FRED) for a period from January 1990 to February 2023 (McCracken & Ng, 2016). The Brent oil index is used instead of the West Texas Intermediate (WTI) contract because previous literature has shown that using Brent oil as a hedge results in a higher hedge effectiveness than using WTI oil (Batten et al., 2021).

In the sample period from January 1990 to December 31 2023, *NBER Business Cycle Dating Committee* (n.d.) has identified four periods of economic recession. These include the periods July 1990 until March 1991, March 2001 until November 2001, December 2007 until June 2009, and February 2020 until April 2020. These periods could be identified as the Gulf War recession, the early 2000s recession, the Great Financial Crisis, and the COVID-19 crisis, respectively. In the remainder of this paper, the term 'recessions' is used with a reference to these time periods.

Monthly equity and oil returns are measured as the difference in the natural logarithm of intermonth closing prices, $R_{it} = ln(P_{i,t}) - ln(P_{i,t-1})$, where $P_{i,t}$ is the closing price of asset $i \in \{s, o\}$ in month t. Monthly returns data is used for two reasons. Firstly, it can counter biases that could arise from daily data (for instance, the bid-ask effect, non-synchronous trading days). Secondly, it corresponds with most macroeconomic data that the HRF model uses is collected at monthly intervals.

The descriptive statistics of S&P 500 and Brent Oil monthly returns are reported in Table 1 for the full sample period from January 1990 to February 2023. At 0.6% per month, the mean return of the S&P 500 is approximately double the mean return of Brent Oil (0.3%). Brent Oil has a substantially higher standard deviation of 0.092 compared to the 0.043 of S&P 500. This can also be observed from the lower minimum and higher maximum monthly return for Brent Oil, and can be seen from the time series plot in Figure 1. Both asset returns have negative skew. This is common in stock markets and means that the there is a higher probability of observing extreme negative returns relative to observing extreme positive returns. From the more negative skew of S&P compared to the skew of Brent Oil, equities seem to display more negative returns (i.e. downside risk) relative to positive returns than oil. Finally, the S&P has a kurtosis of 1.387 and Brent a kurtosis of 3.506. The higher kurtosis indicates that the Brent index has higher fat tailed risk than S&P. The Jarque-Bera test rejects the null hypothesis of normality at the 1% significance level for both oil returns (p-value 5.892e-19) and stock returns (p-value 8.327e-15).

The time series graph of the squared monthly log returns (presented in Figure 1b) is a proxy for how volatility has changed throughout time. Volatility clustering is primarily present around periods of crisis and is stronger for Brent Oil than S&P 500. Based on the presented volatility clustering and the property of returns data to be serially correlated, the conditional volatilities can be modelled by applying GARCH.

Table 2 includes the Pearson pairwise correlation between S&P 500 and Brent Oil monthly log returns for the full sample period, recessionary periods, and non-recessionary periods. The

Table 1: The descriptive statistics (mean, standard deviation (std. dev.), skew, kurtosis), Jarque-Bera's normality test statistic, and Augmented Dicky-Fuller unit root (ADF) test statistic for the S&P 500 and Brent Oil monthly log returns over the full sample period from January 1990 to February 2023.

	Mean	Std. dev.	Skew	Kurtosis	Jarque-Bera	ADF unit root	Obs
S&P 500	0.006	0.043	-0.723	1.387	219.064^{***}	-20.026^{***}	398
Brent Oil	0.003	0.092	-0.578	3.506	64.831^{***}	-12.722^{***}	398

Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively.

Table 2: Pearson pairwise correlations between S&P 500 and Brent Oil returns for the full sample from January 1990 to February 2023, the recessionary months in the sample period and the non-recessionary months in the sample period.

	Full sample	Recessions	Non-recessions
Pearson pairwise Correlation	0.141^{***}	0.322^{**}	0.035

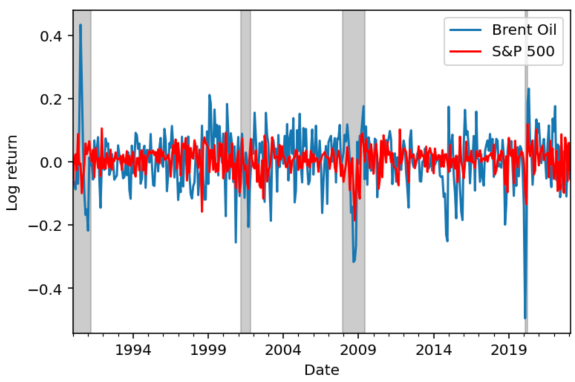
Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively, for the null of no autocorrelation.

latter two are included given the strong similarity between the close prices and returns series plotted in Figure 1 during recessionary periods. An associated test, with a null hypothesis that the distributions are uncorrelated, is applied. During recessions, the pairwise correlation is 0.322 and statistically significant ($p = 0.038$) from the null hypothesis of no autocorrelation at the 5% significance level. During non-recessionary months, the pairwise correlation is not significantly different from zero.

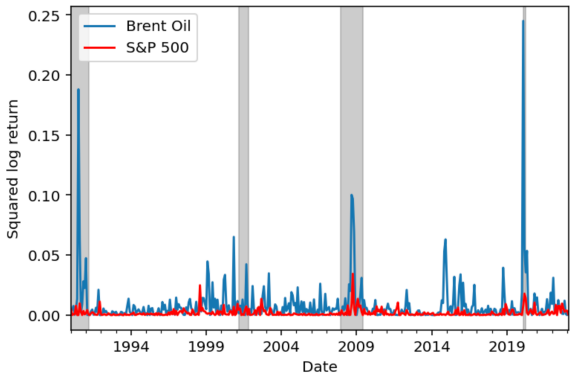
The Augmented Dickey-Fuller (ADF) unit root statistics indicate that monthly returns of both indices are stationary at the 5% significance level (see Table 1). Given stationarity of the time series, the Johansen cointegration test is applied to test the presence of a long term cointegration relationship between energy and stock returns. Based on the trace statistics, the null hypothesis of less than or equal to one cointegration relation is rejected. Thus, taken over the full sample, more than one cointegration relation exists between the asset returns. This indicates a complex long-term relationship among the asset returns and makes modelling time-varying hedge ratios useful.

3.2 Macroeconomic data

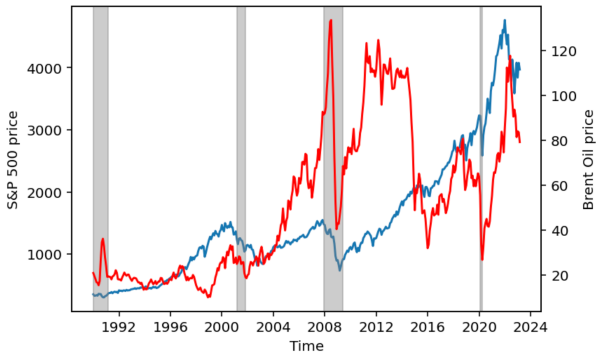
The HRF model requires a large set of macroeconomic and financial variables to make useful tree splits. The core of the used macroeconomic and financial data is retrieved from the Saint Louis branch of the Federal Reserve (McCracken & Ng, 2016). The initial dataset includes 127 variables for a time period from March 1959 until January 2023, which accumulates to a total of 767 monthly time observations per variable. McCracken and Ng (2016) identify eight groups of data in the dataset: Output and income, Labor market, Housing, Consumption, orders and inventory, Money and credit, Interest and exchange rates, Prices, Stock market. To avoid bias in forecasting, one-month lags are taken for each time period. The dataset is transformed using specified transformations for each variable, as elaborated in McCracken and Ng (2016).



(a) Monthly log returns



(b) Squared monthly log returns



(c) Monthly closing prices

Figure 1: Monthly log returns, squared log returns, and closing prices of the S&P 500 (blue) and Brent Oil Index (red) from a period of January 1990 until February 2023. The grey bars indicate the presence of an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

These include taking the first difference, second difference, logarithm, both logarithm and first difference, both logarithm and second difference, and first difference of percentage change.

This dataset was complemented with several indices. The conjecture is that these indices could be relevant in splitting the observations into leaves, as the indices combine much macroeconomic information into one value. The Aruoba-Diebold-Scotti (ADS) index (Aruoba, Diebold & Scotti, 2009) is a business condition tracker. The index includes weekly jobless claims, monthly payroll employment, monthly industrial production, monthly real personal income less transfer payments, monthly real manufacturing and trade sales; and quarterly real GDP). The daily index is downloaded from of Philadelphia (n.d.) for the period of March 1960 to April 2023. In addition, the NFCI weekly is included. This index provides weekly updates on US financial conditions in money markets, debt and equity markets and the banking system. The weekly data is retrieved from Federal Reserve Bank of Chicago (2023) from January 1971 until April 2023. Finally, the change in the US Economic Policy Uncertainty index (S. R. Baker, Bloom & Davis, 2015) is included. This index is relevant for stock market and oil price volatility that results from geopolitics and (energy) policies. The monthly data is retrieved from S. Baker, Bloom and Davis (n.d.) for a period from January 1985 until April 2023.

The transformation of the total dataset is taken from Coulombe (2020). The procedure includes a Principal Component Analysis (PCA) that transform the set of variables into five

uncorrelated principal components that capture the key patterns in the data. The purpose is to twofold. Firstly, the dense information in the principal components make them relevant variables for the HRF model to consider during the splitting procedure. Secondly, the feature weights op the principal components identify the most influential variables that drive macroeconomic variability. Appendix A.2 specifies further details on the decomposition. The results show that volatility, consumer sentiment index, change in NFCI, and the federal funds spreads are important determinants. Several of these variables will be further analysed in Section 5.

4 Methodology

This section discusses this paper's methodology. The hedging objective and forecasting procedure are introduced in Section 4.1 and 4.2, respectively. Section 4.3 details the HRF model and Section 4.4 details the benchmark models. Finally, Section 4.5 elaborates on the performance measures used to evaluate the models.

4.1 Hedging objective

To reduce the downside risks related to their initial equity position, investors seek to hold a portfolio of S&P 500 and Brent oil. For an invested quantity in the S&P 500, the quantity of Brent oil that the investor should invest in to minimise the portfolio variance can be calculated.

Following the derivations of L. H. Ederington (1979), Kroner and Sultan (1993), and Batten et al. (2021), let $r_{j,t}$ denote the monthly log return of asset j at time t, where $j = \{s, o\}$ for the stock and oil index, respectively, and $t = 1, ..., T$. The portfolio return of a portfolio with these two assets can be given by:

$$r_{p,t} = r_{s,t} - \beta_t r_{o,t}, \tag{1}$$

where β_t represents the time-varying hedge ratio. Consequently, the k-step-ahead conditional variance of the portfolio return from Equation 1 is:

$$Var(r_{p,t}|Q_{t-k}) = Var(r_{a,t}|Q_{t-k}) + \beta_t^2 Var(r_{o,t}|Q_{t-k}) + 2\beta_t Cov(r_{a,t}, r_{o,t}|Q_{t-k}). \tag{2}$$

Minimising the k-step-ahead portfolio variance by setting the derivative of Equation 3 equal to 0 results in time-varying minimum variance hedge ratio being:

$$\beta_{optimal,t} = \frac{Cov(r_{s,t}, r_{o,t}|Q_{t-k})}{Var(r_{o,t}|Q_{t-k})}. \tag{3}$$

This equals the conditional, one-step-ahead OLS estimate for the slope coefficient in the following linear relationship between the asset returns at time t:

$$r_{s,t} = \alpha_t + \beta_t r_{o,t} + \epsilon_t, \tag{4}$$

where α_t and β_t are the time-varying intercept and slope coefficients, and ϵ_t is the error term.

In practice, investors are interested in the optimal (minimum-variance) portfolio weights. These can be obtained by dividing the coefficients in Equation 1 by the sum of the coefficients.

For ω_s denoting the portfolio weight of the S&P 500 and ω_o the portfolio weight of Brent Oil, the weighted portfolio return, $r_{P,t}$, is given by:

$$r_{p,t} = \frac{1}{1-\beta_t} r_{s,t} - \frac{\beta_t}{1-\beta_t} r_{o,t} = \omega_{s,t} r_{s,t} + \omega_{o,t} r_{o,t}. \tag{5}$$

4.2 Forecasting procedure

A rolling and expanding window method are used for analysing the out-of-sample performance of the models. In these approaches, periodic model refittings are common. Basher and Sadorsky (2016) model the dynamic hedge ratio between stocks and oil by forecasting 1000 one-step-ahead dynamic conditional correlations using a rolling window approach. To account for periodic changes in the structure of the training set, the GARCH models are refit every 20 observations.

Coulombe (2020) applies the MRF model to forecast quarterly macroeconomic targets, including real GDP and unemployment. He uses an expanding window estimation with direct 1-, 2-, 4-, 6- and 8-quarter-ahead forecasting horizons. He refits the MRF and benchmark models every two years to account for new predictive structures and nonlinear patterns in the data.

Mathematically, let $\hat{\beta}_{t|N} = \hat{F}_N(S_t)$ denote the estimated slope coefficient $\hat{\beta}_t$ at time t. The estimate is obtained from using the information set S_t as input into the specification $\hat{F}_N$. $\hat{F}_N$ denotes the model specification for which the parameters have last been refitted, or re-estimated, in time period N, where $N < t$. In the one-step-ahead rolling window approach, the set S_t is updated for each time period t. Therefore, despite having a model specification that is fixed since time period N, the estimates obtained in each new period are time-varying.

