Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

Elisabeth Skåland Netland
Håkon Melgård Sveen
Ulrik Leinan Baksjøberget

Probabilistic forecasting of the equity risk premium using quantile machine learning

Master's thesis in Industrial Economics and Technology Management
Supervisor: Einar Belsom
June 2022

NTNU
Norwegian University of Science and Technology

Elisabeth Skåland Netland
Håkon Melgård Sveen
Ulrik Leinan Baksjøberget

Probabilistic forecasting of the equity risk premium using quantile machine learning

Master's thesis in Industrial Economics and Technology Management
Supervisor: Einar Belsom
June 2022

Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

NTNU
Norwegian University of
Science and Technology

Preface

This thesis concludes our Master of Science degree in Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU). The paper depicts the evaluation of tree-based machine learning models in producing probabilistic forecasts of the equity risk premium. We have all immersed ourselves in artificial intelligence and finance during our studies and we are pleased to have been able to combine these exciting disciplines in our master's thesis.

We would like to express our appreciation to our supervisor, Associate Professor at NTNU, Einar Belsom, for great guidance and useful critiques. We very much appreciate all meetings and enthusiastic encouragement. Nevertheless, we would like to stress that we alone are responsible for the content of this paper and any errors.

Elisabeth Skåland Netland
Håkon Melgård Sveen
Ulrik Leinan Baksjøberget

June 2022, Trondheim

Abstract

We evaluate the tree-based machine learning algorithms quantile regression forest (QRF) and quantile gradient boosting (QGB) in the out-of-sample probabilistic forecasting of the equity risk premium (ERP), conditioned on an established set of predictive variables. We predict both the monthly ERP and its long-term level, i.e. its next 1-year and 5-year monthly average. To assess their performances, we compare the models against two selected benchmark models, historical unconditional quantiles and lasso quantile regression. For the 1-month and 1-year point estimates, the models struggle to outperform the benchmark models. On the other hand, QRF and QGB perform well in producing probabilistic forecasts, but they are not significantly better in comparison to the benchmark models. For the 5-year predictions, both QRF and QGB perform significantly better than the benchmark models when predicting both point estimates and all the prediction intervals up to the 60% interval. Thus, we find these models to be valuable for predicting the long-term level of the ERP. The evaluation of feature importance indicates that some of the variables are more important than others. For the 5-year ERP prediction, QRF and QGB show the output gap to be particularly useful.

Sammendrag

Vi predikerer sannsynlighetsfordelingen til risikopremien i aksjemarkedet (ERP) ved hjelp av de tre-baserte maskinlæringsalgoritmene quantile random forest (QRF) og quantile gradient boosting (QGB). For å trene algoritmene brukes et velkjent og etablert datasett fra litteraturen. I tillegg til å predikere den månedlige risikopremien, predikerer vi fremtidig ett- og femårs gjennomsnittlig månedlig risikopremie. For å evaluere modellene sammenlignes de med to referansemodeller, en historisk modell og en regularisert kvantilregresjonsmodell.

Ved predikering av månedlig og ettårs risikopremie oppnår verken QRF eller QGB bedre punktestimater enn referansemodellene. Likevel oppnår modellene gode resultater for predikering av sannsynlighetsfordelingen til risikopremien. Tester for statistisk signifikans viser imidlertid at modellenes prediksjoner ikke er signifikant bedre enn referansemodellene. Ved predikering av femårs gjennomsnittlig månedlig risikopremie predikerer både QRF og QGB betydelig bedre enn referansemodellene, både for punktestimater og for prediksjonsintervallene, særlig opp til 60%-intervallet. Vi fastslår dermed at maskinlæringsmodellene er verdifulle for predikering av punktestimater og sannsynlighetsfordelingen til den langsiktige risikopremien.

I tillegg evalueres enkeltvariablenes evne til å predikere ERP, og resultatene indikerer at noen variabler er viktigere enn andre. Variabelen produksjonsgap (*ogap*) skiller seg ut som spesielt nyttig i predikering av langsiktig femårs gjennomsnittlig risikopremie, for både QRF og QGB.

Contents

1 Introduction

The equity risk premium (ERP) is defined as the expected excess equity return over a risk-free alternative. Looking forward, the expected ERP cannot be directly observed and needs to be estimated. Due to the premium's importance in for example asset pricing and cost of capital calculation, its predictability is of great interest to both theorists and practitioners. There is no consensus on how to estimate the premium, but as stated by Brotherson et al. (2015) and Damodaran (2021), the historical unconditional mean is the most common approach. The topic of ERP prediction is well examined in financial literature, but the results are mixed. While some argue that certain variables or models are valuable in forecasting it, others disagree, finding no variable or model superior to the basic historical unconditional mean. Despite an already extensive literature, new variables and models are constantly being tested to improve the estimation of the expected ERP.

We explore a new approach for predicting the risk premium, utilising tree-based machine learning models for probabilistic forecasting. Focusing on quantile predictions, we investigate and predict the whole conditional distribution, and not only the conditional mean focused on in prior research. We utilise an established set of predictive variables, giving us the opportunity to focus the evaluation on the predictive models and compare the results to previous research. Further, we focus on out-of-sample prediction due to its relevance to real prediction tasks. The main contribution of our work is threefold: (i) we extend the common goal of predicting point estimates of the ERP by exploring probabilistic forecasting using quantile predictions, (ii) generally, we add research on the applicability of quantile random forests and quantile gradient boosting on financial data, and specifically on the topic of ERP prediction, and (iii) we contribute to existing literature with a new assessment of the established variables´ importance when predicting the ERP. In the rest of this section, we position our work in the landscape of ERP predictions.

