Erasmus University Rotterdam

Thesis MSc Econometrics and Management Science

Erasmus School of Economics

July 19, 2023

Probabilistic Forecasting of Stock Returns using Distributional Regression Tree-Based Models

Authors

Saskia de Jong (471244)

Supervisor

Dr. O. Kleen (ESE)

Second Assessor

Dr. M. Zhelonkin

Abstract

Distributional forecasting is gaining traction in the financial research area over the recent years, as it already did for some other research areas. Probabilistic forecasts of returns can be used in an economic structure to for example maximise utility functions in investment allocation. This research compares distributional forecast methods with the Continuous Ranked Probability Score for the one-step-ahead forecasts of continuously compounded returns of the S&P500 index. Several classic parametric benchmark methods are considered: Historical Simulation and ARMA-GARCH models with and without an extra explanatory variable. These are compared with recently developed regression tree-based models: Distribution Forests of Schlosser et al. (2019) and Distributional Adaptive Soft Trees of Umlauf and Klein (2022). This research also initiates the combination of these tree-based models with a Beta-transformed Linear Pool. The Distributional Adaptive Soft Tree with underlying Student t distribution performs best out-of-sample.

Keywords: Distributional forecasts, Probabilistic forecasts, Regression Trees, Distribution Forest, Distributional Adaptive Soft Trees, Continuous Ranked Probability Score, Beta-transformed Linear Pool, Ensemble, continuously compounded returns, ARMA, GARCH, Historical Simulation, Certainty Equivalent Rate

Contents

1 Introduction

One of the most researched topics, and a very relevant one in finance, is in what assets one is going to invest. What every investor has in common is that their investment choices depend on what movements the asset returns are going to make in the future. Therefore accurately forecasting return values in any form, particularly point-forecasting of the return value, is the subject of a lot of researches. However, investors are not rational and will not simply invest in the asset with the highest expected point-forecast. Investors also take into account a sense of risk that they preferably avoid. This theoretically results in agents all having their own utility functions dependent on their specific risk aversion. Some of those utility functions only depend on what value such an asset will take in the future, and for those a point-forecast-method suffices. However, research in finance towards distributional forecasts is gaining traction. These forecasts do not only predict the expected value of this asset, but predict the full distribution of that asset in a future point in time. These distributional forecasts of asset returns can then be used to decide on the investment allocation of more types of agents, and their utility functions. In order to do that properly, these distributional forecasting methods should be as accurate and reliable as possible.

The modelling of the future density of a dependent variable is a topic for which interest is growing in a lot of research areas besides the financial research areas. All areas in which not only future values are of interest, but the corresponding prediction intervals as well, are using these types of models. Several recent examples of these probabilistic research subjects are: the chance of natural hazards happening like regional floods (Kiran and Srinivas, 2021), the minimisation of the chance of running out of stock in stores (Ulrich et al., 2021), the forecasting of prediction intervals of intra-day electricity prices (Klein et al., 2023) and uncertainty around health infections like seasonal influenza (Brooks et al., 2018).

Within finance there are numerous subjects for which distributional forecasting models are used as well. It can for example be used for forecasting the government bond yield spreads (Mikis et al., 2022), forecasting the stock market liquidity with high frequency data (Luo et al., 2013) and forecasting financial variables like inflation and output growth (Greenwood-Nimmo et al., 2012). This research however looks at probably one of the most intensively researched subjects in finance: continuously compounded returns.

There are multiple ways to form these probabilistic forecasting models. A widely used method, because it is very simple to implement and low-maintenance is Historical Simulation: empirical non-parametric distributions are made out of sorting all available historical observations. On the other hand, parametric models are used more often even-though they take more computation time. A much used method is the ARMA-GARCH type of model. The GARCH model, introduced by Engle (1982) and Bollerslev (1986), is often used to model the volatility parameter. Moreover, many alterations of the original GARCH model are explored in distributional forecasting: for example asymmetric GJR-GARCH (as done by Abadir et al. (2022)), TGARCH (as implemented by Cai and Stander (2019)) and EGARCH (as done by Hoogerheide (2012)). Next to that, the idea of implementing external economic explanatory variables in the variance equation was the fundament of the GARCH-X model (Apergis, 1998). When this volatility forecasting method is combined with the ARMA method of Box and Jenkins (1976), one can obtain distributional forecasts. Recently Yao et al. (2023) implemented explanatory variables in the ARMA-GARCH model to forecast the distribution of Chinese stocks.

In addition to the many analytical parametric models available, a recent interest has been drawn towards the possibility of using machine learning regression trees to form distributional models. Classical regression trees are used for point-forecasting, and therefore a few adjustments had to be developed in order to make them appropriate for distributional forecasting. This movement origins from wanting to use the benefits of tree-based methods, such as automatic selection of explanatory variables and allowing for intricate interactions between them. Moreover, the model is formed in a computationally inexpensive way. Schlosser et al. (2019) introduce the first Distribution Forest, based on a parametric counterpart: Generalized Additive Models for Location Scale and Shape (GAMLSS) of Rigby and Stasinopoulos (2005). The idea is that every distribution parameter that has to be forecasted is explained by explanatory variables, not only the point-forecast. This Distribution Forest method has been developed for precipitation forecasting, but is already used in some other research areas (e.g. Vasseur and Aznarte (2021), Khatin-Zadeh et al. (2023)). It has not yet been employed in financial forecasting often, and that is what our research is going to do.

The previously mentioned forests are based on multiple regression trees, of which each of them are formed based on strict decision rules of the explanatory variables. All trees therefore impose a step-function, which is then smoothed out by combining the trees. In point-forecasting, a new territory of research developed concerning soft decision rules in regression trees. The first soft splitting rule to

grow a tree was incorporated by Ciampi et al. (2002), to overcome the problem of obtaining only a nearly smooth function with hard splitting rules. Isroy et al. (2012) introduce the first multivariate soft splitting rule. Umlauf and Klein (2022) decided to combine the tree-based distributional idea of Schlosser et al. (2019) with a soft splitting rule for the smoothness of the function. This method is called Distributional Adaptive Soft Trees, and is shown to work better for probabilistic solar cycle forecasting. This research will see if this result holds in financial forecasting.

Over time there has been a lot of discussions if economic variables are able to make financial forecasts more accurate, and if so, what economic variables. To fully make use of the advantage of tree-based methods in selecting explanatory variables, this research implements a big number of economic explanatory variables. Welch and Goyal (2008) made a list of economic variables that had extensively been researched in notable papers. This research incorporates almost all of their economic variables, in line with the approach of other papers (Rapach et al., 2010; Campbell and Thompson, 2008). The three factors of Fama and French (1993) are implemented as well. This research will examine which of these explanatory variables have most explanatory power in our set-up.

Multiple density forecasts can also be combined. Those are called ensembles or pools. Combining multiple models can have advantages as some models interpret signals in a different way than other models. One could just take an equal weight combination model, but those lack calibration (Gneiting and Ranjan, 2013). Ranjan and Gneiting (2010) introduced a beta transformation of a linear combination of models. In this way, weights can be assigned to the separate models according to their forecasting power. To the best of the authors knowledge, there has not yet been made an ensemble of multiple distributional tree-based models. This research is the first to propose this ensemble model.

When incorporating all these types of distributional models, one has to decide on a performance measure to compare them. This is however different than a performance measure in point-forecasting. Gneiting and Raftery (2007) elaborate on what criteria a scoring rule should enclose. A scoring rule for a distributional model has to be proper and should motivate a researcher to be honest. The Continuous Ranked Probability Score (CRPS) of Matheson and Winkler (1976) satisfies those criteria and is therefore used to compare the models in this research.

The main question that this research aims to answer is: *Do (combined) distributional tree-based methods improve the one-month-ahead investment choices of investors compared to choices made with simple distributional (non-)parametric models?* In order to be able to answer this question, an investment set-up has to be clarified. This research roughly follows the set-up of Zhao (2013), meaning that an investor can only invest in one risky asset and the risk-free rate. Investors are imposed to have power utility functions (CRRA), for which full distributional forecasts are needed to optimise investment allocations. The risky asset that is considered in this research is the monthly continuously compounded return of the S&P500 index. Whether or not better distributional forecasting methods improve the investment choices is measured in the difference of the Certainty Equivalent Rate of return (CER) between the formed portfolios with the best tree-based and benchmark method.

The sub-questions that help to answer the research question are formulated as follows:
- *What simple distributional (non-)parametric model performs best in terms of average out-of-sample Continuous Ranked Probability Score?*
- *Do Distributional Adaptive Soft Trees perform better in terms of average out-of-sample Continuous Ranked Probability Score than Distributional Forests in this financial set-up?*
- *What explanatory variables drive the results of the tree-based methods?*
- *Does the proposed ensemble of tree-based methods perform better than the separate tree-based methods in terms of average in- and out-of-sample Continuous Ranked Probability Score?*

The simple distributional (non-)parametric models, alias the benchmarks, that are used in this research are the Historical Simulation, ARMA-GARCH and ARMA-GARCH-X models. The tree-based models considered are Distributional Forests, Distributional Adaptive Soft Trees and a Beta-transformed Linear Pool of those. The models are formed using the train data-set, of which the dates range from January 1970 to December 2014. Then their performance is compared for the test data-set that ranges from January 2015 to December 2022.

The research is structured as follows: In Section 2 the data is further elaborated on and their sources are named, Section 3 provides a methodological explanation of all our distributional models implemented as well as the scoring rule and the investment optimisation, Section 4.1 contains the results of the distributional forecast methods and Section 4.2 displays the results of the value of the investment choices. Section 5 draws a conclusion and puts forward future research ideas.

2 Data

The variable of interest in this research is the continuously compounded return of the monthly S&P500 index (y_t). These returns are computed following this formula:

$$y_t = 100\ln\left(\frac{p_t}{p_{t-1}}\right) \qquad \text{for} \quad t \in (1, 636) \tag{1}$$

Where p_t is the price of the S&P500 index of month t, calculated as the average of the daily prices[1] during that month. The monthly return data used ranges from January 1970 up until December 2022, resulting in 636 observations. This period is split into a train data-set (January 1970 up until December 2014; 540 observations) and a test data-set (January 2015 up until December 2022; 96 observations). This split is used during the whole research.