Coulombe (2020) also analyses the time-varying parameters over a period of 25 years when only two MRF refits are performed. His results show that as the time since the last refit increases, the optimal estimated parameter values deviate more from (the credible region of) the optimal parameters. Specifically, out-of-sample structural breaks, like the 2008 Financial Crisis, are difficult to be modelled in this approach. Whereas cyclical behaviour can be modelled, he concludes that the size and level of the variations has evolved exogenously, which forces the MRF to update the set of estimated F repeatedly. This ensures that the (importance of the) nonlinearities in the dataset are constantly re-evaluated.

Given the substantial volatility in the energy market and the presence of several recessions, and potentially structural breaks, this paper follows Coulombe (2020) and Basher and Sadorsky (2016) in evaluating the HRF and benchmark models using a one-step-ahead rolling window forecast in which the models are refit every 24 observations. The estimation window is varied at a fixed rolling window of 120 and 180 days, and an expanding window that starts at 180 days as minimal window. This paper follows Batten et al. (2021) in analysing the performance of the models for a monthly rebalancing period. The results of other rebalancing windows are shown in Appendix B.1.

4.3 Hedging Random Forest

This section outlines the Heding Random Forest (HRF) model that was developed by Van der Bij et al. (2023). The aim of the model is to use macroeconomic variables to estimate optimal

time-varying parameters, while penalising the transaction costs that result from changing parameter estimates between different (consecutive) times. The HRF is designed to create sets of homogeneous leaves to which a time-invariant OLS regression is applied. In each regression, the optimal hedge ratio is estimated for the set of training variables. The linear part allows for better extrapolation of the HRF as compared to a regular random forest models.

The splitting process is as follows. For the sample at each node l, splits are made by selecting the optimal variable S_j from the variable space S to split the sample into two child nodes l_1 and l_2. The threshold value c at which the split is optimal must also be found. The optimal j^* and c^* are determined by adding the weighted sum of squared errors in the right and left child node for a potential splits. Hence, following (Van der Bij et al., 2023), the tree fitting procedure of the HRF in each node l can be represented as:

$$\begin{aligned}\min_{j\in J^-, c\in\mathbb{R}} \Bigg[&\min_{\beta_1} \sum_{\{t\in l_1 | S_{j,t}\leq c\}} \eta(t;\zeta)(r_{1,t} - \alpha_1 - r_{2,t}\beta_1)^2 + \\ &\min_{\beta_2} \sum_{\{t\in l_2 | S_{j,t}> c\}} \eta(t;\zeta)(r_{1,t} - \alpha_2 - r_{2,t}\beta_2)^2 \Bigg],\end{aligned} \tag{6}$$

where the weights, $\eta(t;\zeta)$, introduce a source of regularization that was created by Coulombe (2020). These weights introduce a regularisation source for period t, as a weight of $\zeta < 1$ is placed on observations $t-1$ and $t+1$ and a weight of $\zeta^2 < 1$ for observations $t-2$ and $t+2$. These weights are set to zero during the splitting to reduce computation time.

Equation 6 is iteratively applied to all the created child nodes until a stopping criterion is met. In the leafs, the optimal beta value is then estimated. Estimating a time-varying β_t implies changing the portfolio weights in each period, called portfolio rebalancing. This results in high transaction costs that arise from selling part of the overweight asset and/or buying part of the underweight asset. As a result of this process, investors incur transaction costs, like trading fees and crossing the bid-ask spread. The HRF model smoothens the time-varying β_t estimates in the estimation phase by incorporating transaction costs in a similar way as a ridge or lasso penalty. The specification of this penalty is based on (Hautsch & Voigt, 2019) and uses the L_1-norm to proxy the rebalancing costs. The L_1-norm of the difference between the target β_t in period t and the previous estimated value β_{t-1} penalizes fluctuations of the estimated parameters, hence pulling the β_t estimate towards that of β_{t-1}. The L_1-norm imposes a stronger penalization on turnover than a quadratic penalization, and is hence found to be more realistic (Hautsch & Voigt, 2019). Given that the set of assets is denoted as A, transaction costs are given by:

$$v_{L_1}(\{\omega_{i,t}, \omega_{i,t-1}, c_{i,t}\}_{i\in A}) = \sum_{i\in A} c_{i,t}|\omega_{i,t} - \omega_{i,t-1}|, \tag{7}$$

with $c_{i,t}$ being a cost parameter for asset i at time t. Given that in this context $A = \{s, o\}$, and assuming that the portfolio is fully invested (i.e. $\omega_{s,t} + \omega_{o,t} = 1$) and that transaction costs are constant throughout time t (i.e. $c_{i,t} = c_i$ for $i \in A$), Equation 7 can be simplified to:

$$v_{L_1}(\omega_{s,t}, \omega_{s,t-1}, c_s, c_o) = (c_s + c_o)|\omega_{i,t} - \omega_{i,t-1}|, \tag{8}$$

for $i \in \{o, s\}$.

In the leave nodes, the time-invariant OLS is applied in combination with the transaction cost penalty from Equation 8:

$$\begin{aligned} \min_{\beta} \sum_{t \in l(j,c)}^{N} \eta(t;\zeta)(r_{1,t} - \alpha - r_{2,t}\beta)^2 + \lambda^* v_{L_1}(\omega_{2,t}, \omega_{2,t-1}, c_1, c_2) = \\ \min_{\beta} \sum_{t \in l(j,c)}^{N} \eta(t;\zeta)(r_{1,t} - \alpha - r_{2,t}\beta)^2 + \lambda|\omega_{2,t} - \omega_{2,t-1}|, \end{aligned} \tag{9}$$

where the weights $\eta(t;\zeta)$ are non-zero to ensure smoothing of the estimated β, and λ is tuning parameter for the sensitivity of the transaction costs. Note that λ equals $\lambda^* \times (c_1 + c_2)$, but since both λ^* from Equation 9 and $(c_s + c_o)$ from Equation 8 are constants, this distinction is irrelevant during the tuning. Since this paper considers a cross-asset hedge with only two assets, the OLS regressions in the leaves only have one regressor $r_{o,t}$. Hence, Equation 9 does not contain a ridge penalty.

As multiple time periods t are included in a leaf, there are different time periods $t-1$ that must be known to calculate the transaction cost penalty and determine the optimal value of β_t in that leaf. Hence, the β_t estimates are not determined chronologically. Given the large size and lack of guaranteed convexity, this creates an infeasible optimisation to solve. The heuristic used to overcome this problem is by taking the average estimated β from all periods $t-1$ over the preceding trees. For the first tree, OLS estimates are used. Though the estimates of the first tree will be biased towards OLS, this effect fades as more trees are added. The final random forest contains 250 trees per model. After this number, the results seem to converge.

4.3.1 Variable importance

Understanding what variables impact the time-varying path of the slope estimates is useful in practice. Coulombe (2020) developed specific variable importance measures that are applicable to the MRF and HRF model. [1] Let VI_{OOS} be the out-of-sample variable importance measure that is based on the standard out-of-bag variable importance measure from the random forest literature (Wei, Lu & Song, 2015). The VI_{OOS} is calculated by randomly removing one feature, S_j, from the total set S, and comparing the forecasting accuracy of this model to the model that is based on the full set of variables, S. The difference in overall fit is compared via the Root Mean Squared Prediction Error (RMSPE). $\text{VI}_{\beta_{k,j}}$, another variable importance measure, is calculated in a similar way as VI_{OOS}, but considers how much the path of β_k changes when variable S_j is randomly removed in the forest part. Using these measures, Coulombe (2020) finds that from the total set of predictors S, the number of important variables rarely exceeds more than 3 or 4 variables.

Since the out-of-sample forecasts are based on rolling or expanding windows, the out-of-sample variable importance can only be evaluated for a relatively small number of observations. Consequently, a static approach is applied to find the variable importance. Both the train and

[1] These measures were previously not supported in the HRF package. This paper's package development contributions and an updated Python package are included in the supplementary material to this paper.

test set consist of 50% of the sample. This allows for sufficient out-of-sample observations to evaluate the variable importance. The train set is from June 1990 until November 2006 and the test set from December 2006 until January 2023.

4.3.2 Parameter settings

The hyperparameters λ and ζ in Equation 6 can be tuned. λ is the sensitivity of the HRF to include the transaction cost in the determination of the optimum hedge ratio. ζ denotes the sensitivity of the model to the inclusion of the weighted regularisation, where $\zeta = 0$ means no regularisation. For $\lambda = \text{inf}$, the HRF approaches an OLS model. For $\lambda = \zeta = 0$, the HRF approaches the local linear forest from Friedberg et al. (2020). For $\lambda = 0$, the HRF reduces to the MRF model from Coulombe (2020). The out-of-sample performance of the HRF is evaluated for varying transaction cost sensitivities. The sensitivities that are evaluated are: $\lambda \in \{0, 1, 5, 10, 25, 50, 100\}$. Due to computational constraints, the tuning of ζ is not considered. $\zeta = 0.5$ is used based on Van der Bij et al. (2023) and Coulombe (2020).

4.4 Benchmark models

The performance of the HRF model will be assessed against several benchmark models that are discussed in this section. These include Least Squares estimation (Section 4.4.1) and four multivariate GARCH models for which promising results have been established in the literature. These include DCC-GARCH (Section 4.4.2), ADCC-GARCH (Section 4.4.4), DCC-GJR-GARCH (Section 4.4.3), and GOGARCH (Section 4.4.5). These multivariate GARCH models have been implemented using R, with packages rugarch (Galanos, 2022b) and rmgarch (Galanos, 2022a).

4.4.1 Ordinary Least Squares

As discussed in Section 4.1, the optimal hedge ratio equals the slope estimate obtained from the Ordinary Least Squares (OLS) regression in Equation 4. Hence, this estimation method will be used. The coefficient estimate is time-invariant within a re-fitting window, but will follow a step-wise pattern over the full sample.

4.4.2 DCC-GARCH

Equation 3 implies that the conditional variance and covariance should be estimated to obtain the optimal hedge ratio. Therefore, the Dynamic Conditional Correlation GARCH (DCC-GARCH) model (R. Engle, 2002) is one of the multivariate GARCH benchmark models used. Contrary to the Constant Conditional Correlation GARCH (CCC-GARCH) specification, the DCC-GARCH allows for time varying conditional correlation. The former model is left out as a benchmark, given that the constant correlation assumption does not hold in most cases (Arouri et al., 2011; Chang et al., 2011).

The DCC GARCH model consists of two steps. Firstly, the volatilities are estimated using GARCH. Secondly, the conditional correlations are estimated. The listed derivations follow

R. Engle (2002). Basher and Sadorsky (2016) apply the DCC-GARCH model for stock and oil returns. Based on their results, an AR(1) process for the mean return equation is estimated.

Given $r_t = [r_{s,t}, r_{o,t}]$ is a 2×1 vector containing two asset return series of the S&P 500 index and Brent oil index, an AR(1) process for r_t conditional on the information set Q_{t-1} can be written as:

$$r_t = \mu + \alpha r_{t-1} + \epsilon_t. \tag{10}$$

Following the notations and specifications in R. Engle (2002) and (), the residuals from Equation 10 can be modelled as:

$$\epsilon_t = H_t^{1/2} z_t, \tag{11}$$

where H_t is the 2×2 conditional covariance matrix. This matrix can be expressed in a 2×2 conditional correlation matrix, R_t, and a 2×2 diagonal matrix, D_t, with conditional, time-varying standard deviations, $h_{i,t}^{1/2}$ for $i \in \{s, o\}$, on the diagonal:

$$H_t = D_t R_t D_t, \tag{12}$$

$$D_t = diag(h_{s,t}^{1/2}, h_{o,t}^{1/2}). \tag{13}$$

Given a GARCH(1,1) specification, the conditional variances $h_{i,t}$ can be expressed as:

$$h_{i,t} = \omega_{0i} + \omega_{1i}\epsilon_{i,t-1}^2 + \omega_{2i} h_{i,t-1}, \tag{14}$$

with $i \in \{s, o\}$. Estimating these GARCH(1,1) parameters to obtain conditional volatility estimates is the first step of the DCC procedure. To account for non-normality in the residuals, the DCC is estimated with a multivariate t-distribution.

The second step includes the estimation of the conditional correlations. The matrix R_t from Equation 12 can be expressed in terms of a 2×2 symmetric positive definite matrix, Q_t, which has the conditional (co)variances $q_{i,j,t}$ $(i, j \in \{o, s\}, i \neq j)$ as its elements, and a transformed matrix Q_t^*:

$$R_t = Q_t^* Q_t Q_t^*, \tag{15}$$

$$Q_t = (1 - \theta_1 - \theta_2)\overline{Q} + \theta_1 z_t z_{t-1}' + \theta_2 Q_{t-1}, \tag{16}$$

$$Q_t^* = diag(q_{s,t}^{-1/2}, q_{o,t}^{-1/2}). \tag{17}$$

$\overline{Q}$ from Equation 16 is the 2×2 unconditional correlation matrix of the standardised residuals $z_{i,t}$ (which can be denoted as $z_{i,t} = \epsilon_{i,t}/\sqrt{h_{i,t}}$ following Equation 11). The parameters θ_1 and θ_2 from Equation 16 are non-negative scalar parameters that capture the effect of previous shocks and the effect of previous dynamic conditional correlations, respectively. For $\theta_1 + \theta_2 \leq 1$, the DCC-GARCH is mean reverting. The conditional correlations for the asset returns at time t give the hedging ratio β_t as defined in Equation 3 and are estimated as follows:

$$\rho_{s,o,t} = \frac{q_{s,o,t}}{\sqrt{q_{s,s,t}}\sqrt{q_{o,o,t}}}. \tag{18}$$

4.4.3 DCC-GJR-GARCH

As an extension to symmetric GARCH model in Equation 20, Glosten et al. (1993) create the Glosten Jagannathan Runkle (GJR) GARCH model. By including an asymmetric GARCH effect in modelling individual asset volatility dynamics, the model better accounts for an asymmetric leverage effect. This means that financial markets, equity markets in particular, lose money when uncertainty or volatility rises. In the GJR specification, the conditional volatility of the returns sis given by:

$$h_{i,t} = \omega_{0i} + \omega_{1i}\epsilon^2_{i,t-1} + \omega_{2i}I_{\{\epsilon_{i,t-1}<0\}}\epsilon^2_{i,t-1} + \omega_{3i}h_{i,t-1}, \tag{19}$$

where $I_{\{\epsilon_{i,t-1}<0\}}$ is an indicator function that equals one if $\epsilon_{i,t-1} < 0$ and zero otherwise, for $i \in \{s, o\}$. In the second step, i.e. modelling the conditional correlation dynamics, the regular DCC model is used. These two steps together create the DCC-GJR GARCH model. Using a DCC-GJR specification to find the optimal hedge ratio between stocks and oil returns has shown good in-sample performance (Batten et al., 2021).

4.4.4 ADCC-GARCH

The ADCC GARCH model (Cappiello et al., 2006) is another extension to the DCC GARCH models. It models asymmetric effects in terms of both correlation and volatility. In the first step of modelling the assets' conditional volatilities, the ADCC follows the GJR GARCH specification, i.e. the asymmetric effects are modelled using Equation 19. In the second step, an asymmetric term is added when modelling the dynamic conditional correlations. This is reflected in the correlation evolution matrix, Q_t from Equation 15, by:

$$Q_t = (\overline{Q} - A'\overline{Q}A - B'\overline{Q}B - G'\overline{Q}^{-}G) + A'z_{t-1}z'_{t-1}A + B'Q_{t-1}B + G'z^{-}_t z'^{-}_t G, \tag{20}$$

where A, B and G are scalars, i.e. not asset-specific, asymmetric and smoothing parameters. z^{-}_t are standardised errors for which it holds that $z^{-}_t = max(z^{-}_t, 0)$. Q and Q^{-} are the unconditional correlation matrices of z_t and z^{-}_t, respectively.

4.4.5 GOGARCH

Besides conditional correlation modelling, an alternative way of multivariate GARCH modelling is factor GARCH models (R. F. Engle, Ng & Rothschild, 1990). These models assume that the returns are generated by unobserved underlying factors that are conditionally heteroskedastic. The orthogonal GARCH (OGARCH) model (Alexander, 2001) uses uncorrelated and independent factors. The linear mapping of the factors to observations is orthogonal. Moreover, estimating the covariance matrices from the principal components reduces the dimensionality and computational burden. Van der Weide (2002) generalise the OGARCH model by allowing for non-orthogonal mappings, thereby creating the generalized OGARCH (GOGARCH) model. Given the use of uncorrelated and independent factors, the GOGARCH model is more flexible than other multivariate GARCH models (Basher & Sadorsky, 2016). The GOGARCH model is estimated using the rugarch package in R.

Theoretically, the GOGARCH model (Van der Weide, 2002) has the following specifications. The returns r_t can be modelled as a function of the conditional mean (μ_t) and an error term (ϵ_t) as follows:

$$r_t = \mu_t + \epsilon_t \tag{21}$$

The innovations in Equation 21 are mapped to the unobservable, uncorrelated and independent factors f_t:

$$\epsilon_t = Af_t, \tag{22}$$

where matrix A denotes the linear mapping. The rows of matrix A display the assets and the columns display the factors (f_t) from Equation 22. The matrix A can be decomposed into an unconditional covariance matrix Σ and an orthogonal (rotation) matrix U as given by:

$$A = \Sigma^{1/2}U. \tag{23}$$

Following (Broda & Paolella, 2009) and (Basher & Sadorsky, 2016), the matrix U is estimated using independent component analysis (ICA). The factors f can be specified as:

$$f_t = H_t^{1/2}z_t, \tag{24}$$

where random variable z follows the multivariate affine negative inverse Gaussian (MANIG) distribution, with mean 0 and variance 1 (Basher & Sadorsky, 2016). The unconditional distribution of factors, f_t, satisfy $E(f_t) = 0$ and $E(f_t f_t') = I$. Combining Equations 22, 23, and 24 yields:

$$r_t = m_t + AH_t^{1/2}z_t \tag{25}$$

Thus, the conditional covariance matrix of the error terms can be given as:

$$\Sigma_t = AH_tA' \tag{26}$$

4.5 Performance measures

This section describes the performance measures that are used to compare the models. The discussed measures include the hedging effectiveness (Section 4.5.1), which is based on the Value-at-Risk (Section 4.5.2) and the expected shortfall (Section 4.5.3), and the utility gain of the model portfolio compared to the unhedged portfolio (Section 4.5.4).

4.5.1 Hedging effectiveness

The hedging effectiveness (HE) (L. Ederington, 1979) considers the percent risk reduction from the hedging strategy compared to the unhedged scenario, where a higher HE means the hedge provides a bigger risk reduction. Consistent with this definition, Sukcharoen and Leatham (2017) define the HE at time t as:

$$HE_t = \left(1 - \frac{Risk_{hedged}}{Risk_{unhedged}}\right) \times 100, \tag{27}$$

where $Risk_{hedged}$ denotes a risk measure for the return of the hedged portfolio and $Risk_{unhedged}$ denotes a risk measure for the return of the unhedged portfolio. In this context, the return of the unhedged portfolio is the S&P 500 return and the return of the hedged portfolio can be given by Equation 5.

This paper uses two downside risk measures: the Value-at-Risk (Section 4.5.2), and Expected Shortfall (Section 4.5.3). These risk measures are selected as the downside risk of portfolio returns seems more relevant to investors than the upside risk.

To test for significant differences between the HEs and several other results, a paired t-test is performed (Sukcharoen & Leatham, 2017). It compares the means of two related treatments and tests the null hypothesis of zero mean difference between the two groups against the alternative hypothesis of nonzero mean. Normality of the distributions can be assumed from the Central Limit Theorem with sufficient (208) observations.

4.5.2 Value-at-risk

The value-at-risk (VaR) measures the largest potential loss over a certain period of time for a particular confidence level p. The one-period estimates for the VaR with a confidence level p of the P&L are computed by multiplying the sample standard deviation with the pth percentile of the theoretical distribution (Hull & White, 1998). The sample standard deviation of the P&L is estimated over a one-year backward-looking moving window. The obtained standard deviation estimate is scaled with the p% percentile of the theoretical distribution of the P&Ls. Given the number of observations exceeds 200, the standard normal distribution seems an approporiate approximation (Central Limit Theorem). The sequence of VaR estimates is given by:

$$\hat{VaR}_p = \sigma_{MW=12} \times z_p, \tag{28}$$

where $\sigma_{MW=12}$ is the sample standard deviation of the monthly portfolio P&Ls, and z_p is the pth quantile of the standard normal distribution. Given that the percentile z_q is independent of the sample, the ratio $\hat{VaR}_p$ values equals the ratio of standard deviations. Hence, Equation 27 based on the VaR equals:

$$HE_{\hat{VaR},t} = \left(1 - \frac{\sigma_{MW=12,hedged,t}}{\sigma_{MW=12,unhedged,t}}\right) \times 100. \tag{29}$$

Given the independence of the HE with respect to the selected confidence level, one VaR-based HE value is calculated for all confidence levels.

The Kupiec POF test (Kupiec et al., 1995) is applied to evaluate the accuracy of the measured VaR estimates against actual gains and losses. This process is called backtesting and tests the null hypothesis that the observed violation rate $\hat{\alpha}$ equals the theoretical violation rate α that is suggested by the confidence level. Given a sample of T observations, the test statistic suggested

by (Kupiec et al., 1995) is:

$$\begin{aligned} POF &= 2\log\left(\left(\frac{1-\hat{\alpha}}{1-\alpha}\right)^{T-I(\alpha)}\left(\frac{\hat{\alpha}}{\alpha}\right)^{I(\alpha)}\right) \sim \chi^2(1), \\ \hat{\alpha} &= \frac{1}{T}I(\alpha) = \frac{1}{T}\sum_{t=1}^{T} I_t(\alpha). \end{aligned} \tag{30}$$

where $I_t(\alpha)$ equals 1 if the portfolio return at time t is lower than the VaR_t at a confidence level α. If $\hat{\alpha} = \alpha$, the Kupiec test statistic equals zero, and the null hypothesis is not rejected, i.e. there is no statistically significant evidence that the VaR is inaccurate. If the violation rate $\hat{\alpha}$ differs significantly from α, then the underlying VaR risk measure likely understates or overstates the portfolio's underlying risk level.

4.5.3 Expected Shortfall

The expected shortfall (ES) is the expected loss given that a loss exceeds the VaR. For the dependent variable $r_{s,t}$, the ES at confidence level p is given by:

$$ES_p = -E[r_{1,t}|r_{1,t} \leq -VaR_p]. \tag{31}$$

Combining Equation 27 and 31 for the empirical $\hat{VaR}_p$ gives:

$$HE_{\hat{ES},t,p} = \left(1 - \frac{\hat{ES}_{hedged,t,p}}{\hat{ES}_{unhedged,t,p}}\right) \times 100. \tag{32}$$

In this paper, the $\hat{ES}_p$ is calculated based on the $\hat{VaR}_p$, where $p \in \{0.90, 0.95, 0.99\}$.

4.5.4 Utility

Besides risks, investors are also interested returns and transaction costs. To compare different results in terms of returns, risk and transaction costs, this paper uses Batten et al. (2021) as inspiration to calculate the one-period-ahead utility of each model. The difference is that Batten et al. (2021) uses the variance as a risk measure, whereas the VaR will be used here to focus on downside risk. For portfolio p, let $r_{p,t}$ be the portfolio return at time t, $Risk(r_p)$ be a risk measure at time t, and $T\bar{C}_{p,t}$ the absolute difference in transaction costs between period t and $t-1$. The monthly utility of holding the portfolio at time t is then given by:

$$U_t(p) = r_{p,t} - \gamma Risk(r_{p,t}) - TC_{p,t}, \tag{33}$$

where γ is the sensitivity of the investor to the VaR, i.e. the investor's risk aversion. This papers follows Batten et al. (2021) in using $\gamma \in \{3, 6, 12\}$ to evaluate different levels of risk aversion. A higher value of γ corresponds with more risk aversion. Following Chen and Sutcliffe (2012), transaction costs $TC_{p,t}$ are calculated as the sum of the absolute changes in the dynamic hedge ratios.

5 Results

This section displays the results. Firstly, some preliminary results are discussed (Section 5.1). Section 5.2 reports and analyses the HE results. Section 5.3 analyses the utility results. Finally, Section 5.4 discusses the variable importance of the HRF models.

5.1 Preliminary analysis

The focus of this paper is one-step-ahead, i.e. monthly, rebalancing. Appendix B discusses the robustness of this specification by checking the HE for quarterly, semi-annual, and annual rebalancing. In short, the results seem to be robust to the choice of the rebalancing frequency. Note that the advantage of lower transaction costs in the HRF model decreases as the rebalancing frequency increases. For lower rebalancing frequencies, the potential costs that arise from misspecifying the optimal hedge ratio (i.e. a decrease in portfolio returns and increase in portfolio risk) are more likely to exceed the additional rebalancing costs.

The Kupiec test shows that the null hypothesis that the computed VaR matches the theoretical value can be rejected for several specifications for the 95% and most specifications for the 99% significance levels. Based on these results, the performance will be evaluated using the expected shortfall and the VaR with the 90% significance level.

Multivariate GARCH models are used to capture the serial correlation in the multivariate return series. For a well-specified model, there should be no serial correlation in the squared residuals. Applying the Ljung-Box test for no serial correlation in this rolling window setting is unfeasible. Given that the applied GARCH models have been specified in accordance with previous literature (Batten et al., 2021), the models are assumed to be calibrated correctly.

5.2 Hedging effectiveness

Table 3 illustrates the mean HE values of all benchmark and HRF models, where the HE measures are based on the VaR and ES (90%) hedging objective. Bold indices indicate that that model has the highest hedging effectiveness compared to the other models in their group (Bench or HRF) under an equal estimation window specification.

For the VaR hedging objective, the OLS and HRF models with $\lambda \in \{50, 100\}$ display the highest mean HE values, indicating the most reduction of downside risk. For the ES (90%), the HRF model with $\lambda = 100$ has the highest HE amongst all other HRFs. The ADCC, GOGARCH and DCC have the highest HE from the benchmark models for a 120, 180, and expanding window (respectively).

For all models, the estimation window of 120 observations has a higher HE compared to the window of 180 observations and the expanding window for the VaR objective. Pairwise t-test results for the VaR HE of each model between the different estimation windows are displayed in Table 8. For most benchmark models, the estimation window of 120 observations provides a significantly higher HE than the other windows. This can most likely be attributed to the frequently changing volatility dynamics (Figure 1b). Moreover, Figure 4 shows the that the estimated time-varying correlations between the returns are highly volatile. Therefore, a shorter training window might be preferred to capture more recent volatility patterns that could be

Table 3: Mean HE based on the Value-at-Risk for the benchmark models (Bench), and HRF models with sensitivity $\lambda \in \{0, 1, 5, 10, 25, 50, 100\}$. These results are based on 208 one-step-ahead forecasts with an estimation window of 120 and 180 months, and an expanding window (exp).