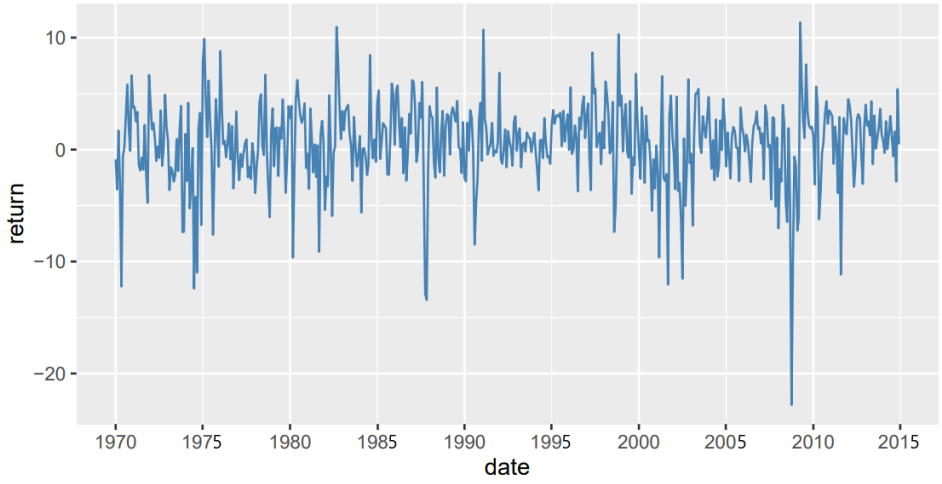


Figure 1: **Historical continuously compounded returns.** This figure shows the monthly continuously compounded returns of the S&P500 index for the train data-set, ranging from January 1970 until December 2014.

In Figure 1, the continuously compounded returns over the train data-period are displayed. As one can see, those returns take both positive and negative values and seem to be centered around zero. In Table 1, the descriptive statistics are reported. Around 2008 the biggest drop can be seen in the return values, which makes sense as the Global Financial Crisis took place there.

[1]Recovered from Wharton Research Data Services: CRSP Index File on the S&P500

Table 1: Descriptive statistics of the continuously compounded returns

	Mean	Std. Dev.	Max	Min
y_t	0.577	3.713	11.352	-22.804

Note: The table reports the descriptive statistics of the monthly continuously compounded returns of the S&P500 index over the train data-set, ranging from January 1970 until December 2014.

When showing the values of the return train data-set against how often those values occur, one obtains a histogram shown in Figure 2. The line shown in the figure is the normal distribution with corresponding mean and standard deviation incorporated. As one can see the data has a somewhat more negative skew than the normal curve, and it seems to be more leptokurtic.

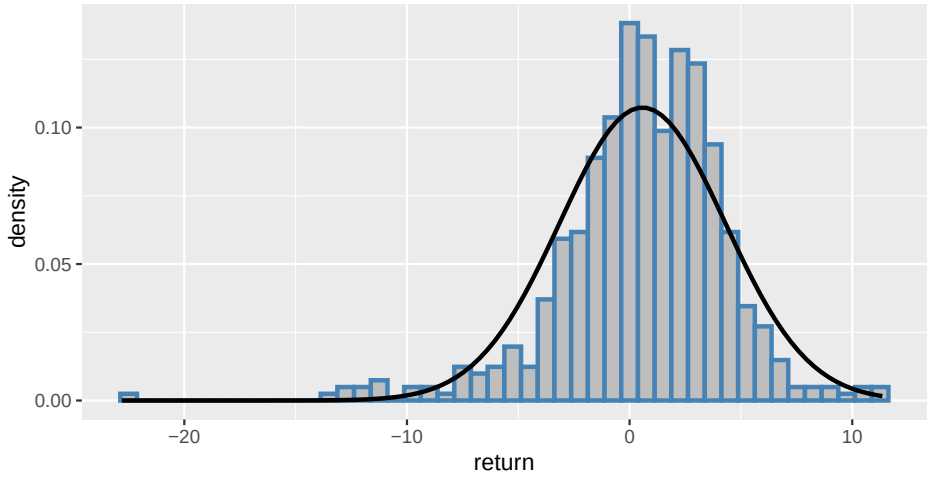


Figure 2: **Histogram of continuously compounded returns.** This figure shows a histogram of the values of monthly continuously compounded returns of the S&P500 index for the train data-set, ranging from January 1970 until December 2014. Next to this histogram, a normal distribution with the mean and standard deviation of the data-set is plotted

In selected models, explanatory variables are included. Here we will elaborate on how those variables are formed. In general, we incorporated the economic variables of Welch and Goyal (2008) as well as the three notorious Fama&French factors from Fama and French (1993). Some of these explanatory variables are composed using other variables which are not included as explanatory

variables themselves. These are specifically the dividends and the earnings[2]. The explanatory variables incorporated are the following:

- Dividend Price Ratio (d/p): the difference between the log of dividends and the log of prices(p_t)
- Dividend Yield (d/y): the difference between the log of dividends and the log of lagged prices(p_{t-1})
- Earnings Price Ratio (e/p): the difference between log of earnings and log of prices(p_t)
- Dividend Payout Ratio (d/e): the difference between log of dividends and log of earnings
- Book to Market Ratio (b/m)[3] the ratio of book-to-market value of the Dow Jones Industrial Average
- Net Equity Expansion ($ntis$)[3]: the twelve-month moving sums of net issues by NYSE listed stocks divided by the total market capitalization of NYSE stocks
- T-bills Rate (tbl)[3]: Interest rate on a secondary market 3-Month Treasury Bill
- Long Term Yield (lty)[3] : Long-term government bond yields
- Long Term Rate of Return (ltr)[3]: Long-term government bond returns
- Term Spread (tms)[3]: the difference between lty and tbl
- Default Yield Spread (dfy)[3]: the difference between corporate bond yields rated BAA and AAA
- Default Return Spread (dfr)[3]: the difference between the return on long-term corporate bonds and ltr
- Inflation ($infl$)[3]: the one month lagged Consumer Price Index, as it is released at the end of the month
- Size Premium ($FF1$) [4]: The first Fama&French factor, also known as Small minus Big (SMB)
- Book-to-Market value Factor ($FF2$)[4]: The second Fama&French factor, also known as the Value Premium / High minus Low (HML)
- Excess Return on the market Factor ($FF3$)[4]: The third Fama&French Factor, also known as

[2]Both the Dividends and the Earnings are retrieved from Robert Shillers website: http://www.econ.yale.edu/ shiller/data.htm

[3]These monthly variables are obtained from the updated data sheet published on Amit Goyal's website, https://sites.google.com/view/agoyal145, and therefore have the same data sources as used in Welch and Goyal (2008)

[4]Factors retrieved from Kenneth R. French's website under *"Changes in CRSP data"*, (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html)

sensitivity to the market

- Sum of Squared Returns (ssr): a proxy of the variance, the sum of squared daily returns on the S&P500 index. This is calculated by taking the average of the squared daily returns and multiplying it with 21 to take into account the different lengths of trading months.

In Table 6 in Appendix C.1, an overview is shown of the explanatory variables and their abbreviations which are used throughout this paper.

3 Methodology

This section explains all mathematical theory used in this research. To answer our research question, this research uses a non-parametric and several parametric distributional forecast methods, combination methods and performance evaluations. First the benchmark models are elaborated on; Historical Simulation and several selected ARMA-GARCH models. Then the tree-based methods of Schlosser et al. (2019) and Umlauf and Klein (2022) will be theoretically substantiated, which are named the Distributional Forest method and the Distributional Adaptive Soft Tree method respectively. Thereafter, a Beta-transformed Linear Pooling method is incorporated to combine these tree-based methods. We will subsequently explain the scoring rule that measures the predicting performances of all single models together with the tests that are implemented to see whether the differences are significant. The scoring rule implemented is named the Continuous Ranked Probability Score. All models are constituted using only the train data-set, and compared out-of-sample using the test data-set with one-step-ahead forecasts. The models are not recalibrated intermediately.

After obtaining the best probabilistic tree-based forecasting method, a set-up to compare its benefits against the best benchmark method in an economic application is shown. Following Zhao (2013), a portfolio is formed of the risk-free rate and the asset of interest, here the S&P500 index. The weights of the portfolio are formed with the provided distributional forecast models and the Certainty Equivalent Rate of return is compared to analyse the additional value of a better forecasting model.

The programming of all models has been done in R, version 4.3.0. For an overview of all coefficients and variables used, look at Table 7 in Appendix C.2.

3.1 Historical Simulation

As a benchmark method, we will incorporate a widely used non-parametric model which is called Historical Simulation (e.g. Hendricks (1996)). This method computes an empirical CDF using historical observations. First the dependent variable, the monthly continuously compounded returns, is sorted in order of magnitude:

$$y_1 < y_2 < ... < y_{J-1} < y_J \tag{2}$$

In which J is the number of observations in the train data-set. Then, the empirical CDF is computed as follows:

$$F^{HS}(y) = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_{[y_j < y]} \tag{3}$$

This comes down to counting the number of observations that are lower than y in the historical data, divided by the total number of observations in the data at that point. As our research concerns one-step-ahead data, every month an observation is added to the ordered list in Formula 2. The advantage of this method is that it does not make assumptions about the distribution. For small samples however, every added observation makes a big difference in estimation, making this method less robust.

3.2 ARMA-GARCH method

For the formation of point-forecasts of a financial time series, an Autoregressive Moving Average (ARMA) model is often implemented (Box and Jenkins, 1976). This model combines previous values of the dependent variable (y_t) with the previous white noise disturbance terms (ϵ_t). In mathematical notation the ARMA(p, q) model is denoted as follows:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \upsilon_j \epsilon_{t-j} + \epsilon_t \tag{4}$$

In which α is a constant term that is included to allow for a non-zero mean, p the number of previous dependent variables and q the number of previous disturbance terms included. The AR and MA coefficients, ϕ_i, υ_i, are to be estimated. We will use an ARMA(1,1) model.

To form a distributional forecast with the ARMA model, an additional part is needed to forecast the volatility of the time series. A method often combined with it is the General Autoregressive Conditionally Heteroscedastic (GARCH) model of Bollerslev (1986), which was originally devised for modelling heteroscedasticity in timeseries. Its property to be able to directly forecast the volatility

parameter is used when forming probabilistic forecasts. To keep the computation simple we only consider lags of one in the volatility modelling, resulting in a GARCH(1,1) model as shown in Formula 5. These lag orders are shown to capture the dynamics accurately (Bollerslev, 1986).

$$\begin{aligned} \epsilon_t &= z_t \sigma_t \\ \sigma_t^2 &= \psi + \chi \epsilon_{t-1}^2 + \zeta \sigma_{t-1}^2 \end{aligned} \tag{5}$$

In which σ_t^2 is the conditional variance, ψ the long-run volatility and z_t the innovation; an independent and identically distributed randomized variable with a mean of zero and a variance of one. The coefficients of the GARCH and ARCH terms, χ, ζ, are to be estimated.

This model allows for time varying terms in the disturbance term ϵ_t. The GARCH model implies symmetry between the amplitude of the previous error term and the current volatility, regardless of the sign of the previous error. It therefore assumes that our time series respond the same to negative and positive shocks. Even though this assumption is not always correct for return data, this distinction is less important when using less-frequent data, like monthly data.

When adding an explanatory variable for accuracy in the variance modelling, the model is called a GARCH-X model. We will incorporate the sum of squared daily returns (ssr) as an explanatory variable in our monthly GARCH(1,1) model as a proxy for the realized volatility. This GARCH(1,1)-X model is shown below:

$$\begin{aligned} \epsilon_t &= z_t \sigma_t \\ \sigma_t^2 &= \psi + \chi \epsilon_{t-1}^2 + \zeta \sigma_{t-1}^2 + \xi ssr_{t-1} \end{aligned} \tag{6}$$

For all GARCH models, the distribution that is implemented for z_t is the assumed distribution of the GARCH residuals. Often a standardized normal distribution is assumed when implementing a GARCH(1,1) model as it fits the restrictions for the mean and variance of z_t. Another distribution that is shown to work better when there is leptokurticity in the data (i.e. fatter tails in the residuals) is the standardized Student t distribution. Bollerslev (1987) introduces this distribution in the GARCH method and shows that this non-normality works for leptokurtic financial data.