Model		Hedging objective					
		VaR			**ES 90%**		
		120	180	exp	120	180	exp
Bench	**OLS**	**3.690**	**2.932**	**2.587**	8.024	6.915	2.626
	DCC	2.152	-2.245	-2.144	7.122	5.807	**6.155**
	DCC-GJR	1.594	-4.806	-3.012	6.625	3.824	3.739
	GOGARCH	3.430	0.447	-1.231	4.244	**8.983**	1.571
	ADCC	2.017	0.430	1.113	**12.185**	1.682	3.665
HRF	**0**	2.683	1.743	0.020	3.973	4.074	1.248
	1	2.202	1.178	-0.749	0.696	3.563	0.158
	5	2.414	2.279	2.079	0.044	1.197	0.926
	10	2.787	2.652	2.410	5.274	2.224	5.381
	25	3.054	2.803	2.421	6.422	5.848	2.296
	50	3.419	**3.346**	2.409	5.911	5.784	2.271
	100	**3.682**	2.931	**2.595**	**8.027**	**6.922**	**2.514**

Note: bold HE values indicate that the model has the highest HE amongst all models that belong to the same group (i.e. Bench or HRF) and are evaluated for the same estimation window.

more relevant for the one-step-ahead forecast.

For the HRF models, the difference between all the three estimation windows is insignificant for the VaR (Table 8) and significant for some models for the ES. This result is unexpected. The HRF for the 180 and exponential window are expected to (significantly) outperform the 120 window, as random forests usually perform better if trained on more data. Figure 2 shows a time series plot of the HE of the HRF model ($\lambda = 100$) for different estimation windows. Three potential reasons for the mentioned results could be identified from the graph. Firstly, the difference between the estimation window lengths might be insufficient to obtain statistically significant differences in the hedge ratios and, consequently, HE. This is partially confirmed by Figure 2, which shows that after 2016, the HE of the rolling windows with 120 and 180 observations closely follow each other. Secondly, the HE of two different windows could follow different paths, but have the same mean HE. Figure 2 shows that around the extreme values, the HEs across models are different. Since the effects for the minima are negated by the maxima, the resulting means of the series are not (significantly) different. Thirdly, the volatile time-varying relationship between the asset returns (Figure 4) might imply that adding data further in the past does not add relevant information. This could be especially because of the short forecasting horizon and short model refit horizon.

t-test results for the 120 window show that the ES HEs are significantly different for most models (Appendix B.3). ADCC shows a strong outperformance compared to all other models, but this result seems relatively unrobust given the low value for larger windows. Again, OLS shows similar performance to the HRF model with $\lambda = 100$.

For further analysis, an estimation window of 120 observations is considered. For this window, a paired t-test is performed between all benchmark and HRF models (Appendix B.3). The

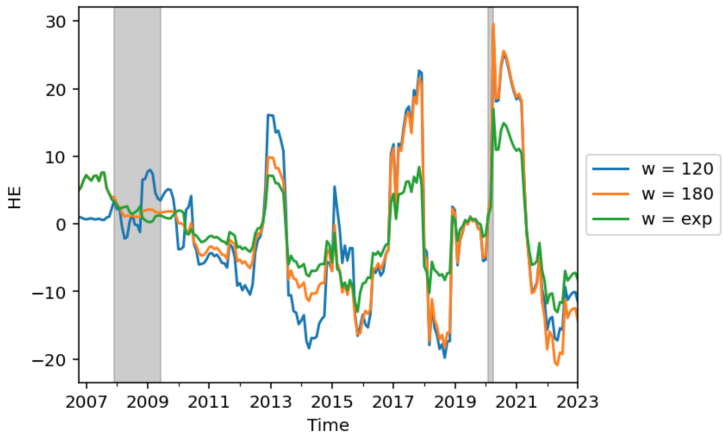
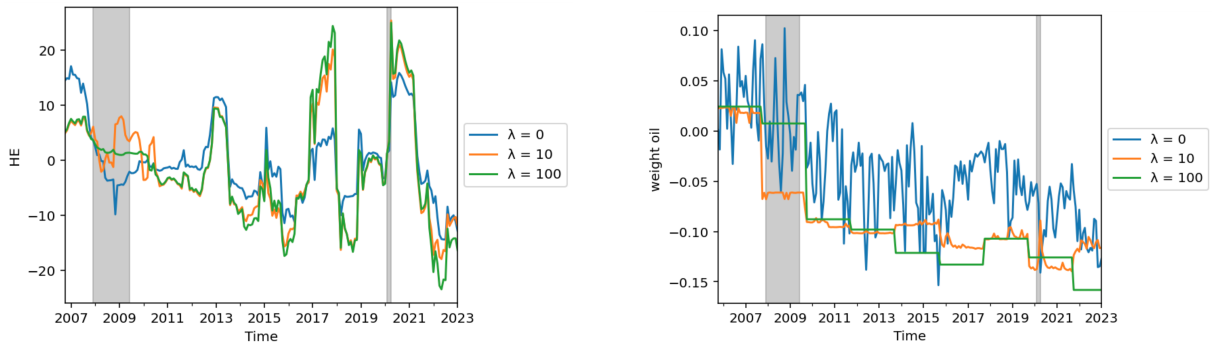


Figure 2: Time series plot of the HE for the HRF models with $\lambda = 100$. The estimation window is varied for an exponential window and a rolling window of 120 months and 180 months. Vertical grey bars indicate the months during which there is an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

results show that none of the models is significantly different from all other models at the 1% and 5% significance level, and only two pairs are significant at the 10% level. Given that HRF models differ insignificantly across λ values, the further analysis is simplified by considering $\lambda \in \{0, 10, 100\}$.

5.2.1 HRF models

The dynamics behind the mean HE for an estimation window of 120 observations are analysed here. Figure 3 shows the time-varying HE and weights for different values of λ. Figure 3b shows that a higher value of λ corresponds with less volatile weights. For the highest sensitivity to transaction costs, $\lambda = 100$, the bi-annual refitting causes a step-wise movement that is comparable to fitting a constant in every refit period. The weights of $\lambda = 10$ and $\lambda = 100$ are closely aligned form 2010 to 2014 and from 2017 to 2021, but show discrepancies between 2008 and 2010.



(a) Hedging effectiveness HRF models

(b) Weights HRF models

Figure 3: Time series plots of the HE and optimal weights of HRF models with $\lambda \in \{0, 10, 100\}$. Monthly rebalancing and an estimation window of 120 months are used. Vertical grey bars indicate the months during which there is an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

During the 2008 Great Financial Crisis, the path of the weights and HE differ between the HRF models. From 2008 to 2010, the hedge weight for $\lambda = 10$ falls to -0.06 at the start of the crisis. This is likely caused by the elevated volatility in S&P prices at the end of 2007 (Figure 1b). For $\lambda = 100$, these dynamics are captured at the next model refit two years later. It seems that the threshold of the $\lambda = 10$ model was exceeded such that the optimal weight was decreased, but that this did not (yet) happen for the $\lambda = 100$, given it considers higher transaction costs. After the second refitting moment in mid-2009, the models reach similar weights. The impact of this deviation on the HE is visible in Figure 3a. During the Great Financial Crisis, the HRF without transaction cost penalty experiences a negative HE. The volatilities of the stock and oil returns were high during this period (Figure 1) and the HRF model tries to capture this volatility by constantly rebalancing the portfolio weights (Figure 3b). However, short and frequent periods of stock rebounds during a period of decreasing returns and economic downturn could cause the weights to be frequently misspecified during the next month, resulting in increased risk exposure. Figure 1c and 3b show that even though oil prices fell sharply after September 2008, i.e. the fall of Lehman Brothers, the HRF with = 0 occasionally fixes weights above 0. A higher transaction cost penalty limits this flexibility. With the portfolio weight of the hedge being close to 0, the HRF with $\lambda = 100$ mimics the unhedged position. This results in an HE around zero during the recession. For $\lambda = 10$, the negative weight of -0.06 first slightly worsens and subsequently improves the downside risk of the portfolio compared to the unhedged position. This likely arises due to the strong positive returns of oil at the start of 2008 and the subsequent negative returns of oil at the end of 2008.

Whereas the mean HE of all HRF models is around 2 (Table 3), Figure 3a shows that there there are several periods for which the HEs sharply rise above and fall below zero. The latter implies that the constructed hedged portfolios have more (downside) risk than the unhedged portfolios. This occurs from 2014 until 2017, during 2018, and after the start of 2021. During the former and latter period, the S&P exhibits stable positive returns with low downside risk. Oil prices during these periods are more volatile, especially towards the upside. Given that oil has negative portfolio weights, this introduces downside risk for the portfolio and results in a negative HE compared to the relatively stable S&P 500. The negative HE in 2018 is caused by the strong positive co-movements of the oil and S&P returns. The negative portfolio weights limit the upside potential and introduce downside risk. The HRF with $\lambda = 0$ captures these dynamics by more frequent rebalancing of the weights to values closer to the optimal weight level. During these periods, the models with $\lambda = 10$ and $\lambda = 100$ only make small adjustments in the weights (Figure 3b), resulting in a lower HE of these models compared to a sensitivity of 0. Hence, the increased model flexibility mitigates the decrease in the HE by making timely adjustments to the weights.

The periods of positive HE peaks occur from the end of 2016 to the start of 2018 and from the start of 2020 until mid 2022. During both time periods, there is a strong positive correlation between the general asset price trends (Figure 1) caused by increased economic growth and reduced oil supplies. The positive asset co-movement combined with the negative portfolio weight of oil reduce potential losses of the S&P 500. This reduces the downside risks during these periods and improves the HE. The HRF model with $\lambda = 0$ mitigates the peaks in HE

compared to the $\lambda = 10$ and $\lambda = 100$ models (Figure 3a). The most flexible model could capture some small fluctuations in the asset trends that cause disturbances in modelling the general macroeconomic co-movement.

From the above analysis, it could be interpreted that sensitivity values between 0 and 10 might result in a more optimal balance. Further analysis in Appendix C shows that intermediate values of λ do not improve both the periods of high and periods of low HE without the other deteriorating.

5.2.2 Benchmark models

The performance of the benchmark models is further analysed in this section. In general, the multivariate GARCH models follow a similar HE path compared to the HRF models, with more extreme minima and maxima.

Figure 4 shows the one-step-ahead time-varying conditional correlations between stock and oil returns as modelled by the multivariate GARCH models. The correlations modelled by the ADCC, DCC, DCC-GJR, and GOGARCH models show similar patterns until 2016. In general, the correlations follow an upward trend until around 2009, followed by a downward trend until 2016. This matches the results reported in Basher and Sadorsky (2016). After 2015, the GOGARCH model shows a different, substantially more volatile, pattern than the three other models.

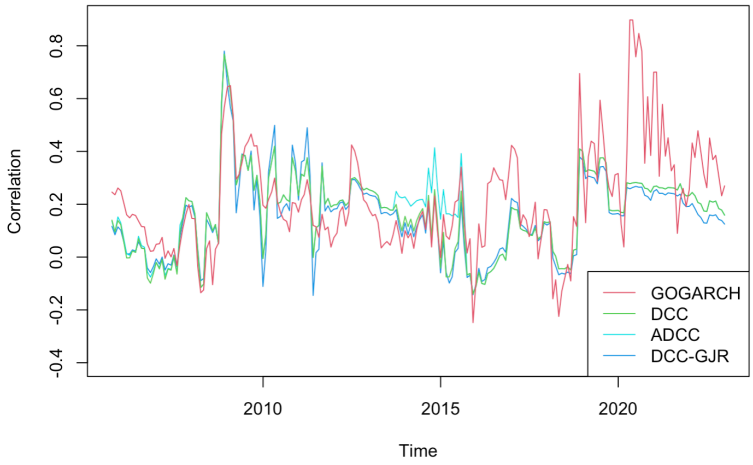
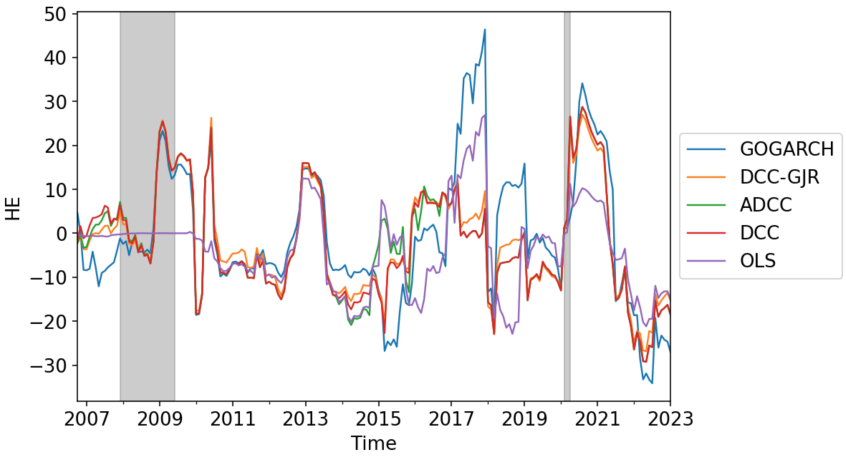
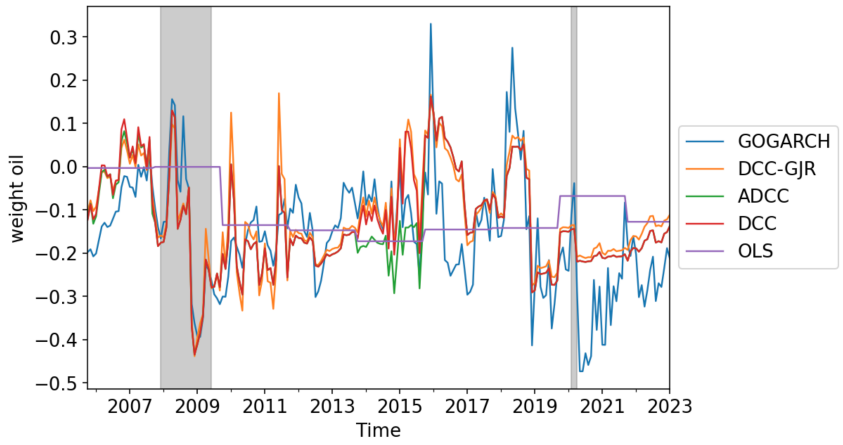


Figure 4: Time series plots of the monthly dynamic conditional correlation as calculated by the multivariate DCC, DCC-GJR, GOGARCH, and ADCC models based on an estimation window of 120 observations. GARCH models are refit once every 24 observations. The (A)DCC and DCC-GJR assume a multivariate t-distribution, the GOGARCH model a MANIG distribution. Different estimation windows are considered in each of the subplots.

During the Great Financial Crisis, the benchmark models show a peak the modelled conditional correlations. This strong positive correlation captures the strong (negative) co-movement between oil and stock prices (Figure 1c). Figure 5 shows the impact of the conditional correlation estimation on the weights and HEs of the models. The positive estimates result in a negative hedge weight and a positive HE during the second half of 2008 and 2009. The OLS estimate of the optimal weight, and the resulting HE, are close to zero during this period.



(a) Hedging effectiveness



(b) Portfolio weights of oil

Figure 5: Time series plots of the HE (Figure 5a) and weights (Figure 5b) of the multivariate benchmark models. An estimation window of 120 months and $\gamma = 3$ are used. Vertical grey bars indicate the presence of an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

Several discrepancies between the multivariate GARCH models can be identified. The correlations of the ADCC and GOGARCH model deviate from the DCC(-GJR) models from 2014 until 2016 (Figure 4). During this period, the S&P 500 exhibits stable positive returns with low downside risk. Oil prices are highly volatile, with a strong oil price decrease at the start of 2014. The DCC and DCC-GJR capture these movements by modelling a weaker conditional correlations that are closer to zero (Figure 4). Based on asymmetries in the returns data and the increase of correlations in highly volatile periods (Table 2, Figure 1b), ADCC could capture the higher volatility of oil returns by estimating a relatively high correlation. In 2014, oil prises sharply fell and S&P 500 prices remained high. The more positive correlation modelled by the ADCC results in a lower HE during 2014. In 2015, the S&P returns exhibit short periods of negative returns, which causes the HE of the S&P 500 to exceed that of the DCC(-GJR) model.

After 2016, GOGARCH models a more volatile and generally higher conditional correlation between the assets, resulting in periods of higher and lower HE compared to the other multivariate GARCH models. During this period, prices of the S&P 500 and Brent oil generally move together, though Brent Oil exhibits some big deviations from the trend at the end

of 2017 and 2020 (Figure 1). GOGARCH computes the time-varying conditional correlations from the estimated individual asset volatilities. DCC-GARCH explicitly models the dynamic conditional correlations by assuming a common volatility structure for the returns. This makes the correlations modelled by the GOGARCH model more flexible in modelling (differences in) the idiosyncratic volatilities of the individual assets. Moreover, several big deviations in the oil price (volatility) could lead the DCC model to estimate lower common correlation patterns. The GOGARCH can model higher correlations given that these big deviations only impact few observations and most observations show a strong co-movement between asset returns. The impact of these differences is illustrated in Figure 5 and show GOGARCH is especially effective in modelling the discussed 2017 and 2020 deviations of Brent oil compared to the other multivariate GARCH models.

Finally, for the period after 2020, the weights and HEs of all benchmark models show similar paths as the paths modelled by the HRF models.

5.3 Utility differences

Table 4 reports the mean monthly percentage returns, mean 90% Value-at-Risk, mean transaction costs, and the mean difference between the utility of the specified model and the utility of the unhedged S&P 500 position (Equation 4.5.4). The unhedged portfolio achieves the highest average return and the second lowest VaR. These results are expected, given that hedging usually limits downside risks at the cost of some upside potentials. Moreover, mean transaction costs are highest for the GOGARCH model, which also follows from Figure 5b. Following Batten et al. (2021), the sensitivity of the utility to the included VaR, γ, is set to 3, 6, and 9, where 3 corresponds to low risk aversion and 9 to high risk aversion.

The utility differences of the multivariate GARCH models and HRF models with $\lambda \in \{0, 1, 5\}$ are mostly negative, whereas utility differences of the other models are mostly positive. This former result follows from the lower return, high transaction cost and similar downside risk levels between the multivariate GARCH portfolios and unhedged portfolio. Increasing γ from 3 to 9 leads to insignificant improvements in the utilities of the multivariate GARCH models (Appendix D.2). As investors become more risk averse, the utility of holding the portfolio from these benchmark models does not improve given the similar VaR levels between the unhedged portfolio and multivariate models. The OLS model and HRF models with $\lambda \in \{25, 50, 100\}$ show positive and significantly increasing utility differences.

Pairwise significances of the utility differences across models are analysed for $\gamma = 6$ using a t-test (Appendix D.2). This value of γ is selected to avoid any assumptions on the risk aversion of investors. Three distinct model groups can be identified. For each of these groups, the utilities differ insignificantly between all model pairs in the group and differ significantly from the other models. The first group consists of the OLS and a HRF with $\lambda \in \{25, 50, 100\}$. The second group consists of the unhedged portfolio and HRF model with $\lambda \in \{1, 5, 10\}$. The final group consists of the four multivariate GARCH models. Based on the results of the t-test, Figure 6 displays the time-varying utilities for the DCC, OLS and HRF with $\lambda = 0$, i.e. one from each of the mentioned groups. Plots of all models are included in Appendix D.1.

Utility differences for the DCC follow the general pattern of the HE from Figure 5a, but

Table 4: The mean return, mean VaR, mean transaction costs (RC), and mean utility differences under $\gamma \in \{3, 6, 9\}$ for the benchmark models ('Bench') and HRF models ($\lambda \in \{0, 1, 5, 10, 25, 50, 100\}$)

Model		Return	VaR	TC	Utility difference		
					$\gamma = 3$	$\gamma = 6$	$\gamma = 9$
Unhedged		0.551	-5.151	0.000	0.000	0.000	0.000
Bench	**OLS**	0.466	-4.963	0.101	0.366	0.930	1.494
	DCC	0.518	-5.130	2.596	-2.623	-2.561	-2.500
	DCC-GJR	0.518	-5.157	2.871	-3.022	-3.043	-3.063
	GOGARCH	0.510	-5.148	4.466	-4.704	-4.696	-4.688
	ADCC	0.476	-5.133	2.489	-2.559	-2.507	-2.455
HRF	**0**	0.498	-5.054	1.365	-1.108	-0.819	-0.531
	1	0.508	-5.061	0.669	-0.442	-0.175	0.093
	5	0.498	-5.075	0.406	-0.252	-0.026	0.200
	10	0.476	-5.066	0.261	-0.100	0.154	0.408
	25	0.467	-5.039	0.083	0.158	0.493	0.827
	50	0.496	-5.017	0.101	0.236	0.638	1.039
	100	0.465	-4.963	0.101	0.365	0.928	1.492

).

frequently exhibit strong downward spikes. These downward spikes are observed during periods with big changes in the weight of oil, i.e. high transaction costs, but decreases or relatively small increases in the downside HE (e.g. in 2008 and 2019).

The utility patterns of the OLS and HRF model with $\lambda = 0$ are substantially less volatile than the patterns of the DCC. The utility paths between 2013 and 2017 show less negative spikes than the time-varying HE, especially for the OLS model. A potential reason is that the oil returns decreased sharply during this period (Figure 1c). The negative portfolio weight of oil results in a positive portfolio return (Figure 5b), which positively impact the time-varying utility during this period.

For all three models, positive utilities can be observed at the end or immediately after an NBER recession (Figure 6).

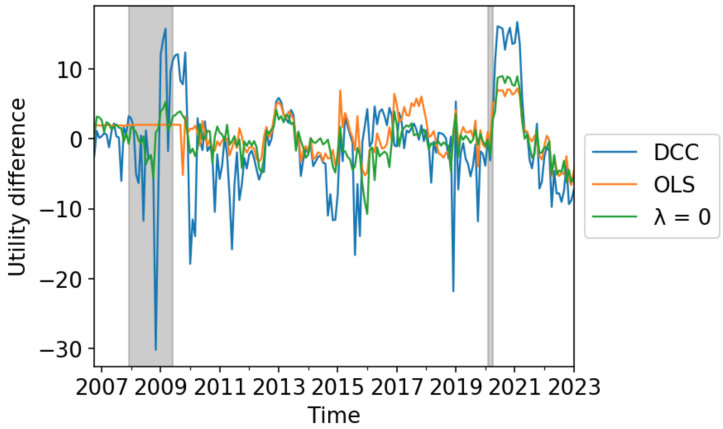


Figure 6: Time series plots of the time-varying realised utilities based on the one-step-ahead forecasts for the DCC, OLS and HRF model with $\lambda = 0$. The estimation window is set to 120 observations and $\gamma = 3$. Vertical grey bars indicate the months during which there is an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

5.4 Variable importance

This section presents and analyses the variable importance results. Table 5 reports the 10 variables with the highest variable importance that are found based on a static forecast with a train and test set that are 50% of the data.

Table 5: Name, description and category of the top 10 most important variables, as given by the HRF variable importance based on the out-of-sample variable importance (VI_{OOS}) and out-of-sample variable importance (VI_{β}) generated by a static forecast with 50% train and 50% test data.

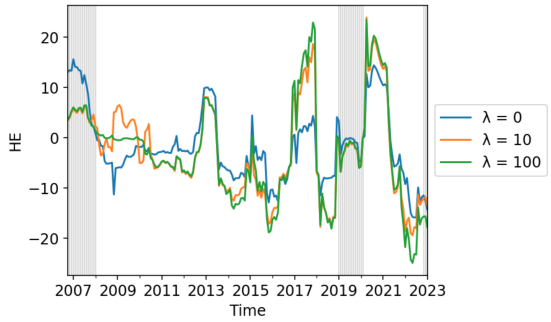
Name	**Description**	**Category**	$\mathbf{VI_{OOS}}$	$\mathbf{VI_{\beta}}$
S&P: indust	S&P's Price Index: Industrials	Stock market	0.173	0.094
S&P 500	S&P's Price Index: Composite	Stock market	0.081	0.076
CONSPI	Consumer credit to Personal Income	Money and credit	-	0.158
M2SL	M2 Money Stock	Money and credit	-	0.083
IPB51222S	Industrial Production: Residential Utilities	Income	0.053	0.071
EXJPUS	Exchange rate JPY US	Exchange rates	0.169	0.087
WPSID62	Producer Price Index: Crude Materials	Prices	0.171	0.087
UEMPMEAN	Average Duration of Unemployment (Weeks)	Labor market	-	0.083
PC1	Negatively related to the spread of Treasuries and corporate bonds minus Federal Funds	Interest rates	0.188	0.087

The categories of the top 10 most important variables show that there are different types of macroeconomic or financial variables that determine the relationship between oil and stocks. The variables together cover all categories that are identified in McCracken and Ng (2016): stock market, money and credit, income, exchange and interest rates, prices and the labor market. This illustrates that the relationship between oil and stocks is complex and can be modelled by a set of mutually reinforcing variables.

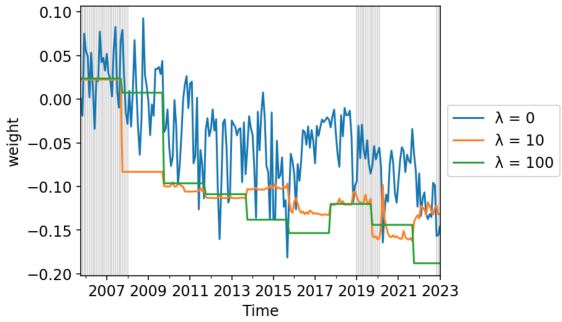
The VI_{OOS} and VI_{β} show that the composite and industrial S&P index, the Industrial Production: Residential Utilities, and Producer Price Index: Crude Materials are important variables for determining the next period hedge ratio. These variables all have strong links to the S&P 500 and Brent Oil index that are analysed in the hedge. Moreover, the average duration of unemployment is strongly related to recessionary periods and show results similar to what has been analysed in Section 5. The remaining variables require more specific considerations.

PC1 denotes the first principal component that was derived in Appendix A.2. The value of the PC is negatively related to the performance of the 10-Year Treasury Constant Maturity Minus Federal Funds Rate (T10YFFM). This refers to the spread between the interest rate on the 10-year U.S. Treasury bond and the short-term overnight interest rate at which banks lend money to each other. Figure 7a and 7b show that increases in the spread, known as inverted yield curves, occur before the 2008 and covid-19 recession. These inverted yield curves are often a signal that economic recessions are expected. Although Figure 7a shows that this does not have a one-directional effect for the HE of the HRF model, it is important for investors to consider this information carefully.