Both distributions are incorporated as they are represented in Yaya et al. (2014). The associated distribution formulas (f) and log-likelihoods (ℓ) for the randomized variable z_t for both models (ARMA-GARCH$^{\text{normal}}$ & ARMA-GARCH$^{\text{student-t}}$) are shown in Formula 7. Then all of the parameters, including the degrees of freedom ν, are estimated by maximising the corresponding log likelihood function using the train data-set.

$$\text{ARMA-GARCH}^{\text{normal}}: \quad f(z_t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z_t^2\right)$$
$$\ell(z_t) = -\frac{1}{2}\left(N\log(2\pi) + \sum_{t=1}^{N} z_t^2\right)$$

$$\text{ARMA-GARCH}^{\text{student-t}}: \quad f(z_t,\nu) = \Gamma\left(\frac{\nu+1}{2}\right)\left[\Gamma\left(\frac{\nu}{2}\right)\right]^{-1}[\pi(\nu)]^{-\frac{1}{2}}\left(1+\frac{z_t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$
$$\ell(z_t,\nu) = -\frac{1}{2}\left\{N\log\left(\frac{\pi(\nu-2)\Gamma(\nu/2)^2}{\Gamma((\nu+1)/2)^2}\right) + (\nu+1)\sum_{t=1}^{N}\log\left[1+\frac{z_t^2}{\nu-1}\right]\right\} \tag{7}$$

In which $\Gamma()$ is the gamma function. This model is incorporated using the *rugarch* R-package by Ghalanos (2022).

3.3 Distributional Forests

The method of setting up a Distributional Forest to create probabilistic forecasts is invented for the usage of precipitation forecasting by Schlosser et al. (2019). The concept of a Distributional Forest is to use tree-based learning methods to asses the relation of the dependent variable with different explanatory variables, instead of using a classic parametric model to capture this. It changes the conventional regression trees, which are used to form point forecasts, into a distributional form by adding the ideas of Generalized Additive Models for Location, Scale and Shape (GAMLSS) modelling. The fundament of the GAMLSS model of Rigby and Stasinopoulos (2005) is that for all location, scale and shape parameters of a given distribution, the possible relations with explanatory variables are examined. To make the distributional model more smooth, a number of trees are combined into one Distributional Forest. This method can therefore be seen as a distributional-allowing random forest.

Simultaneously, the model of Schlosser et al. (2019) adopts multiple benefits of random forests. Theoretically it is able to capture the sudden abrupt changes that our return data contains, as well as the non-linear smoothness in the movements. Probably the biggest benefit however, is that the selection of all explanatory variables is done by the model itself in a relatively quickly way, while also considering intricate interactions between the explanatory variables. In our research we do not know if and which of the numerous variables have explanatory power yet, much less about their

dependencies. Therefore this method is a better fit for this problem than for example the parametric GAMLSS model itself.

Because of the latter benefit, we incorporate all possible explanatory variables discussed in Section 2, enumerated in Table 6. Two lags of the dependent variable are added as explanatory variables as well: the value of a month ago and the value of a year ago to take into account possible seasonality (y_{t-1}, y_{t-12} respectively). To evaluate which variables drive the results of the forest, the variable importance is assessed in hind sight by breaking all associations in the formed forest with each separate explanatory variable. Thereafter the difference in performance score will be analysed. This is done over the train data-set, as the forest is built over that data.

One of the fundamental choices of this method is on which distribution family the Distributional Forest is based. In the research of precipitation there are no negative values allowed so Schlosser et al. (2019) implement a left-censored Gaussian distribution. This research deviates from this approach as our time series of simple returns do take negative values. We apply two non-censored distributional families to form their respective Distributional Forests: the Gaussian distribution family (NO) and the Student t distribution family (stT). Both distributions have their own associated log-likelihood formulas $\ell^k(\theta^k; Y)$ where $k \in \{\text{NO, stT}\}$. The estimation of their distribution parameters θ^k on the train data-set is often done by using Maximum Likelihood Estimation (MLE). A goodness-of-fit measure for an individual observation with a set distribution parameters is found when using the MLE first order condition for one observation, and is called the score function in the framework of Schlosser et al. (2019):

$$s(\theta; y_t) = \frac{\partial \ell}{\partial \theta}(\theta; y_t) \tag{8}$$

However, the idea of the distribution tree is that only one distribution function with a defined set of distribution parameters θ is not sufficient to capture the data. The explanatory variable space is split into B segments, forming total segment space $\mathcal{B}_b$ with $b = 1, ..., B$. All segments have their own local distributional parameters (θ^b). The way these segments are partitioned is by assessing scores of possible partitions recursively:

1. In the current sub-sample, estimate $\hat{\theta}$ using MLE

2. For each possible explanatory variable, test for associations of the scores $s(\theta; y_t)$ and that explanatory variable.

3. Split that sub-sample based on the explanatory variable with the strongest association. The splitting value of the variable is obtained by calculating which split results in the greatest difference in log-likelihood. The associations on which the variable is chosen can be computed in multiple ways, where our research follows Schlosser et al. (2019). This means that the permutation tests of Hothorn et al. (2006) are incorporated.

4. Repeat the above steps for all sub-samples until the tree is full-grown or a stopping criteria is reached.

Stopping criteria that can be incorporated are minimum association with an explanatory variable, minimal segment size (set to 10) and minimal observations to perform a split (set to 30). As pre-pruning is not recommended by Hastie et al. (2001), we did not incorporate a minimum association criteria.

For the formation of the Distributional Forests, H of those distributional trees are trained on subsets of the covariates in the train data-set and afterwards combined. The number of explanatory variables in such a subset is in our research set equal to the rounded up square root of the number of all possible explanatory variables as recommended by Schlosser et al. (2019), so $\lceil\sqrt{19}\rceil$. By combining the trees, the imposed steps of the individual decision trees are smoothed out to match the smooth dependent data. As more trees do not lead to overfitting (Biau and Scornet, 2015), we use $H = 1000$. This is also driven by the conclusion of Probst and Boulesteix (2018), that setting H larger as long as its computational feasible and classical error measures are considered, it is better.

After combining the trees, making predictions with this optimised forest is not as simple as sending the new values of the covariates (x^{new}) down the tree anymore. The Distributional Forest method of Schlosser et al. (2019) incorporates averaged nearest neighbour weights from all trees for all train observations as follows:

$$w_t(x^{new}) = \frac{1}{H}\sum_{h=1}^{H}\sum_{b=1}^{B^{(h)}}\frac{\mathbb{1}\left((x_t \in \mathcal{B}_b^h) \wedge (x^{new} \in \mathcal{B}_b^h)\right)}{|\mathcal{B}_b^h|} \tag{9}$$

In which $B^{(h)}$ is the number of segments in tree h, $|\mathcal{B}_b^h|$ the number of observations in section b of tree h, t the observations of the learning sample and $\mathbb{1}$ an indicator function. The weights from Formula 9 take a value between 0 and 1.

The actual predicted distributional parameters then become:

$$\hat{\theta}(x^{new}) = argmax_{\theta \in \Theta} \sum_{t=1}^{n} w_t(x^{new}) \cdot \ell(\theta; y_t) \tag{10}$$

These parameter estimates give the distributional forecast wanted. The method is incorporated in this research using the R-packages *partykit* (Hothorn and Zeileis, 2015), *disttree* (Moritz et al., 2019) and *GAMLSS* (Rigby and Stasinopoulos, 2005)

3.4 Distributional Adaptive Soft Trees

A new method that follows up from the previously described Distributional Forests are Distributional Adaptive Soft Trees, shaped by Umlauf and Klein (2022). Instead of making the predicted distribution smooth by combining multiple trees, this method incorporates a multivariate soft split rule to make an initial smooth tree. The earlier mentioned method used hard splits that lead up to imposed steps. Umlauf and Klein (2022) argue that the combination of multiple trees only finds a nearly smooth function at best. A soft splitting rule means that the step function is replaced by a soft discriminator without an exact split point. The soft splitting rule that they have proposed is the AdaSoRT, which improves the point-forecasts of soft regression trees. This method is elaborated on in Appendix A.

Just like Schlosser et al. (2019) altered the GAMLSS model of Rigby and Stasinopoulos (2005) to form distributional forecasts with hard splitting rules, Umlauf and Klein (2022) altered the same GAMLSS model with their AdaSoRT splitting rule to allow for distributional forecasts.

In GAMLSS, a distribution family is fitted for distributional forecasting by modelling all distribution parameters in vector θ. These distribution parameters, $\theta_1, ..., \theta_k$ are each connected with an explanatory-variable predictor with known connector functions. These explanatory variable predictors (η_k) can be formed using the AdaSoRT (Appendix A), and are then given as:

$$\eta_k = \eta_k(X_k; \beta_l, \Omega_k) = N_k(X_k, \omega_k)\beta_k \tag{11}$$

In which $N_k(\cdot)$ is the design matrix of the AdaSoRT, Ω_k is the set of weights associated with the paths in this AdaSoRT, X_k is the matrix containing the explanatory variables and β_k the matrix containing the node values. The values of the variables in matrices Ω_k and β_k have to be estimated for all k, so for all distribution parameters. These are found following the in-detail described theoretical road map of Umlauf and Klein (2022), which is summarized below:

1. Intercepts β_{0k} are initialized and predictors are set equal to them: $\eta_k^{(0)} = \beta_{0k}$

2. Soft split the root nodes for all k models: Estimate a set of optimal weights $\omega_k = \{\omega_{1k}\}$ with Maximum Likelihood determining the first split and resulting in the design matrices $N_{1k}(X_k, \omega_k)$. Then split the design matrix in small two column matrices per split by introducing a new index $c = 1, ..., \frac{G_k}{2}$ with G_k the number of nodes in the tree for distribution parameter k: $N_k(X_k, \omega_k) = (1, N_{1k}(X_k, \omega_k), ..., N_{\frac{J_k}{2}k}(X_k, \omega_k))$. Then define a score factor $u_k = \partial\ell(\beta, \omega; y, X)/\partial\eta_k$ and working weights $W_{kk} = -\text{diag}(\partial^2\ell(\beta, \omega; y, X)/\partial\eta_k\partial\eta_k'$.
   Then update the predictor by using Formula 12 in which t is set to 1 for the root split.

$$\begin{aligned} \eta_k^{(t+1)} &= \eta_k^{(t)} + N_{1k}(X_k, \omega_k)\beta_{1k} \qquad \text{in which} \\ \beta_{1k} &= (N_{1k}(X_k, \omega_k)'W_{kk}N_{1k}(x_k, \omega_k) + \rho I)^{-1}N_{1k}(X_k, \omega_k)'u_k \end{aligned} \tag{12}$$

   Where ρI is to ensure numerical stability as explained by Umlauf and Klein (2022).

3. Grow the tree by doubling all columns of the design matrix as in Step 2. This means that the weights and corresponding parameters are $\omega_k = \{\omega_{1k}, \omega_{2k}, \omega_{3k}\}$ and β_{2k}, β_{3k}. Calculate the latter with Formula 12. Take note: in this step you only update the explanatory-variable predictor with the best soft split of the two in terms of log-likelihood.

4. Update the weights by using Maximum Likelihood: use the first and second order derivatives of the log-likelihood with respect to the weights. To avoid overfitting, a shrinkage term is included in the form of $\lambda_k\omega_{rk}'\omega_{rk}$. The smaller one sets this λ_k, the closer the tree stays to the actual data which could result in overfitting. In our research we follow the heuristic approach of setting this hyper-parameter at first very high, and decreasing it until the tree is continuously growing with low AIC differences. This resulted in a value of $\lambda = 10$.

Repeat step 3 and 4 until convergence is reached, meaning that there is no improvement in terms of the Akaike Information Criterion (Akaike, 1974) when growing the tree further.

In line with our approach for the formation of the Distributional Forests (Section 3.3), we again incorporate all possible explanatory variables discussed in Section 2, enumerated in Table 6. The two lags of the dependent variable (y_{t-1}, y_{t-12}) are added as well. The same data-manipulation is operated to evaluate the variable importance in these trees over the train data-set.

In their research Umlauf and Klein (2022) also make forests of their Adaptive Soft Decision Trees,

however they do not elaborate on them much. As they conclude that one tree works better for data with less than 2000 observations, and our data contains less, we also only incorporate one tree. In their research, one soft tree outperforms the full Distributional Forests of Schlosser et al. (2019).

The method is incorporated in this research using the R-packages *softtrees* (Umlauf, 2022)) and *GAMLSS* (Rigby and Stasinopoulos, 2005).

3.5 Ensemble forecasts

Combining multiple distributional forecast methods forms an ensemble forecast method. This ensemble method takes advantage of the positive aspects of different methods. For example, when one method is a very conservative density estimator, and the other is very aggressive, one could try to balance those out. There are multiple ways to form these ensemble methods. One could decide to simply take the average of different methods, which implies an equal weight ensemble. Over time multiple ways to determine and optimise those weights were formed, also in non-linear ways. Ranjan and Gneiting (2010) introduce a beta transformation on a linear pooling system and call this method the Beta-transformed Linear Pool (BLP). This BLP uses the CDF of the beta density with shape parameters α and β as follows:

$$F_{\alpha,\beta}(y) = B_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right) \tag{13}$$

In which $F_1, ..., F_k$ are the cumulative distributions of the different methods that you are combining, and $w_1, ..., w_k$ are the corresponding non-negative weights that all have to sum up to 1. Bogner et al. (2017) use this formula to derive the following PDF function of the BLP:

$$f_{\alpha,\beta}(y) = \left(\sum_{i=1}^{k} w_i f_i(y)\right) b_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right) \tag{14}$$

Where f_i are the density functions of the different methods and $b_{\alpha,\beta}$ the PDF of the beta density with its shape parameters. To find the optimal ensemble method, the parameters in this formula $(\alpha, \beta, w_1, ..., w_k)$ are estimated by maximizing the corresponding log-likelihood function over the train data-set. This log-likelihood function is displayed in formula 15, and is derived as in Bogner et al. (2017).

$$\begin{aligned}\ell(w_1, ..., w_k : \alpha, \beta) &= \sum_{j=1}^{J} \log(f(y_j)) \\ &= \sum_{j=1}^{J} \log\left(\sum_{i=1}^{k} w_i f_{ij}(y_j)\right) + \sum_{j=1}^{J} \log\left(b_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_{ij}(y_j)\right)\right)\end{aligned} \tag{15}$$

In which J is the number of observations in the train data-set. After optimising the parameters, one obtains a pooled density forecaster. We will use this pooling system to combine the tree-based methods.

3.6 Statistical Evaluation of Forecasts

To asses the performance of the distributional forecasts, one should implement a scoring rule. A scoring rule tests the performance of the full distribution forecast, not only the performance of the point forecast. As Gneiting and Raftery (2007) argue, a scoring rule should be proper to be a good comparison resource between predicting models. The scoring rule which will be used to compare the one-step-ahead model forecasting performance on the test data-set is the mean Continuous Ranked Probability Score (CRPS). This scoring rule is theoretically attractive and satisfy the requirements of Gneiting and Raftery (2007).

The CRPS is introduced by Matheson and Winkler (1976), and is proposed to calculate a performance score of a distributional forecast F compared to the actual value y. This score is computed as follows:

$$CRPS_t(F, y_t) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}(y_t \leq z))^2 dz \tag{16}$$

In which $\mathbb{1}(\cdot)$ is an indicator function. A lower value for the CRPS score indicates a better forecast.

The CRPS is computed by implemented the *scoringRules* R-package of Jordan et al. (2022). This package uses closed-form expressions for different distributions of the above CRPS formula. For the models that incorporate a normal distribution, the CRPS for time t is calculated by adjusting the CRPS of the standard normal distribution with the forecasted location and scale parameters. The calculation is shown in Formula 17, following Gneiting et al. (2005).

$$
\begin{aligned}
CRPS_t(\Phi, y_t) =& y_t(2\Phi(y_t) - 1) + 2\varphi(y_t) - \frac{1}{\sqrt{\pi}} \\
CRPS_t(F_{\hat{\mu}_t,\hat{\sigma}_t}, y_t) =& \hat{\sigma}_t CRPS_t\left(\Phi, \frac{y_t - \hat{\mu}_t}{\hat{\sigma}_t}\right)
\end{aligned}
\tag{17}
$$

In which $\Phi(y_t)$ is the standard normal CDF function and $F_{\hat{\mu}_t,\hat{\sigma}_t}$ is the obtained distributional forecast.

For the models that incorporate a Student t distribution, the CRPS for time t is calculated by adjusting the CRPS of the standard Student t distribution with forecasted degrees of freedom ν_t, with the forecasted location and scale parameters (μ_t, σ_t). The calculation is shown below, following Jordan et al. (2019).

$$
\begin{aligned}
CRPS_t(F_{\hat{\nu}_t}, y_t) =& y_t(2F_{\hat{\nu}_t}(y_t) - 1) + 2f_{\hat{\nu}_t}(y_t)\frac{\hat{\nu}_t + y_t^2}{\hat{\nu}_t - 1} - \frac{2\sqrt{\hat{\nu}_t}}{\hat{\nu}_t - 1}\frac{B(\frac{1}{2}, \hat{\nu}_t - \frac{1}{2})}{B(\frac{1}{2}, \frac{\hat{\nu}_t}{2})^2} \\
CRPS_t(F_{\hat{\nu}_t,\hat{\mu}_t,\hat{\sigma}_t}, y_t) =& \hat{\sigma}_t CRPS_t\left(F_{\hat{\nu}_t}, \frac{y_t - \hat{\mu}_t}{\hat{\sigma}_t}\right)
\end{aligned}
\tag{18}
$$

In which $F_\nu(y_t)$ is the standard CDF function of the Student t distribution with degrees of freedom ν and $F_{\hat{\nu}_t,\hat{\mu}_t,\hat{\sigma}_t}$ is the obtained distributional forecast.

After calculating the one-step-ahead CRPS scores for all months in the test data-set, we calculate the average CRPS score for a model. These performance scores will be employed to compare all models.

To see whether our regression tree-based models perform better than our best benchmark model, a test is performed on all separate CRPS scores. Other than looking at the differences in CRPS score, the significance of the differences are examined using a Diebold-Mariano test (Diebold and Mariano, 1995). Originally this test is used with forecast errors of predictive models, but it is also used to compare scoring rules with each other in distributional forecasting. Gneiting and Katzfuss (2014) implements it to compare CRPS scores as well, using the following test statistic:

$$
\begin{aligned}
t^{\text{DM}} =& \sqrt{Q}\frac{CRPS^{\text{model1}} - CRPS^{\text{model2}}}{\hat{\sigma}} \qquad \text{in which} \\
\hat{\sigma}^2 =& \frac{1}{Q}\sum_{t=1}^{Q}\left(CRPS_t^{\text{model1}} - CRPS_t^{\text{model2}}\right)^2
\end{aligned}
\tag{19}
$$

In which Q is the number of observations in the train data-set and $\hat{\sigma}$ an estimator of the standard deviation of the CRPS scoring rule difference. The test statistic is asymptotically standard normal,

following the conclusions of Gneiting and Katzfuss (2014). Our research implements a two-sided test, that concludes which model is significantly better performing based on the sign of the test-statistic if the null-hypothesis is rejected. This test is incorporated using the *dm.test* function of the R-package *forecast* (Hyndman et al., 2023).

To also take into account the interdependency of our different models, we implement a model confidence set of Hansen et al. (2011) of all our models. This method forms a set of all models that contains the best model with a pre-specified significance level, which is set to 10% in our research. It can be seen as a generalization of the Diebold-Mariano test that is computed as follows:

$$t_k^{\text{MCS}} = \frac{d_{CRPS^{\text{model}k}}}{\sqrt{\hat{\text{var}}(d_{CRPS^{\text{model}k}})}} \tag{20}$$

In which $d_{CRPS^{\text{model}k}}$ is the loss of model k compared to the average losses of all other models incorporated. This test is incorporated in R using the *rugarch* package by Ghalanos (2022).

3.7 Economic Interpretation of Distributional Models

After calibrating our benchmark models and our research focus models, we will use the best forecasting models of each group to compare in an economic application. Investors are interested in making portfolio decisions. They make these decisions based on maximising their utility functions, which partly depend on the risk aversion of the specific investor. Very risk-averse investors invest in the risk-free rate, while others want to invest in more risky assets. Next to that, some of these utility functions need full distributional forecasts of those risky assets instead of point-forecasts only to decide on the investment weights.

To assess the economic use of better distributional forecasts, this research follows the set-up of Zhao (2013). For convenience, only two assets are taken into account to allocate weights over for investing: the risk-free rate r_t^f, and the S&P500 index (a risky asset). For the risk-free rate, the 3-month Treasury Bill data is used (*tbl*). To decide the allocation of these assets for the next period ($\omega_t^{r^f}$, $\omega_t^{S\&P500}$), a utility function is maximised. Moreover, we restrict the research so that there is no short selling allowed: $\omega_t^{r_f}$, $\omega_t^{S\&P500} \geq 0$. Investors are assumed to have a power utility function over their wealth in the next period, because the full distributional forecast of the risky asset is used in assigning the optimal weights with this function (Zhao, 2013). This is also called Constant

Relative Risk Aversion (CRRA):

$$U(W_{t+1}) = \begin{cases} \frac{W_{t+1}^{1-\gamma}}{1-\gamma} & \text{for} \quad \gamma \neq 1 \\ \log(W_{t+1}) & \text{for} \quad \gamma = 1 \end{cases} \tag{21}$$

This function depends on the relative risk aversion coefficient γ and the wealth of the next period (W_{t+1}). This wealth is defined as follows:

$$\begin{aligned} W_{t+1} &= W_t(1 + \omega_t^{rf} r_t^f + \omega_t^{S\&P500} R_{t+1}) \\ &= W_t(1 + \omega_t R_{t+1}^e + r_t^f) \end{aligned} \tag{22}$$

In which R_{t+1} and R_{t+1}^e are the return and the excess return on the S&P500 index respectively, ω_t the weight of the risky asset using ω_t^{rf} = 1 - $\omega_t^{S\&P500}$ as there are only 2 investment options.