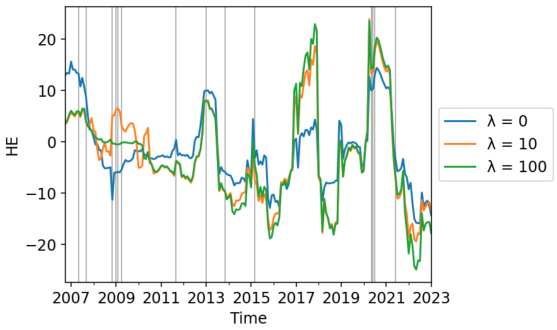
Consumer credit to Personal Income (CONSPI) indicates change in consumer confidence, where a decrease in CONSPI could be caused by fall in consumer confidence. Figure 7e and 7f show the HE and weight as a function of the 5th percentile of periods for which log difference



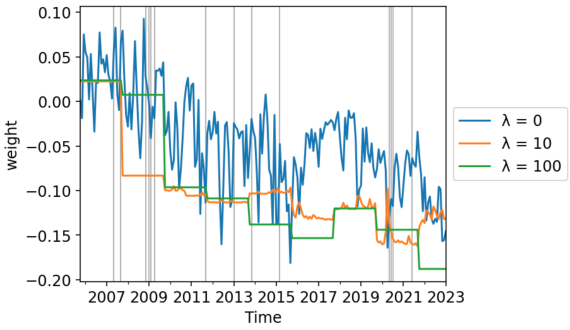
(a) T10YFFM HE, above 80th percentile



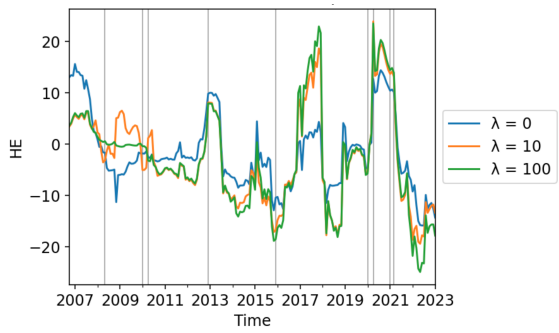
(b) T10YFFM weights, above 80th percentile



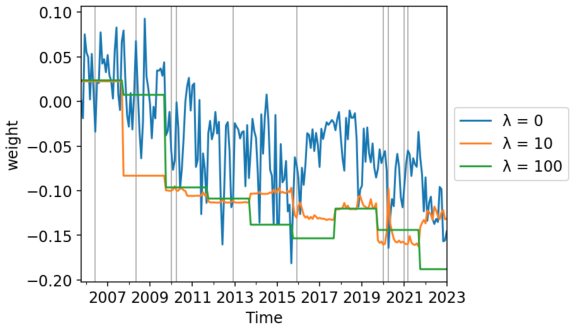
(c) M2 HE, below 5th percentile



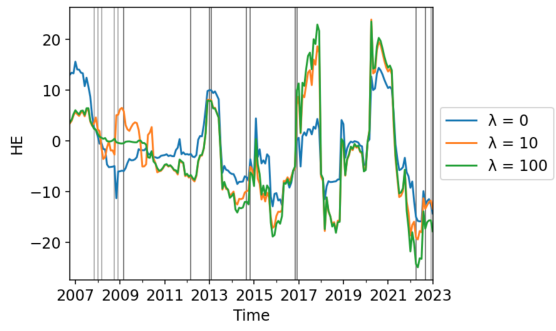
(d) M2 weights, below 5th percentile



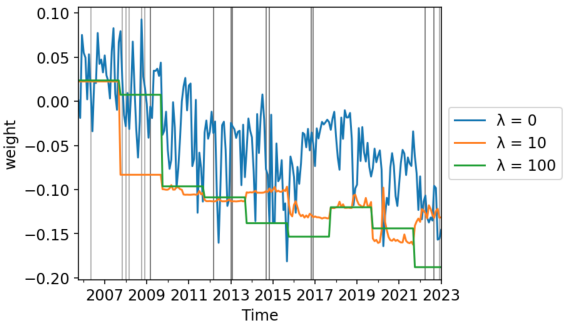
(e) CONSPI HE, below 5th percentile



(f) CONSPI weights, below 5th percentile



(g) JPYUSD HE, below 5th percentile



(h) JPYUSD weights, below 5th percentile

Figure 7: Time series plots of the HE of the HRF models with $\lambda \in \{0, 10, 100\}$. An estimation window of 120 observation is used. Grey bars indicate periods in which the log difference of the displayed variable is above or below the indicated percentile.

of the CONSPI are lowest, i.e. most negative. In 2010, 2013 and 2021, periods of low CONSPI change (indicated by grey bars) are followed by a short period of decreasing HE. This is primarily caused by the volatility of oil increasing compared to that of S&P. One outlier, i.e. the increase after 2016, could be attributed to a low volatility and increasing prices of both oil and stocks. Therefore, CONSPI seems to have a (slightly) negative short-term impact on the variables.

The M2 money supply is a money supply measure that includes currency in circulation and deposits that are easily convertible to cash. A decrease of this money supply measure could be caused by increased savings, decreased borrowing, and contractionary monetary policies. Figure 7d shows that most periods in the lowest 5th percentile of log differences are immediately followed by a more negative oil weight. This is unexpected, as a decrease in M2 money supply is mostly paired with bad macroeconomic performance. The weight of the hedge is expected to increase in such circumstances. Figure 7c shows that periods directly following a negative M2 supply decrease also result in a decreasing HE. Hence, it seems that (extreme) decreases in the M2 supply could indicate that the performance of the HRF could decline in the future. Switching to more conservative weights or the unhedged portfolio could be considered.

Finally, the exchange rate between the Japanese Yen and US dollar is identified as an important variable. The JPYUSD currency pair differentiates itself from other currency pairs because both currencies are relatively safe. Figure 7g shows periods that fall below the 5th percentile of the exchange rate. In analysing the HE after these periods, no one-directional effect is found. A potential reason for this is that the extreme exchange rate changes are often caused by outliers or tail events that might not be directly related to the presented hedge. Rather, the exchange rate could mimic the more general pattern that the economy and S&P follow. E.g. a low JPYUSD exchange rate suggests a strong USD, which is related to positive economic growth, i.e positive S&P returns.

6 Conclusion

This paper evaluates the performance of the Hedging Random Forest (HRF) in forecasting the optimal one-step-ahead time-varying hedging ratios for the S&P 500 and Brent Oil index. The performance is compared to several benchmark models: OLS, DCC-GARCH, ADCC-GARCH, DCC-GJR-GARCH and GOGARCH. One-step-ahead hedge ratios are constructed from October 2006 until January 2023 using a rolling window technique. Different estimation windows are considered. The performance of these models is evaluated using the hedging effectiveness and a measure of utility difference that incorporates returns, risks and transaction costs.

For all models, a smaller estimation window of 120 observations results in a higher HE compared to a fixed window of 180 observations and an expanding window. For the benchmarks, these differences are significant and can likely be attributed to the relevance of more recent data in modelling the volatile correlations. For the HRF models, these differences are not significant.

The HRF model is unable to significantly outperform OLS and benchmark models in terms of only the hedging effectiveness based on the VaR or ES. Lower values of the transaction cost penalty λ result in smoother HE paths, i.e. patterns with higher negative and lower positive extreme values, compared to higher values of λ. The results suggest that relatively high transaction cost penalties should be used in periods with high volatility, whereas low transaction

cost sensitivities should be used in case the hedged asset is less volatile than the hedge. During periods of steady S&P 500 growth and volatile oil prices, the unhedged portfolio outperforms all hedged portfolios and low λ penalties can decrease the weight of the hedge, resulting in lower risks. However, during periods of macroeconomic uncertainty, e.g. the 2008 Great Financial Crisis and covid-19 crisis, low values λ seem too flexible. A potential reason that was identified is the rebound effects for asset prices, leading to frequent misspecifications because the optimal weight estimates are lagged.

HRF models with high λ significantly outperform the unhedged portfolio, HRF models with lower λ values, and multivariate GARCH models in terms of the utility measure. However, HRF models with high λ fail to provide significantly better results compared to OLS. The lower downside risk of OLS and the HRF models leads to an increasing utility difference as risk aversion increases. Multivariate GARCH models have a significantly lower utility than the unhedged portfolio due to high transaction costs and a relatively small improvement of the downside risk.

Finally, analysis of the variable importance of the HRF shows that the relationship between stock and oil returns is caused by a diverse set of macroeconomic and financial variables. Analysis has shown that the JPYUSD exchange rate, M2 money stock and 10-year Treasury spread are important variables for investors to consider.

7 Discussion

Despite the HRF model failing to outperform the simple OLS method in terms of hedging effectiveness and utility, several research limitations and promising directions for future research are outlined in this section.

7.1 HRF application

One-step-ahead hedge ratios are constructed using a varying estimation window, i.e. a rolling window (120 and 180 observations) and expanding window. Models are refit once every 24 observations. The choice for this refit window has been justified in Section 4.2, but its robustness has not been further analysed. Random forests are adaptive and can exhibit a learning process throughout time. Other methods, like OLS, cannot do this. Hence, less frequent refitting and larger forecasting windows could result in an improved HRF performance compared to other benchmark models. A different refit window could be applied in the context of a rolling window or a static window. Moreover, additional estimation windows could be analysed. Although this study is not about finding the optimal estimation length of the rolling window, no significant differences between window of 120, 180 and the expanding window were found. This might suggest that the windows are too similar, such that increasing the window size could result in significant differences and a clearer directional effect for the model performance. In this respect, it could be interesting to consider ways to combine forecasts across estimation windows (Pesaran & Pick, 2011). The Python code for this implementation can be found in the supplementary material to this paper.

This research considers one series of refits that are 24 observations apart. Based on this

refitting, 24 possible series can be generated (each one starting in a different period). Although computing each of these forecast series and taking the average over each of the forecasts for one observation could provide more robust results, this was not researched. An issue that arises is that this approach results in changing OLS estimate during every observation. This does not correspond with the practical implementation of the strategy in practice and is, therefore, not considered.

The stock and oil hedge problem that is studied in this paper has been well studied in previous literature. This paper focuses on this specific problem given the recent social relevance of oil prices that have not (yet) been researched in the literature, and this paper's goal to provide a detailed comparison of the out-of-sample forecasting performance of the new HRF with established models that have previously performed well. An interesting extension would be to relate this problem to clean energy, e.g. by studying the hedge of the S&P Global Clean Energy Index using Brent oil. Another interesting extension would be to hedge stocks with two (or more) assets, e.g. oil and gold. Gold is a safe asset that generally exhibits a low (or negative) correlation with oil returns.

7.2 HRF technicalities

This research compares the performance of multiple transaction cost sensitivities λ during the out-of-sample performance. For different hedges and forecasting designs (e.g. different estimation window, forecasting window, refit frequency), different optimal values of λ can be found. Hence, the reason for analysing different sensitivities out-of-sample is not necessarily to find the specific value of λ that generates the best model fit or HE, neither in a train set, nor in a test set. Different values are analysed to analyse and understand the dynamic paths of the optimal hedge ratio. Hence, although more tuning of λ was not feasible due to computational constraints, this is not considered as a big limitation of this research.

The results of this paper show that different transaction cost sensitivities show different different performances during different times. Although the goal of this research was not to provide a complex ensemble method in which transaction costs are altered based on specific regimes, this could be an interesting avenue of further research.

7.3 Variable importance

The variable importance of the HRF is analysed using the variable importance measures introduced by Coulombe (2020). These measures specifically consider the evolution of the beta estimate rather than the model fit. Given that variable importance does not measure how each variable improves the model's accuracy, some of these variables are further analysed by plotting the time-varying HE as a function of the variables. Other measures, like SHapley Additive exPlanations (SHAP) values, could quantify how each predictor value influences each observation's classification. Attempts to implement this measure and other components (like ALE plots) have not succeeded. Since the HedgingRandomForest class is a custom class, more Python development of the class is required to implement these. This was beyond the scope of this research.

References

Alexander, C. (2001). Orthogonal garch. *Mastering risk*, 21–38.

Arouri, M. E. H., Jouini, J. & Nguyen, D. K. (2011). Volatility spillovers between oil prices and stock sector returns: Implications for portfolio management. *Journal of International money and finance*, *30*(7), 1387–1405.

Aruoba, S. B., Diebold, F. X. & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, *27*(4), 417–427.

Athey, S., Tibshirani, J. & Wager, S. (2019). Generalized random forests.

Baker, S., Bloom, N. & Davis, S. J. (n.d.). *Economic policy uncertainty.* Website: `www.PolicyUncertainty.com`.

Baker, S. R., Bloom, N. & Davis, S. J. (2015). Measuring economic policy uncertainty. no. w21633. *Nat. Bur. Econ. Res.*.

Basher, S. A. & Sadorsky, P. (2016). Hedging emerging market stock prices with oil, gold, vix, and bonds: A comparison between dcc, adcc and go-garch. *Energy Economics*, *54*, 235–247.

Batten, J. A., Kinateder, H., Szilagyi, P. G. & Wagner, N. F. (2017). Can stock market investors hedge energy risk? evidence from asia. *Energy Economics*, *66*, 559–570.

Batten, J. A., Kinateder, H., Szilagyi, P. G. & Wagner, N. F. (2021). Hedging stocks with oil. *Energy Economics*, *93*, 104422.

Bloniarz, A., Talwalkar, A., Yu, B. & Wu, C. (2016). Supervised neighborhoods for distributed nonparametric regression. In *Artificial intelligence and statistics* (pp. 1450–1459).

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

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

Broda, S. A. & Paolella, M. S. (2009). Chicago: A fast and accurate method for portfolio risk calculation. *Journal of Financial Econometrics*, *7*(4), 412–436.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging. *The annals of Statistics*, *30*(4), 927–961.

Cappiello, L., Engle, R. F. & Sheppard, K. (2006). Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial econometrics*, *4*(4), 537–572.

Chang, C.-L., McAleer, M. & Tansuchat, R. (2011). Crude oil hedging strategies using dynamic multivariate garch. *Energy Economics*, *33*(5), 912–923.

Chen, F. & Sutcliffe, C. (2012). Better cross hedges with composite hedging? hedging equity portfolios using financial and commodity futures. *The European Journal of Finance*, *18*(6), 575–595.

Coakley, J., Dollery, J. & Kellard, N. (2008). The role of long memory in hedging effectiveness. *Computational statistics & data analysis*, *52*(6), 3075–3082.

Coulombe, P. G. (2020). The macroeconomy as a random forest. *arXiv preprint arXiv:2006.12724*.

Ederington, L. (1979, 03). The hedging performance of the new futures market. *The Journal of Finance*, *34*,, 157-170. doi: 10.2307/2327150

Ederington, L. H. (1979). The hedging performance of the new futures markets. *The journal of finance*, *34*(1), 157–170.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, *20*(3), 339–350.

Engle, R. F., Ng, V. K. & Rothschild, M. (1990). Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills. *Journal of econometrics*, *45*(1-2), 213–237.

Federal Reserve Bank of Chicago. (2023, 18th June). *Chicago fed national financial conditions index [NFCI].* Retrieved from FRED, Federal Reserve Bank of St. Louis. (`https://fred.stlouisfed.org/series/NFCI`)

Friedberg, R., Tibshirani, J., Athey, S. & Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, *30*(2), 503–517.

Galanos, A. (2022a). rmgarch: Multivariate garch models. [Computer software manual]. (R package version 1.3-9.)

Galanos, A. (2022b). rugarch: Univariate garch models. [Computer software manual]. (R package version 1.4-9.)

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

Hautsch, N. & Voigt, S. (2019). Large-scale portfolio allocation under transaction costs and model uncertainty. *Journal of Econometrics*, *212*(1), 221–240.

Hothorn, T., Lausen, B., Benner, A. & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in medicine*, *23*(1), 77–91.

Hull, J. & White, A. (1998). Value at risk when daily changes in market variables are not normally distributed. *Journal of derivatives*, *5*, 9–19.

Kroner, K. F. & Sultan, J. (1993). Time-varying distributions and dynamic hedging with foreign currency futures. *Journal of financial and quantitative analysis*, *28*(4), 535–551.

Kupiec, P. H. et al. (1995). *Techniques for verifying the accuracy of risk measurement models* (Vol. 95) (No. 24). Division of Research and Statistics, Division of Monetary Affairs, Federal ....

McCracken, M. & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business Economic Statistics*, *34*(4), 574-589. Retrieved from `https://EconPapers.repec.org/RePEc:taf:jnlbes:v:34:y:2016:i:4:p:574-589`

NBER Business Cycle Dating Committee. (n.d.). `https://www.nber.org/research/data/us-business-cycl` (Accessed: June 16, 2023)

of Philadelphia, F. R. B. (n.d.). *ADS: Aggregate Data Sources.* `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/ads`. ([Accessed: June 19, 2023])

Pesaran, M. H. & Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics*, *29*(2), 307–318.

Sukcharoen, K. & Leatham, D. J. (2017). Hedging downside risk of oil refineries: A vine copula approach. *Energy Economics*, *66*, 493–507.

Tan, Z., Yan, Z. & Zhu, G. (2019). Stock selection with random forest: An exploitation of

excess return in the chinese stock market. *Heliyon*, *5*(8), e02310.

Van der Bij, E., Ter Horst, T., Palim, P. & Zegwaart, D. (2023). *Hedging random forest.* (Unpublished seminar work)

Van der Weide, R. (2002). Go-garch: a multivariate generalized orthogonal garch model. *Journal of Applied Econometrics*, *17*(5), 549–564.

Wei, P., Lu, Z. & Song, J. (2015). Variable importance analysis: a comprehensive review. *Reliability Engineering & System Safety*, *142*, 399–432.

Yahoo Finance, x. (2023). *S&p 500 historical data.* Retrieved from `https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC` (Accessed on June 24, 2023)

A Data preprocessing procedure

A.1 Data transformations

Table 6: Transformations of the macroeconomic data as specified by (Coulombe, 2020)

Transformation	Reason	Method
8 lags of r_1	Endogenous SETAR like Dynamics	-
t	Exogenous structural change/breaks	-
2 lags of Fred	Fast-Switching behavior	-
8 lags of 5 PCA's of FRED	Compress cross sectional information ex-ante	Usual PCA
2 MAF's for r_2	Compress lag polynomial information ex-ante	PCA on lags

Table 6 shows the further transformations that have been applied to the data to obtain the final set of regressors S. The first three transformations are combined with FRED data discussed in Section 3. This enlarges the number of regressors in the initial set. Subsequently, five principal components are computed. The specific feature weightings of these components can be found in Appendix A.2. Lags of the principal components are also included. Finally, two Moving Average Factors (MAFs) for the Brent oil return r_o are included. These MAFs compress the information from the many lags of r_o into two components. This ensures that r_o can be summarised using only two features. These two features are included in the existing set of variables to create the final set S.

The usefulness of MAFs is further studied in Goulet Coulombe et al. (2020a) and found to help, mostly with tree-based algorithms.

A.2 PCA feature weights

Table 7 contains the features with the highest and lowest feature weighting for each principal component. A list of abbreviations can be found via McCracken and Ng (2016).

Table 7: The five highest and lowest features and corresponding feature weights from the five principal components that are added to the data. The features for each principal component are listed in ascending order of their weights.

PC: negative HWI		**PC: volatility**		**PC: sentiment**		**PC: Federal Funds Spread**		**PC: negative ΔNFCI**	
HWI	-1.000	UMCSENTx	-0.108	UNRATE	-0.023	BAAFFM	-0.571	Change NFCI	-0.994
Change ADSI	-0.003	Change ADSI	-0.059	NONBORRES	-0.0188	AAAFFM	-0.53	VIXCLSx	-0.03
UMCSENTx	-0.003	AWHMAN	-0.032	Change NBPUI	-0.011	T10YFFM	-0.458	PERMITMW	-0.021
CUMFNS	-0.001	CES060000007	-0.029	PERMITW	-0.011	T5YFFM	-0.329	HOUSTMW	-0.02
UEMPMEAN	-0.001	Change NFCI	-0.025	HOUSTW	-0.011	Change NFCI	-0.076	HOUSTW	-0.018
BAAFFM	0.000	HWI	0.007	AAAFFM	0.055	HOUSTS	0.062	AWHMAN	0.024
CLAIMSx	0.000	UNRATE	0.01	Change ADSI	0.055	PERMITS	0.062	UEMPMEAN	0.026
NONBORRES	0.000	AAAFFM	0.028	BAAFFM	0.071	HOUSTW	0.071	T10YFFM	0.034
UNRATE	0.001	BAAFFM	0.055	VIXCLSx	0.105	PERMITW	0.074	BAAFFM	0.043
VIXCLSx	0.006	VIXCLSx	0.988	UMCSENTx	0.984	UMCSENTx	0.106	AAAFFM	0.0445

B Supplementary material for the HE

This section contains supplementary tables to illustrate the robustness of rebalancing frequency and estimation window specifications.

B.1 HE based on the VaR

Following Sukcharoen and Leatham (2017), a paired t-test is applied to find whether the reported means are statistically (in)significantly different.The VaR-based HEs of the HRF models are all insignificantly different from each other (at the 10% confidence level). The high standard deviations of the HEs (see Section 5.2) likely cause the t-tests to conclude that the mean HEs do not differ significantly between the HRF pairs.

Comparing across the estimation windows of the benchmark models, a rolling window of 120 observations outperforms the one with a size of 180 observations and an exponential window in most scenarios. Except for the GOGARCH and ADCC model (with annual rebalancing frequency), the window of 120 is significantly different from the two other estimation windows. Hence, this window is selected for further analysis. Conclusions on the preferred rebalancing window are slightly more arbitrary. For an estimation window of 120 observations, there are no significant differences between rebalancing frequencies for the DCC and DCC-GJR models. The GOGARCH is best for an annual rebalancing frequency, and the ADCC model is best for quarterly rebalancing. Since

Table 8 reports the HE based on the VaR for the benchmark models and HRF models at different rebalancing windows and different estimation windows. A paired t-test for the hedging effectiveness based on the VaR is performed between all of the model specifications. The paired t-test tests the null hypothesis that the mean HEs of the two provided models are equal. As there are 156 models to compare (13 models times 4 rebalancing windows times 3 estimation windows), the matrix of t-statistics and p-values is included in the supplementary material. From the 3486 HRF model pairs, there are 6 pairs which have significantly different mean HEs (at 10% significance level). For the remaining model pairs, the HEs are not significantly different.

displays the HE values based on the VaR and ES (90%) objective for all HRF and benchmark models, under all considered rebalancing frequencies (monthly, quarterly, semi-annually, annually), and all considered estimation windows (120, 180, expanding).

For the benchmark models, the situation can be further analysed. Comparing across the estimation windows, a window of 120 outperforms a rolling window of size 180 and an exponential window in most cases. With the exception of the GOGARCH model, and the ADCC model for an annual rebalancing frequency, the window of 120 is significantly different than the exponential window and window of 180 observations.

For an estimation window of 120 months, the HE of the DCC and DCC-GJR model do not differ significantly between rebalancing frequencies. The GOGARCH is significantly best for annual rebalancing frequency. The ADCC model is significantly better for quarterly rebalancing compared to monthly and semi-annual rebalancing.

Table 8: Mean HE based on the Value-at-Risk for the benchmark (8a) and HRF models (8b). The rebalancing window (Reb.) is varied between 1 (monthly), 3 (quarterly), 6 (semi-annually), and 12 (annually).The estimation window (Est.) is varied between a rolling window of 120 and 180 months, and an expanding window (exp).

(a) Benchmark models

Reb.	**Est.**	**OLS**	**QTLS**	**DCC**	**DCC-GJR**	**GOGARCH**	**ADCC**
	120	**3.690**	3.405	2.152	1.594	-1.572	2.017
1	180	**2.932**	2.145	-2.245	-4.806	0.447	0.430
	exp	**2.587**	-0.285	-2.144	-3.012	-1.231	1.113
	120	3.690	3.405	3.811	3.017	0.314	**4.429**
3	180	**2.932**	2.145	-2.253	-5.806	0.695	-0.701
	exp	**2.587**	-0.285	-1.878	-2.866	-0.820	1.321
	120	**3.690**	3.405	1.399	1.469	2.090	1.610
6	180	2.932	2.145	-2.963	-6.115	-0.037	-0.095
	exp	**2.587**	-0.285	-3.637	-4.294	-1.982	0.486
	120	**3.690**	3.405	3.154	3.117	3.517	2.689
12	180	**2.932**	2.145	0.809	-0.474	0.988	2.763
	exp	2.587	-0.285	0.720	0.381	1.588	**2.951**

(b) HRF models with parameter settings $\zeta = 0.5$, $r = 0$, and $\lambda \in \{0, 1, 5, 10, 12, 25, 50, 100\}$.

Reb.	**Est.**	**0**	**1**	**5**	**10**	**25**	**50**	**100**
	120	2.683	2.202	2.414	2.787	3.054	3.419	**3.682**
1	180	1.743	1.178	2.279	2.652	2.803	**3.346**	2.931
	exp	0.020	-0.749	2.079	2.410	2.421	2.409	**2.595**
	120	2.685	2.283	2.237	2.704	3.254	3.419	**3.682**
3	180	2.412	1.612	2.206	2.652	2.602	**3.346**	2.931
	exp	1.276	0.115	1.952	2.352	2.421	2.409	**2.595**
	120	1.954	1.726	2.442	2.781	3.245	3.419	**3.682**
6	180	1.189	0.810	2.320	2.674	2.602	**3.346**	2.931
	exp	0.053	-0.286	1.987	2.424	2.421	2.409	**2.595**
	120	2.481	1.879	2.453	2.745	3.241	3.419	**3.682**
12	180	1.742	1.106	2.279	2.603	2.600	**3.346**	2.931
	exp	1.128	0.204	1.913	2.368	2.421	2.409	**2.595**

B.2 HE based on the ES

Table 9 shows the mean HE based on the ES at the 90% confidence level. The paired t-test between the expected shortfalls of the different model specifications shows that none of the hedging effectiveness values are significantly different.

Table 9: Mean HE based on the Expected Shortfall at the 90% confidence level for the benchmark (9a) and HRF models (9b). The rebalancing window (Reb.) is varied between 1 (monthly), 3 (quarterly), 6 (semi-annually), and 12 (annually).The estimation window (Est.) is varied between a rolling window of 120 and 180 months, and an expanding window (exp).

(a) Benchmark models

Reb.	**Est.**	**OLS**	**QTLS**	**DCC**	**DCC-GJR**	**GOGARCH**	**ADCC**
	120	8.024	3.301	7.122	6.625	4.244	**12.185**
1	180	6.915	2.565	5.807	3.824	**8.983**	1.682
	exp	2.626	0.640	**6.155**	3.739	1.571	3.665
	120	8.024	3.301	9.971	8.969	4.910	**12.496**
3	180	6.915	2.565	7.627	5.201	**8.343**	1.266
	exp	2.626	0.640	8.067	**8.153**	2.631	3.529
	120	**8.024**	3.301	1.824	1.756	5.693	6.469
6	180	**6.915**	2.565	1.945	1.421	-0.961	3.843
	exp	**2.626**	0.640	1.167	1.818	0.168	-0.972
	120	**8.024**	3.301	-1.214	0.417	2.501	4.229
12	180	**6.915**	2.565	0.977	-0.023	-1.915	0.108
	exp	**2.626**	0.640	2.517	2.233	-4.468	-0.073

(b) HRF models with parameter settings $\zeta = 0.5$, $r = 0$, and $\lambda \in \{0, 1, 5, 10, 12, 25, 50, 100\}$.