One can find the optimal weight allocation at time t, ω_t^*, by maximising the investors utility function, giving the following solution:

$$\omega_t^* = \text{argmax}_{\omega_t} \int_{-\infty}^{\infty} \frac{[W_t(1 + \omega_t R_{t+1}^e + r_t^f)]^{1-\gamma}}{1-\gamma} f(R_{t+1}^e | \mathcal{I}_t) dR_{t+1}^e \tag{23}$$

In which $f(R_{t+1}^e|\mathcal{I}_t)$ is the PDF of the excess returns based on all information up until point t. The distributional forecasting models are trained for continuously compounded returns, so they have to be carefully converted following Zhao (2013). They can be calculated following Formula 24.

$$f(R_{t+1}^e|\mathcal{I}_t) = \left| \frac{100}{R_{t+1} + 1} \right| f(y_{t+1}|\hat{\theta}) \tag{24}$$

In which $f(y_{t+1}|\hat{\theta})$ is the forecasted PDF function for the next period from our distributional models.

These optimal one-step-ahead weights are computed for the best tree-based model and the best benchmark model for the full test data-period. With those the utilities of the investor over time are computed (W_t^*) in the test data-period as well, containing Q observations. The certainty equivalent rate (CER) of return is then formed as follows, derived by Zhao (2013) as well:

$$CER^{\text{model}} = \left((1-\gamma) \frac{1}{Q} \sum_{t=1}^{Q} U(W_t^*) \right)^{\frac{1}{1-\gamma}} - 1 \tag{25}$$

This CER value can be seen as a performance measure of the formed investment portfolio, as it is the rate that a risk-free investment should offer to be as valuable to the investor as the current investment. The higher the CER, the better the investment. To compare our distributional forecasting model to the benchmark model in an economic setting we will look at the additional CER our best distributional model delivers.

4 Results

This section display the results obtained in this research. First the results of all the models elaborated on to form distributional forecasts for the test data-set are shown, and conclusions are drawn accordingly. Thereafter, the results of the economic application are presented.

4.1 Results of the Distributional Forecasting Models

In this research of probabilistic forecasting monthly continuously compounded returns, different models are considered: Historical Simulation, several versions of the ARMA(1,1)-GARCH(1,1) model, Distributional Regression Forests, Distributional Adaptive Soft Trees and a Beta-transformed Linear Pool of the latter two. These models are implemented to forecast the continuously compounded returns of the monthly S&P500 index specifically. All models are formed based on data from January 1970 up until December 2014, which is our train data-set. Then the models are compared based on their forecast performance for the test data-set (January 2015 - December 2022). These forecasts are 1-step-ahead forecasts, without re-calibrating the model at different time-points.

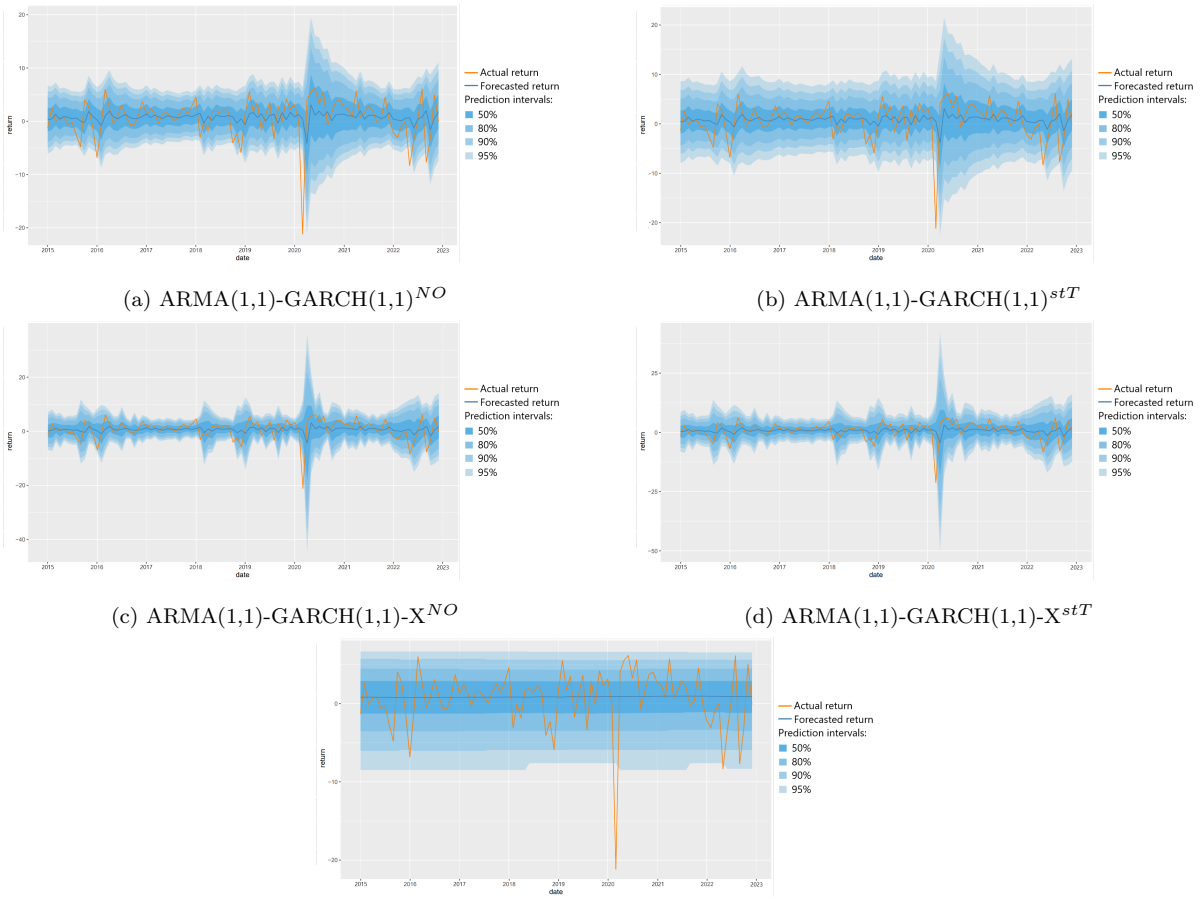
Table 2: Coefficients of the GARCH models

	Mean Model			Variance Model				
	μ	ϕ_1	υ_1	ψ	ζ	χ	ξ	ν
ARMA(1,1)-GARCH(1,1)^{NO}	0.770	-0.090	0.319	0.947	0.145	0.792	N.A.	N.A.
ARMA(1,1)-GARCH(1,1)^{stT}	0.873	-0.104	0.322	0.982	0.089	0.836	N.A	4.775
ARMA(1,1)-GARCH(1,1)-X^{NO}	0.614	-0.110	0.345	1.293	0.000	0.000	6.019	N.A.
ARMA(1,1)-GARCH(1,1)-X^{stT}	0.758	-0.073	0.312	2.178	0.000	0.000	5.376	7.033

Note: The table reports the coefficients of the different GARCH models implemented in Section 3: An ARMA(1,1)-GARCH(1,1) model with and without the additional explanatory variable *ssr*. *NO* stands for the assumption of a normal distribution, *stT* stands for the assumption of a Student t distribution. The AR, MA, ARCH and GARCH coefficients are respectively $\phi_1, \theta_1, \zeta, \chi$. The constant of the mean model is μ, and the one of the variance model is ψ. The shape parameter of the Student t distribution is ν and the coefficient of the explanatory variable is ξ; both only reported when applicable.

To decide on which model is set as our benchmark model, we compare the relatively simple models first among each other. These are the Historical Simulation model, and the discussed versions of the ARMA(1,1)-GARCH(1,1) models. In Table 2 the coefficients of all parametric benchmark models are reported. What can be seen is that when adding *ssr* as an explanatory variable, the ARCH and GARCH coefficients of the variance model become zero. Therefore one can conclude that the

sum of squared daily returns of the previous month has all explanatory power in this regression. Moreover, the AR coefficient is always negative. This indicates that the previous month's value has an opposite effect on the current value.



(a) ARMA(1,1)-GARCH(1,1)^{NO}

(b) ARMA(1,1)-GARCH(1,1)^{stT}

(c) ARMA(1,1)-GARCH(1,1)-X^{NO}

(d) ARMA(1,1)-GARCH(1,1)-X^{stT}

(e) HS

Figure 3: **Probabilistic forecasted returns of possible benchmark models shown with their respective prediction intervals and the actual returns.** This figure shows the 1-step-ahead probabilistic forecasts for the ARMA(1,1)-GARCH(1,1) models with normal and Student t distributions, with and without the explanatory variable *svar* for the test data-period from January 2015 to December 2022. The prediction intervals shown are with certainty of 50%, 80%, 90% and 95%. NO stands for the assumption of a normal distribution, stT stands for the assumption of a Student t distribution. The Historical Simulation method (HS) is displayed as well.

In Figure 3 the density forecasts of all possible benchmark models are shown for the out-of-sample test period, together with the actual realized values of the returns. As expected, the historical simulation does not alter much over time. The other models appear to more or less follow the actual returns with a lag. The models without the explanatory variable in the variance model do not seem

to anticipate on uncertainty quickly enough. As can be observed, both of the models including the explanatory variable *ssr* have more sudden high peaks and decline more rigorously. These prediction intervals respond heavy on previous lows especially, probably because of the magnitude of the lows. When looking at the range of values that the prediction intervals take, the predicted standard deviations seem to over-respond.