Reb.	**Est.**	**0**	**1**	**5**	**10**	**25**	**50**	**100**
	120	3.973	0.696	0.044	5.274	6.422	5.911	**8.027**
1	180	4.074	3.563	1.197	2.224	5.848	5.784	**6.922**
	exp	1.248	0.158	0.926	5.381	2.296	2.271	**2.514**
	120	5.441	-0.483	-0.386	5.112	7.149	5.911	**8.027**
3	180	5.210	2.820	0.921	2.097	4.261	5.784	**6.922**
	exp	2.174	1.154	0.733	5.260	2.296	2.271	**2.514**
	120	4.206	-1.422	-0.020	5.227	7.136	5.911	**8.027**
6	180	2.116	1.382	1.139	2.185	4.263	5.784	**6.922**
	exp	0.611	0.232	0.900	5.354	2.296	2.271	**2.514**
	120	4.103	1.418	0.089	5.261	7.135	5.911	**8.027**
12	180	1.552	1.298	1.296	2.273	4.264	5.784	**6.922**
	exp	0.567	0.362	0.856	5.313	2.296	2.271	**2.514**

B.3 Pairwise t-test

	180	exp
120	4.707	4.232
180		0.290

(a) OLS

	180	exp
120	13.060***	13.599***
180		0.533

(b) DCC

	180	exp
120	12.760***	11.466***
180		0.290

(c) DCC-GJR

	180	exp
120	5.333	5.446
180		-3.488

(d) GOGARCH

	180	exp
120	7.294**	7.234**
180		-0.150

(e) ADCC

	180	exp
120	1.231	4.079
180		2.822*

(f) HRF, $\lambda = 0$

	180	exp
120	1.762	5.819
180		4.328

(g) HRF, $\lambda = 1$

	180	exp
120	0.211	0.578
180		0.348

(h) HRF, $\lambda = 5$

	180	exp
120	0.137	0.350
180		0.234

(i) HRF, $\lambda = 10$

	180	exp
120	0.237	0.715
180		0.449

(j) HRF, $\lambda = 25$

	180	exp
120	0.074	1.176
180		1.205

(k) HRF, $\lambda = 50$

	180	exp
120	0.739	1.318
180		0.401

(l) HRF, $\lambda = 100$

Figure 8: T-statistics for the pairwise t-tests between the different estimation windows for all studied models. A negative t-statistic indicates that the model specification in the columns has a higher hedging effectiveness. Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively.

Table 10: T-statistics and associated significance level for the paired t-test between the HEs with the VaR objective of all benchmark models and the HRF models with parameter settings $\zeta = 0.5$, $r = 0$, and $\lambda \in \{0, 1, 5, 10, 12, 25, 50, 100\}$. The null hypothesis is that the mean HEs of the specific models are equal. A negative t-statistic indicates that the model in the column has a higher HE.

	DCC	**GJR-DCC**	**GOGARCH**	**ADCC**	**OLS**	**QTLS**	**0**	**1**	**5**	**10**	**25**	**50**	**100**
DCC	0.000	0.442	2.510**	0.104	-1.324	-1.213	-0.493	-0.049	-0.256	-0.541	-0.754	-1.075	-1.316
GJR-DCC		0.000	2.180**	-0.335	-1.871*	-1.835*	-1.053	-0.624	-0.838	-1.053	-1.262	-1.603	-1.861*
GOGARCH			0.000	-2.419**	-3.857***	-3.959***	-3.287***	-3.026***	-3.187***	-3.171***	-3.317***	-3.619***	-3.847***
ADCC				0.000	-1.441	-1.344	-0.618	-0.182	-0.388	-0.657	-0.867	-1.190	-1.433
OLS					0.000	0.334	1.110	1.775*	1.513	0.887	0.608	0.264	0.007
QTLS						0.000	0.981	1.853*	1.512	0.712	0.390	-0.016	-0.324
0							0.000	0.669	0.371	-0.112	-0.389	-0.792	-1.099
1								0.000	-0.333	-0.684	-0.960	-1.411	-1.761*
5									0.000	-0.434	-0.718	-1.159	-1.500
10										0.000	-0.252	-0.609	-0.878
25											0.000	-0.343	-0.600
50												0.000	-0.256
100													0.000

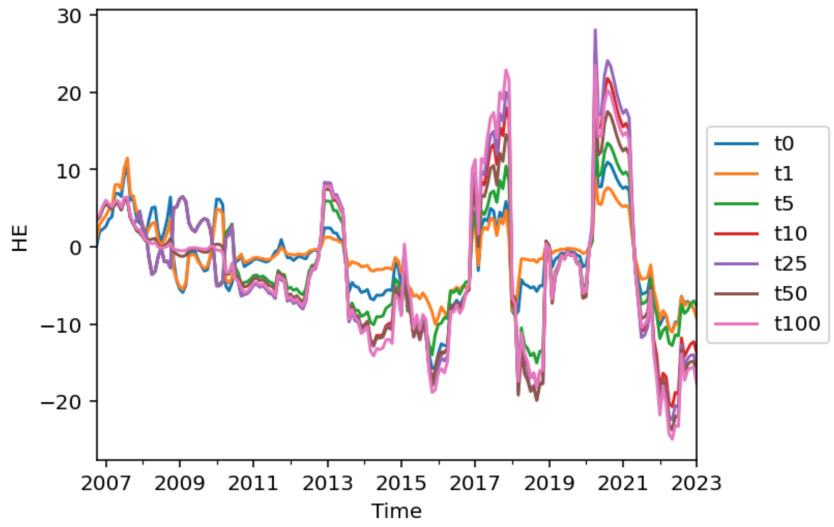
Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively.

Table 11: T-statistics and associated significance level for the paired t-test between the HEs with the ES objective of all benchmark models and the HRF models with parameter settings $\zeta = 0.5$, $r = 0$, and $\lambda \in \{0, 1, 5, 10, 12, 25, 50, 100\}$. The null hypothesis is that the mean HEs of the specific models are equal. A negative t-statistic indicates that the model in the column has a higher HE.

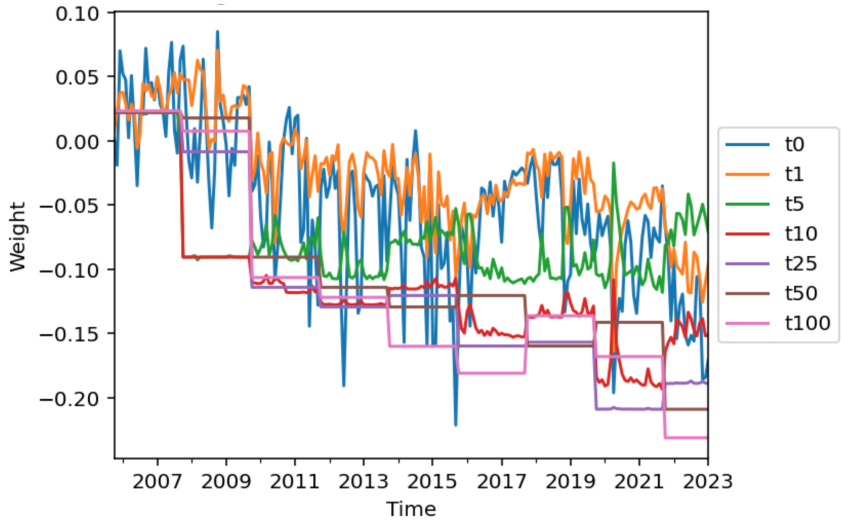
	DCC	**DCC-GJR**	**GOGARCH**	**ADCC**	**OLS**	**0**	**1**	**5**	**10**	**25**	**50**	**100**
DCC	0.000	0.027	-2.024	-10.430***	-13.139***	-4.939***	-2.350***	-4.840***	-10.347***	-12.664***	-5.407***	-13.208***
DCC-GJR		0.000	-2.053***	-10.468***	-13.180***	-4.971***	-2.380***	-4.872***	-10.386***	-12.705***	-5.440***	-13.249***
GOGARCH			0.000	-8.311***	-11.024***	-2.884***	-0.321	-2.783***	-8.240***	-10.541***	-3.350***	-11.093***
ADCC				0.000	-2.848***	5.422***	7.998***	5.531***	0.029	-2.311	4.941***	-2.918***
OLS					0.000	8.180***	10.719***	8.291***	2.862	0.548	7.703***	-0.068
0						0.000	2.566	0.103	-5.366***	-7.678***	-0.471	-8.249***
1							0.000	-2.465	-7.929***	-10.233***	-3.033	-10.788***
5								0.000	-5.475	-7.789***	-0.575	-8.360***
10									0.000	-2.327	4.888	-2.931
25										0.000	7.198***	-0.617
50											0.000	-7.772***
100												0.000

Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively.

C Supplementary graphs for HE based on VaR



(a) Hedging effectiveness



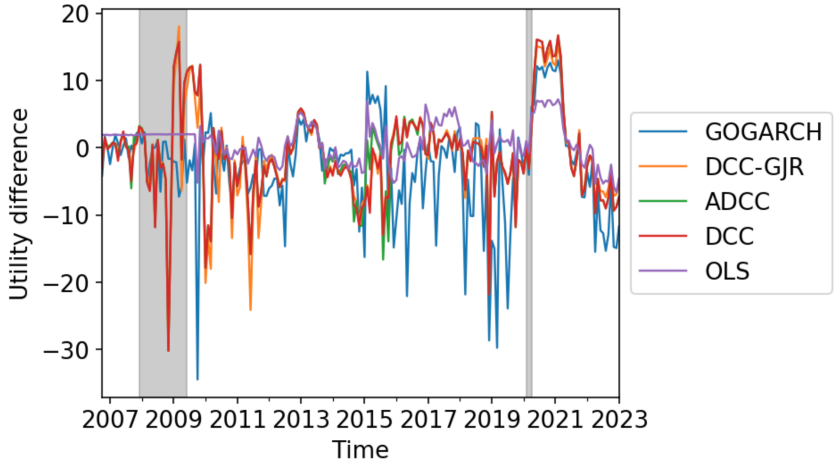
(b) Weights

Figure 9: Time series plots of the hedging effectiveness (Figure 9a) and weights (Figure 9b) of all HRF models. An estimation window of 120 observations is used.

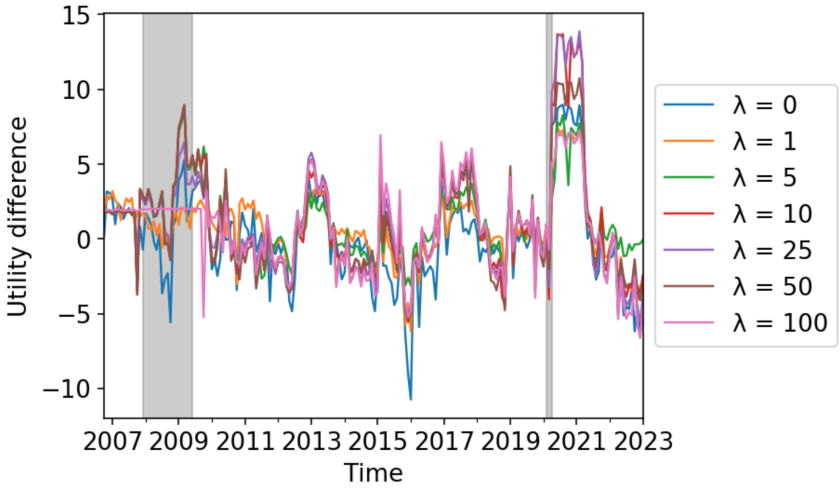
D Utilities

D.1 Time-varying utilities for all models

D.2 Pairwise t-tests



(a) Benchmark models



(b) HRF models, $\lambda \in \{0, 10, 100\}$

Figure 10: Time series plots of the realised utilities based on the one-step-ahead forecasts for the benchmark models (Figure 10a) and HRF models (Figure 10b). An estimation window of 120 observations and risk sensitivity of 3 are used. Vertical grey bars indicate the months during which there is an NBER recession (*NBER Business Cycle Dating Committee*, n.d.).

Table 12: T-statistics and associated significance level for the paired t-test between the utility differences of all benchmark models and all studied HRF models. The null hypothesis is that the mean HEs of the specific models are equal. A negative t-statistic indicates that the model in the column has a higher utility.

	DCC	**DCC-GJR**	**GOGARCH**	**ADCC**	**OLS**	**0**	**1**	**5**	**10**	**25**	**50**	**100**	**unhedged**
DCC	0.000	0.222***	2.639	-0.131	-2.755***	-1.531***	-2.997***	-3.577***	-3.442***	-3.463***	-3.324***	-2.754***	-5.681***
DCC-GJR		0.000	2.442	-0.352	-3.075	-1.834***	-3.331***	-3.916***	-3.752***	-3.773***	-3.640***	-3.074***	-6.072***
GOGARCH			0.000	-2.752	-5.912***	-4.737***	-6.218***	-6.741***	-6.433***	-6.445***	-6.373***	-5.909***	-8.838***
ADCC				0.000	-2.561***	-1.350***	-2.794***	-3.371***	-3.253***	-3.275***	-3.131***	-2.560***	-5.438***
OLS					0.000	2.080***	-0.298***	-1.431***	-1.333**	-1.375	-1.091	-0.002	-6.233***
0						0.000	-2.583**	-3.617***	-3.119***	-3.145***	-2.979***	-2.078***	-8.656***
1							0.000	-1.298	-1.191	-1.237**	-0.923***	0.295***	-7.478
5								0.000	-0.201	-0.256*	0.128**	1.425***	-5.241
10									0.000	-0.048	0.284	1.330**	-3.159
25										0.000	0.332	1.371	-3.047*
50											0.000	1.087	-3.891***
100												0.000	-6.203***
unhedged													0.000

Note: ***, ** and * denote the statistical significance at the 1%, 5% and 10% level, respectively.