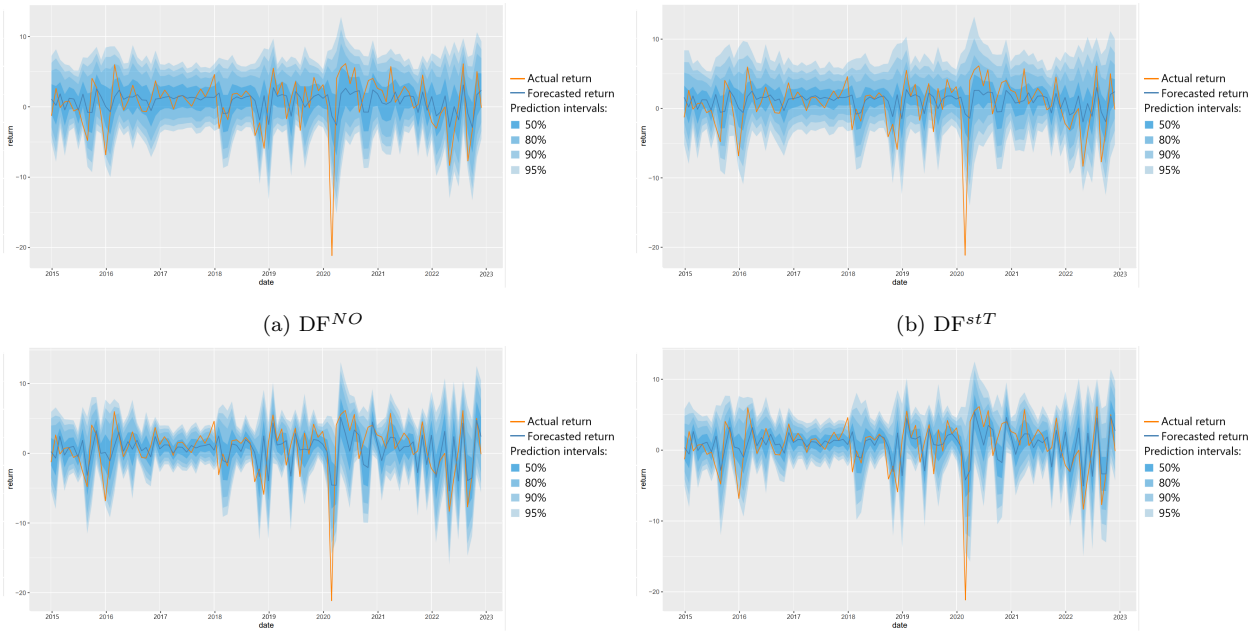
Table 3: Performance statistics based on the CRPS scoring rule

	$\text{CRPS}^{\text{train}}$	$\text{CRPS}^{\text{test}}$
Panel A: Non-parametric benchmark model		
HS	-	1.837
Panel B: Parametric benchmark models		
ARMA(1,1)-GARCH(1,1)^{NO}	-	1.835
ARMA(1,1)-GARCH(1,1)^{stT}	-	1.860
ARMA(1,1)-GARCH(1,1)-X^{NO}	-	1.816
ARMA(1,1)-GARCH(1,1)-X^{stT}	-	1.827
Panel C: Regression tree-based models		
DF^{NO}	1.565	1.653*
DF^{stT}	1.611	1.637**
DAST^{NO}	1.536	1.602
DAST^{stT}	1.529	**1.570*****
Panel D: Ensemble of regression tree-based models		
BLP	1.510	1.578**

Note: The table reports the average out-of-sample Continuous Ranked Probability Score (CRPS) for the Historical Simulation method, several ARMA(1,1)-GARCH(1,1) models, the Distributional Forests, the Distributional Adaptive Soft Trees and the Beta-transformed Linear Pool. *NO* stands for the assumption of a normal distribution, *stT* stands for the assumption of a Student t distribution. *, **, *** signify that the CRPS is different from the best scoring benchmark model with significant levels respectively 5%, 10%, 20% following the Diebold-Mariano test. For the models included in the Beta-transformed Linear Pool, the average CRPS of the train data is displayed as well. The bold printed models are the models that are included in the model confidence set when implementing a significance level of 10%.

The out-of-sample distributional forecasts can however be compared more precisely by using scoring rules. As discussed in Section 3.6, the scoring rule that is used to compare the forecasts is the mean Continuous Ranked Probability Score (CRPS). This mean is calculated over all one-step-ahead forecasts in the test data-set. The lower the score, the better. In Table 3, it can be seen

that the model including the explanatory variable and that assumes the normal distribution obtains the best out-of-sample scores. Therefore, this method is selected as our benchmark method. Pointedly, the CRPS scores of the Historical Simulation are not the worst of all benchmark methods.



(a) DF^{NO} (b) DF^{stT}

(c) DAST^{NO} (d) DAST^{stT}

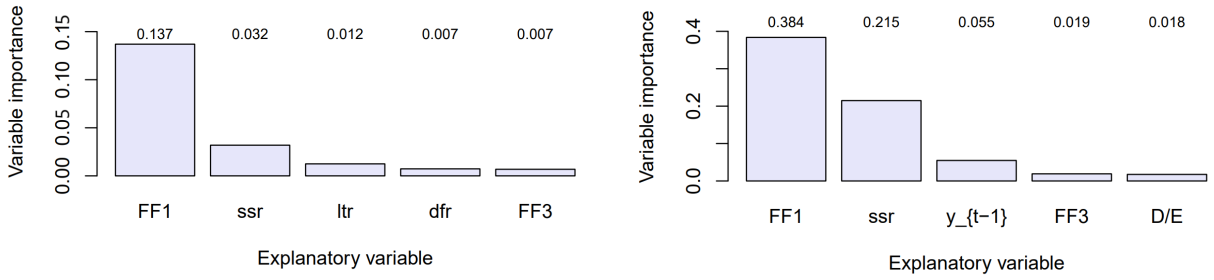
Figure 4: **Probabilistic forecasted returns of the distributional tree-based models shown with their respective prediction intervals and the actual returns.** This figure shows the 1-step-ahead probabilistic forecasts for the Distributional Forests and Distributional Adaptive Soft Trees with normal and Student t distributions for the test data-period from January 2015 to December 2022. The prediction intervals shown are with certainty of 50%, 80%, 90% and 95%. NO stands for the assumption of a normal distribution, stT stands for the assumption of a Student t distribution.

The methods of interest in our research are the tree-based distributional methods. These tree-based models incorporate all explanatory variables summed up in Section 2. Both the Distributional Forests (DF) and the Distributional Adaptive Soft Trees (DAST) are formed with the assumption of normally distributed data and Student t distributed data. Figure 4 shows the out-of-sample distributional forecasts of these methods, and the CRPS scores are stated in Table 3. The DF's appear to be too conservative when looking at the prediction interval graphs, while the DAST's are less cautious. This resonates with the out-of-sample CRPS scores, as the score of the DAST^{stT} is the best of all single models. Overall, all tree-based models have better results in terms of CRPS

than the benchmark models. However, when comparing the CRPS scores over time of the separate regression tree-based models with our best benchmark model (ARMA(1,1)-GARCH(1,1)-X^{NO}), this conclusion appears not to hold. In the table the results of the Diebold-Mariano test are shown for different significance levels. Only the DF^{NO} model is separately significantly different and lower from the benchmark model when implementing a significance level of 5%.

Taking into consideration the interdependency of our models, we have also performed a model confidence set test. In Table 3 the models that are included in the model confidence set are shown in bold. The set was formed with a significance level of 10%, and only contains the $DAST^{stT}$ model. Based on this result, we will use this model as our best regression tree-based model.

While these models are performing better than the benchmark models, we do not yet know which of the underlying variables mainly drive the results. As discussed in Section 3, the variable importance of these trees are assessed in terms of CRPS difference when breaking the associations of the model with specific explanatory variables. This is done by artificially setting the values of this explanatory variable in the test data-set to zero. The five explanatory variables that have the highest variable importance for the best of each forecasting method, DF^{stT} & DAST^{stT}, are shown in Figure 5.



(a) DF^{stT} (b) DAST^{stT}

Figure 5: **Variable Importance of best performing tree-based methods.** This figure shows the 5 explanatory variables with the highest variable importance in terms of additional average CRPS scores. These are shown for the best DF and DAST model. *stT* stands for the assumption of a Student t distribution.

For both of the models, the first Fama&French factor is most important, which is the size premium variable. As expected is the sum of squared daily returns, *ssr*, an important variable as well. What is astonishing is that the lag of the dependent variable is only in the top five for one of the models.

This could however be explained because we look at monthly data, and not higher frequency data. The third Fama&French factor is present in both lists. Next to that, the economic variables "Long Term Rate of Return", "Dividend Payout Ratio" and the "Default Return Spread" are of relatively high importance for the tree-based methods.

The combination of the distributional tree-based models is formed with a Beta-transformed Linear Pool. As far as our knowledge goes, this research is the first to pool multiple tree-based distributional models. The Beta-transformed Linear Pool consists of all four tree-based models.

In Table 4 the weights of these models are displayed together with the optimised shape parameters. Those parameters are obtained through maximum likelihood estimation on the train data-set. As can be seen, the weight of $DASF^{NO}$ is equal to zero and therefore not a part of the BLP. However, the other weights are well distributed, which could indicate that the combination is indeed better performing than a single model. When looking at the CRPS scores over the train data-set (in-sample) in Table 3, one can see that indeed the BLP scores lowest of all models.

Table 4: Optimized parameters of the Beta-transformed Linear Pool

Weights				Shape parameters	
DF^{NO}	DF^{stT}	$DASF^{NO}$	$DASF^{stT}$	α	β
0.519	0.138	0.000	0.343	1.309	1.305

Note: The table reports the optimised parameters of the Beta-transformed Linear Pool, which consists out of the weights of the different models to combine summing up to 1, and the shape parameters α, β. The shape parameters both have to be bigger than zero. NO stands for the assumption of a normal distribution, stT stands for the assumption of a Student t distribution.

In Figure 6, the out-of-sample distributional forecasting results of the BLP can be seen. The CRPS scores are unfortunately not better out-of-sample compared to the $DAST^{stT}$ method. This might be due to the magnitude of the unforeseen negative shock that Covid-19 has brought with it, combined with this model having a tighter distribution. It could also be that the optimisation over the train data-set resulted in over-fitting this ensemble for that period in time. Moreover, it is not included in the formed model confidence set. Consequently the $DAST^{stT}$ method will be implemented in the

next part of our research.

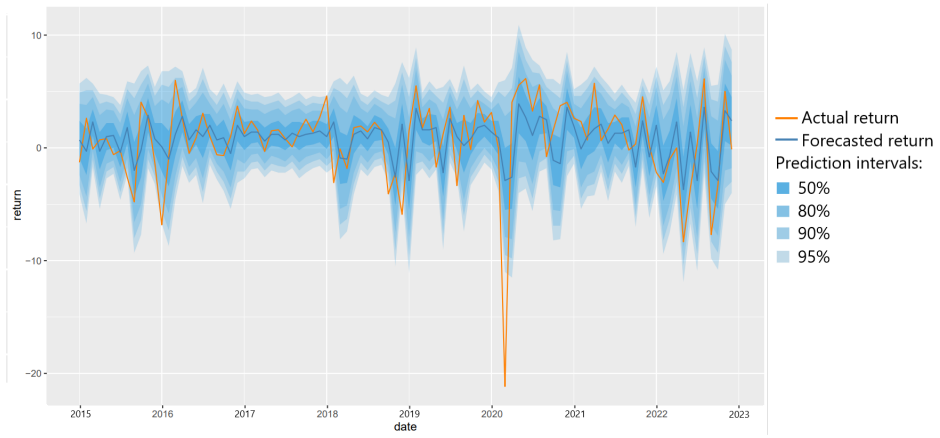


Figure 6: **Beta-transformed Linear Pool.** This figure shows the 1-step-ahead probabilistic forecasts for the Beta-transformed Linear Pool for the test data-period from January 2015 to December 2022. The prediction intervals shown are with certainty of 50%, 80%, 90% and 95%.

To asses whether all of the single parametric models we have used are calibrated right, the out of sample Probability Integral Transform histograms of all models are analysed. These histograms are displayed in Appendix B, and there is concluded that the models are calibrated adequately.

4.2 Results of Portfolio Optimisation

As we now have concluded that the best performing benchmark model is the ARMA(1,1)-GARCH(1,1)-X^{NO} model, and the best tree-based model is the DAST^{stT} model, we are going to compare them in an economic application. Both were best in one-step-ahead out-of-sample forecasting of the distributional model of the continuously compounded returns, and the economic application matches this set-up.

The Certainty Equivalent Returns of both models are displayed in Table 5, together with their difference. The difference shows how much higher a risk free rate should be to replace the formed portfolios with each other. These portfolios were formed with the distributional forecasts of the risky asset: the S&P500 index. The CER's are shown for a number of possible risk aversion coefficients to represent multiple investors. Generally spoken, the more risk-averse an investor becomes, the less beneficial our tree-based model becomes compared to our benchmark model. This can be easily interpreted as a more risk-averse investor is going to invest in the risk-free rate more often already, meaning that the forecast of the risky asset is less important.

However, one exception is striking: when the risk aversion coefficient becomes equal to one, meaning that the utility function becomes a logarithmic utility function, the benchmark model is more desired. This means that the optimal investment weights based on the distributional forecasts models performs better for the benchmark model if an investor has a risk aversion coefficient of one.

Table 5: Certainty Equivalent Rates of return

$\gamma =$	1	2	3	4	5	6	7	8	9	10
DAST^{stT}	0.483	0.573	0.485	0.396	0.329	0.283	0.247	0.225	0.193	0.037
ARMA(1,1)-GARCH(1,1)-X^{NO}	0.594	0.481	0.399	0.335	0.288	0.247	0.220	0.198	0.179	0.191
Difference	-0.111	0.091	0.085	0.062	0.041	0.036	0.027	0.027	0.014	-0.021

Note: The table reports the Certainty Equivalent Returns of the best benchmark model and tree-based model in terms of CRPS; ARMA(1,1)-GARCH(1,1)-X^{NO} and DAST^{stT}. Their differences are reported as well for comparison of additional CER by our optimal model. Multiple risk aversion coefficients (γ) are used for multiple types of investors.

For the majority of the risk aversion coefficients however, the optimal distributional forecasting method results in higher CER's. This shows that a better distributional forecasting method is beneficial in economic applications.

5 Conclusion

This paper argues that our best out-of-sample tree-based model does improve the one-month-ahead investment choices for most risk aversion coefficients compared to the investment choices made based on our best benchmark model. This is found when assuming a power utility function for the investors and only considering the risk-free rate and the S&P500 index as investment options. The best benchmark model used in this investment choice problem is the ARMA(1,1)-GARCH(1,1)-X model with a proxy of the realized volatility as the added explanatory variable. This proxy is the sum of squared daily returns. Moreover, the assumed underlying distribution was the normal distribution for this model. This model outperformed the other possible benchmark models in terms of average out-of-sample Continuous Ranked Probability Score, compared to the Historical Simulation model and the ARMA(1,1)-GARCH models.

Our research shows that all implemented tree-based methods perform better in terms of distributional forecasting continuously compounded returns compared to the benchmarks. These models are all trained on data ranging from January 1970 to December 2014, and tested on data ranging

from January 2015 to December 2022. To answer our research question which tree-based method performs best in this financial application, they are compared among each other. The tree-based models that were used for this research are the Distribution Forests of Schlosser et al. (2019) and the Distributional Adaptive Soft Trees of Umlauf and Klein (2022). Both were formed with an underlying normal distribution and a Student t distribution. In line with their own findings, a single Distributional Adaptive Soft Tree of Umlauf and Klein (2022) performs better in the test-data set than a full grown Distributional Forest. This could be due to the low number of observations in our test data set, i.e. 540 observations. Therefore it would be interesting to look at this finding in bigger data sets. This research could for example be done with higher frequency data instead of monthly continuously compounded returns to look if the findings would still be in line with this research.

For both tree-based methods, the explanatory variables with the highest variable importance are assorted. For both models, the first and third Fama&French factors were among the top variables, as well as the sum of squared daily returns. Other explanatory variables that came up as important were the one month lag of the dependent variable, the long term rate of return, the dividend payout ratio and the default return spread. However, our research only explores 17 economic explanatory variables. We did not explore the option to incorporate any other explanatory variables possibly of a different nature. It could be a huge opportunity, especially for the tree-based models, to incorporate other variables additionally that could influence the continuously compounded returns of assets. As the benefits of the tree based models are the automatic selection of explanatory variables to incorporate and examine the possible interactions between them, I do not think that we have reached the limits of explanatory power yet in these models.

Additionally this research looked at the possibility of combining multiple tree-based distributional methods into one ensemble distributional model. This has been done by implementing a Beta-transform Linear Pool. This method gave us a glimpse at the opportunities of combining these methods as it outperformed all considered models in-sample. However, out-of-sample it did not deliver the best performance score of all models. Therefore, further research can be done to look into another method to make an ensemble of tree-based distributional models which might perform even better. Moreover, the opportunity of a penalizing rule to prevent overfitting on the train data when combining the methods could be explored to optimise the parameters of the ensemble. To the best of the authors knowledge, ensemble methods have not been used for tree-based methods

in any field of research. The opportunities of combining these methods with each other, or maybe even with completely other distributional forecast methods could be explored in asset forecasting, but also in many other research fields.

Furthermore, this research can be build upon to widen the specific research question in various ways. One could for example incorporate more risky assets in the allocation problem to see whether the probabilistic forecasting method forecasts well for all assets, and not only the S&P500 index. It could also be generalized by looking at multiple periods ahead forecasting instead of only one-month-ahead, to make the investment more executable in the real word. Lastly, this research only took into account investors with power utility functions and this could be broadened.

The importance of all of the suggested further researches proofs the relevance of the research-flow to distributional forecasting in financial applications. It justifies why this subject is gaining so much traction over the years and I am excited to see what the developments lead to over the coming years.

References

Abadir, K. M., Luati, A., and Paruolo, P. (2022). Garch density and functional forecasts. *Journal of Econometrics*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Apergis, N. (1998). Stock market volatility and deviations from macroeconomic fundamentals: evidence from garch and garch-x models. *Kredit und Kapital*, 31:400–412.

Biau, G. and Scornet, E. (2015). A random forest guided tour. *TEST*, 25.

Bogner, K., Liechti, K., and Zappa, M. (2017). Technical note: Combining quantile forecasts and predictive distributions of streamflows. *Hydrology and Earth System Sciences*, 21(11):5493–5502.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. (1987). A conditional heteroscedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics*, 69:542–547.

Box, G. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Brooks, L., Farrow, D., Hyun, S., Tibshirani, R., Rosenfeld, R., and Viboud, C. (2018). Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *Plos Computational Biology*, 14(6).

Cai, Y. and Stander, J. (2019). The Threshold GARCH Model: Estimation and Density Forecasting for Financial Returns*. *Journal of Financial Econometrics*, 18(2):395–424.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Ciampi, A., Couturier, A., and Li, S. (2002). Prediction trees with soft nodes for binary outcomes. *Statistics in Medicine*, 21(8):1145–1165.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–55.

Ghalanos, A. (2022). *rugarch: Univariate GARCH models.* R package version 1.4-9.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7(none):1747 – 1782.

Greenwood-Nimmo, M., Nguyen, V. H., and Shin, Y. (2012). Probabilistic forecasting of output growth, inflation and the balance of trade in a gvar framework. *Journal of Applied Econometrics*, 27(4):554–573.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*.

Hendricks, D. (1996). Evaluation of value-at-risk models using historical data. *Economic Policy Review - Federal Reserve Bank of New York.*, 2(1):36–39.

Hoogerheide, L. F. (2012). Density prediction of stock index returns using garch models: Frequentist or bayesian estimation? *Economics Letters*, 116(3):322–325.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.*, 15:651 – 674.

Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *Journal of Machine Learning Research*, 16(1):3905–3909.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.21.

Isroy, O., Yildiz, O., and Alpaydin, E. (2012). Soft decision trees. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1819–1822.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12):1–37.

Jordan, A., Krüger, F., and Lerch, S. (2022). *dscoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts.* R package version 1.0.2, URL https://CRAN.R-project.org/package=scoringRules.

Khatin-Zadeh, O., Hu, J., Banaruee, H., and Marmolejo-Ramos, F. (2023). How emotions are metaphorically embodied: measuring hand and head action strengths of typical emotional states. *Cognition and Emotion*, 37(3):486–498.

Kiran, K. G. and Srinivas, V. V. (2021). Distributional regression forests approach to regional frequency analysis with partial duration series. *Water Resources Research*, 57(10).

Klein, N., Smith, M., and Nott, D. (2023). Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *Journal of Applied Econometrics*, 38(4):493–511.

Luo, J., Chen, L., and Liu, H. (2013). Distribution characteristics of stock market liquidity. *Physica A: Statistical Mechanics and its Applications*, 392(23):6004–6014.

Matheson, J. and Winkler, R. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

Mikis, D. S., Robert, A. R., Nikolaos, G., and Fernanda, D. B. (2022). Principal component regression in gamlss applied to greek–german government bond yield spreads. *Statistical Modelling*, 22(1-2):127–145.

Moritz, N. L., Schlosser, L., and Zeileis, A. (2019). *disttree: Trees and forests for distributional regression*. R package version 0.2-0.

Pearson, E. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30:134–148.

Probst, P. and Boulesteix, A. (2018). To tune or not to tune the number of trees in random forest. 18(181):1–18.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society Ser. B*, 72:71–91.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):1564–1589.

Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., and Senge, R. (2021). Distributional regression for demand forecasting in e-grocery. *European Journal of Operational Research*, 294(3):831–842.

Umlauf, N. (2022). *softtrees: Soft Distributional Regression Trees and Forests.* R package version 1.0.

Umlauf, N. and Klein, N. (2022). Distributional adaptive soft regression trees. *Working paper of the Emmy Noether Research Group at Humboldt Universität zu Berlin, eprint:2210.10389.*

Vasseur, S. and Aznarte, J. (2021). Comparing quantile regression methods for probabilistic forecasting of no2 pollution levels. *Sci Rep 11*, 11592.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21:1455—-1508.

Yao, Y., Huang, Q., and Cai, S. (2023). Daily return distribution forecast incorporating intraday high frequency information in china's stock market. *Economic Research-Ekonomska Istraživanja*, page 36:2.

Yaya, O., Olubusoye, O., and Ojo, O. (2014). Estimates and forecasts of garch model under misspecified probability distributions: A monte carlo simulation approach. *Journal of modern applied statistical methods: JMASM*, 13:479–492.

Zeileis, A. and Kleiber, C. (2023). *"countreg: Count Data Regression"*. R package version 0.2-1.

Zhao, Y. (2013). Forecasting the stock return distribution using macro finance variables. *Job Market Paper.*

A Appendix: AdaSoRT (Umlauf and Klein (2022)

For classical regression trees, Umlauf and Klein (2022) came up with the Adaptive Soft Regression Tree structure on which the Distributional Adaptive Soft Trees described in Section 3.4 builds forward. The different with other soft regression tree structures is that it allows to stop a node from splitting before you reach a terminal node if it does not add much information. The structure, fully designed by Umlauf and Klein (2022), is summarized below.

There are n observations of the dependent variable: $y_i, i = 1, ..., n$. The corresponding feature vector of the i^{th} observation is $x_i = (x_{i1}, ..., x_{iq})$, containing q features. For the tree we have, excluding the ultimate root node of the tree, S nodes: T terminal nodes + M root nodes (nodes that have child nodes). The output-value of a node is denoted as $N_{node}(\cdot)$.

Every root node $m \in M$ is the weighted average of its Left (L) and Right (R) child nodes:

$$N_m(x_i) = N_m^L(x_i)p_m(x_i) + N_m^R(x_i)(1 - p_m(x_i)) \tag{26}$$

In which $p_m(x_i)$ can be seen as the posterior probability of directing y_i to the left node given x_i. These probabilities can be calculated using different mapping functions.

For all nodes, $s \in S$, we define the path from the ultimate root node to node s as $D_{(s)}$ and $\mathcal{D}_{(s)}$ the set of nodes that are in that path. The set of corresponding weights is $\Omega_{(s)}$ and the path probability $P_{(s)}$ is calculated as follows:

$$P_s(x^*, \Omega_s) = \prod_{r \in \mathcal{D}_{(s)}} p_r(x^*)^{d_r}(1 - p_r(x^*))^{1-d_r} \tag{27}$$

In which x^* is a new vector of features, the value of a node is displayed as $N_s(x_i) = \beta_s$ and d_r denotes directions to node s in a binary form (left: 0, right: 1).

A predictor of all individual nodes S and the ultimate root node is then:

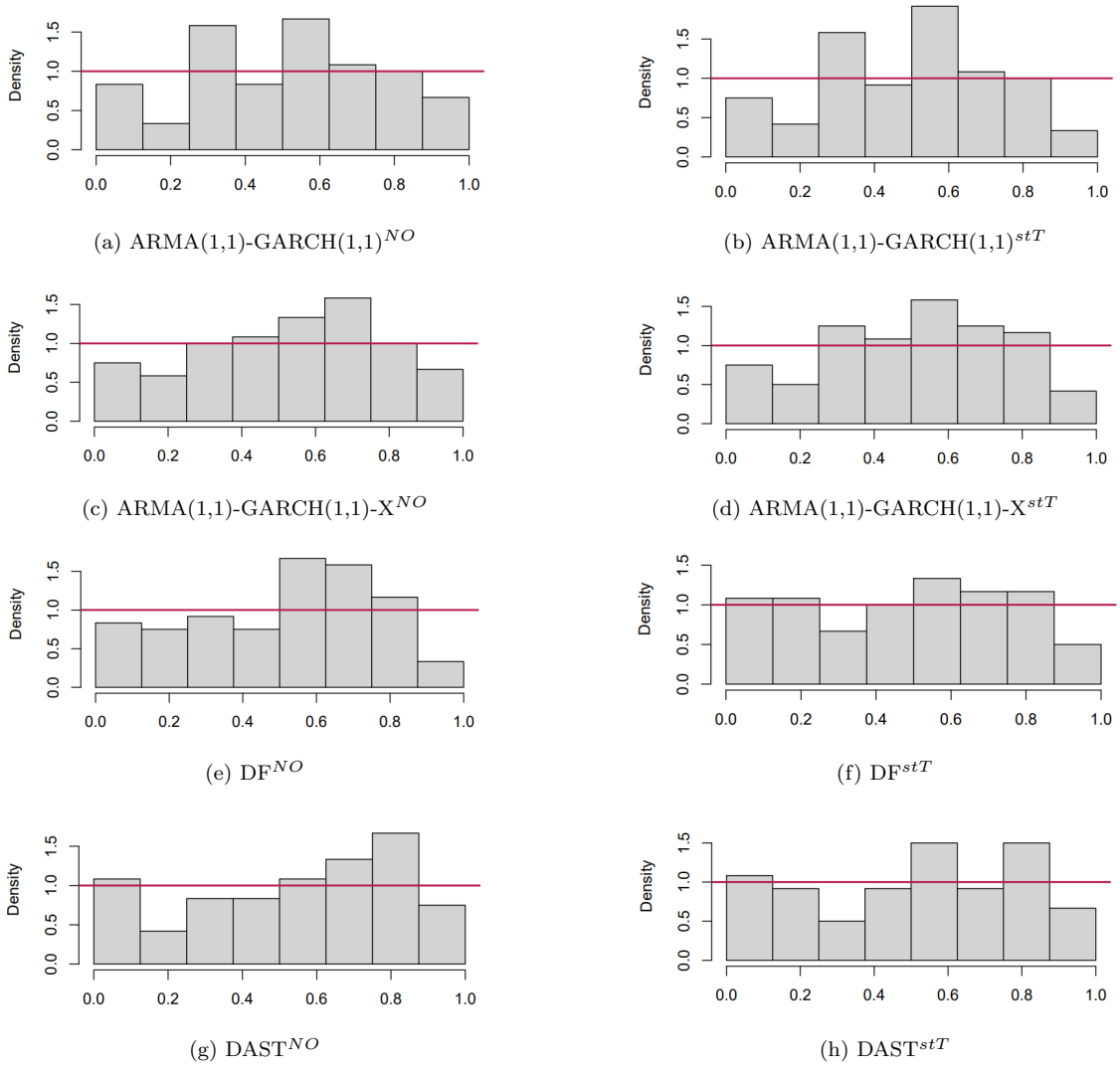
$$\eta(x) = \beta_0 + \sum_{s=1}^{S} P_s(x, \Omega_{(j)})\beta_s \tag{28}$$

Where β_0 is the intercept of the ultimate root node. The AdaSoRT is given in formula 29 where the design matrix ($N(\cdot)$) has as many columns as there are nodes in the tree.

$$\eta = N(X, \Omega)\beta \tag{29}$$

B Appendix: Probability Integral Transform Histograms

To analyse whether the models are well-calibrated, we look at the Probability Integral Transform histograms (PIT-histograms) of all our separate parametric models, like Schlosser et al. (2019). If they are well-calibrated, they should appear almost uniform (Pearson (1938)). As can be seen in Figure 7, the models are overall somewhat uniform. Therefore we conclude that the models are adequately calibrated.



(a) ARMA(1,1)-GARCH(1,1)^{NO}

(b) ARMA(1,1)-GARCH(1,1)^{stT}

(c) ARMA(1,1)-GARCH(1,1)-X^{NO}

(d) ARMA(1,1)-GARCH(1,1)-X^{stT}

(e) DF^{NO}

(f) DF^{stT}

(g) DAST^{NO}

(h) DAST^{stT}

Figure 7: **Probability Integral Transform Histograms** This figure shows the out-of-sample Probability Integral Transform Histograms for all single parametric models employed in this research together with the uniformity line. *NO* stands for the assumption of a normal distribution, *stT* stands for the assumption of a Student t distribution.

These Histograms are formed using the *countreg* R-package (Zeileis and Kleiber (2023)).

C Appendix: Variables

This appendix contains all tables of the appendices.

C.1 Explanatory Variables Abbreviations

This Appendix displays a Table containing all explanatory variables explained in Section 2.

Table 6: List of all explanatory variables used with their abbreviations

Variable Abbreviation	Full Variable Name
b/m	Book to Market Ratio
d/e	Dividend Payout Ratio
d/p	Dividend Price Ratio
d/y	Dividend Yield
dfr	Default Return Spread
dfy	Default Yield Spread
e/p	Earnings Price Ratio
FF1	Fama&French Size Factor
FF2	Fama&French Value Factor
FF3	Fama&French Market Sensitivity Factor
infl	Inflation
ltr	Long Term Rate of Return
lty	Long Term Yield
ntis	Net Equity Expansion
svar	Sum of Squared Returns
tbl	T-bills Rate
tms	Term Spread

Note: The table reports the abbreviations together with the full variable names. The way these variables are computed is explained in Section 2, the Data section. The sources are noted for each individual variable there as well.

C.2 Overview of Variable Denotation

This Appendix displays a Table containing all variable denotations used throughout this research together with their meanings.

Table 7: Variable Denotations

	Variable Meaning	Section		Variable Meaning	Section
α	Shape parameter of Beta distribution	3.5, 4.1	B	Number of segments	3.3
β	Shape parameter of Beta distribution	3.5, 4.1	$\mathcal{B}_b$	Segment space	3.3
β_k	Node values of tree k AdaSoRT	3.4, A	$D_{(s)}$	Path towards node s in tree	A
γ	Risk aversion coefficient	3.7, 4.2	$\mathcal{D}_{(s)}$	Nodes on path $D_{(s)}$	A
$\Gamma()$	Gamma function	3.2	G_k	Number of nodes in AdaSoRT tree k	3.4
ζ	Coefficient of ARCH term	3.2, 4.1	H	Number of trees in Distributional Forest	3.3
η	Explanatory variable predictor	3.4, A	J	Number of observations in train data-set	3.1, 3.5
θ	Set of distribution parameters	3.3, 3.4, 3.7	M	Number of root nodes in AdaSoRT	A
ϵ_t	White noise disturbance term	3.2	$N()$	Design matrix	3.4, A
λ	Shrinkage parameter	3.4	p_t	Price of S&P500 index in month t	2
μ	Mean	3.6, 4.1	Q	Number of observations in test data-set	3.6, 3.7
ν	Degrees of Freedom of the Student t distribution	3.2, 3.6, 4.1	r^f	Risk-free rate	3.7
ξ	Coefficient of explanatory variable: ssr	3.2, 4.1	R_t	Return S&P500 index	3.7
ρ	Penalization term	3.4	R_t^e	Excess return S&P500 index	3.7
σ_t	Volatility	3.2, 3.6	S	Number of nodes in AdaSoRT	A
υ_i	Coefficients of MA terms	3.2, 4.1	T	Number of terminal nodes in AdaSoRT	A
ϕ_i	Coefficients of AR terms	3.2, 4.1	U()	Utility function	3.7
Φ	Standard normal distribution	3.6	w	weights	3.3, 3.4, 3.5, 3.6, 3.7, A
χ	Coefficient of GARCH term	3.2, 4.1	W_t	Wealth at point t	3.7
ψ	Long-run volatility	3.2, 4.1	x	Set of explanatory variables	3.3, 3.4, A
Ω	Matrix of weights associated with AdaSoRT paths	3.4, A	y_t	Continuously compounded return of S&P500 index at month t	2, 3.1, 3.3, 3.4, 3.5, 3.6, 3.7, 4.1
$b_{\alpha,\beta}$	PDF of beta distribution	3.5	z_t	Innovation term in GARCH model	3.2

Note: The table reports all variables used throughout this research with there denotation, and in which section they are used.