Traditionally, macroeconomic variables such as dividend-to-price (see, e.g. Fama and French (1998); Ang and Bekaert (2007); Cochrane (2008)), earnings-to-price (see, e.g. Goyal and Welch (2008)), inflation (see, e.g. Campbell and Vuolteenaho (2004)), interest rates (see, e.g. Campbell (1987)), and risk measures (see, e.g. Guo (2006)) have been proposed as better predictors of the expected ERP than the historical mean. Several papers discuss whether these variables have any predictive power at all and, if so, to what extent. To compare and reassess the fragmented research under a common framework, Goyal and Welch (2008) conduct a comprehensive examination of several suggested predictors of the ERP and find that none perform well in predicting the ERP in-sample (IS) nor out-of-sample (OOS). An extensive amount of research has attempted to validate their findings afterwards or improve them by suggesting new variables with better predictive performance. Several claim to find variables that are better, see e.g. Li et al. (2013) who suggest using the implied ERP as a predictor and Neely et al. (2014) who suggest technical indicators. In Goyal et al. (2021), the authors reassess their original suggested predictive variables, as well as a new set of suggested variables from literature published after their original work, finding none of the variables, neither old nor new, to have significant IS or OOS predictive performance.

Due to the inconsistent predictive performance of univariate models, combination forecasts have in general been suggested to increase forecast accuracy, see, e.g. Clemen (1989). Therefore, as an alternative to the univariate examination by Goyal and Welch (2008), Rapach et al. (2010) suggest integrating the forecasts based on single variables, which results in a strong OOS predictive power against the historical mean. Duarte and Rosa (2015) propose combining forecasts from 20 univariate and multivariate models based on historical mean, discounted cash flow models, time series regression, cross-sectional regression, and surveys of professionals in order to predict the expected ERP. However, the paper does not assess the combined model's OOS predictive power. Neely et al. (2014) further investigate a regression model that combines both fundamental and technical predictors through principal components, resulting in forecasts that are superior to the univariate regressions of the predictors separately.

Artificial intelligence, and machine learning in particular, for financial analysis is starting to become a well-established connection, especially for stock return and price movement prediction (of the more recent, see e.g. Ballings et al. (2015), Vijh et al. (2020), Rapach and Zhou (2020), and Basak et al. (2019)). Though, in predicting the ERP, machine learning models are less utilised. Wolff and Neugebauer (2019) investigate tree-based machine learning approaches for equity market predictions using the provided dataset from Goyal and Welch (2008), but do not examine the premium. Their results in predicting equity market returns with machine-learning models are mixed. A comprehensive comparative study of machine learning models to predict single asset risk premium and the aggregated market premium is later performed by Gu et al. (2020) who use linear models (ordinary least squares and elastic net), tree-based models (random forests and boosted regression trees), neural nets, and dimensionality reduction techniques (principal component regression and partial least squares). They find large economic gains when utilising machine learning forecasts, especially by applying trees and neural nets. This research also partly uses the available dataset from Goyal and Welch (2008).

The studies discussed above provide mixed results for ERP prediction. This might be due to the chaotic nature of stock returns, proven to unsettle even state-of-the-art predictive models. In other areas subject to chaotic processes, there has been a growing interest in probabilistic forecasting techniques, i.e. predictive models that can quantify their uncertainty. Dawid (1984) argues that forecasts ought to be probabilistic by nature, and inspired by his work, several recent studies have covered this material (see, e.g. Gneiting and Raftery (2007), Gneiting et al. (2007), Gneiting (2008), and Gneiting and Katzfuss (2014)). Gneiting and Katzfuss (2014) define a probabilistic forecast as a predictive probability distribution over future and unknown quantities of events of interest, with the aim of maximizing the sharpness subject to reliability of the predictive distributions or equivalently minimizing the prediction intervals, based on available information.

As probabilistic forecasts incorporate uncertainty, this method has been proposed in recent decades to be beneficial for financial forecasting of stock prices (see, e.g. Onkal and Muradoglu (1994)) and macroeconomic forecasting of inflation rates (see, e.g. Garrat et al. (2003)). Following this line of thought, a probabilistic forecast of the ERP is more interesting than only a point estimate that only

describes one possible outcome. The additional information given by the probabilistic forecasts provides the basis for better decision making. As an example, the practical study by Alessandrini et al. (2014) shows that when trading future wind energy production, using probabilistic wind power predictions can lead to higher economic gains than using deterministic forecasts alone. Curiously, we find that there is little research on probabilistic ERP forecasting, despite the fact that there are several studies of point estimation ERP forecasts.

Quantile regression, initially suggested by Koenker and Basset (1978), is a technique that enables probabilistic forecasting. Meligkotsidou et al. (2014) connect the empirical evidence of non-normally distributed stock returns, the risk premium´s exhibition of time-varying volatility, excess kurtosis, and negative skewness, to the bad predictive performance of the variables investigated in Goyal and Welch (2008). Rather, they suggest using the quantile regression technique to model the relationship between a set of variables and the quantiles of the ERP. Applying this technique on ERP prediction, the authors find the same set of variables investigated in Goyal and Welch (2008) to be better at predicting certain conditional quantiles of the ERP, than predicting the conditional mean. Meligkotsidou et al. (2019) continue the research on quantile regression applied to ERP prediction by implementing a multivariate complete subset quantile regression framework for predicting the risk premium, resulting in significant OOS predictive power compared to both the historical mean, and the earlier suggested combination of single-variable quantile regression predictions (see Meligkotsidou et al. (2014)).

Combining the quantile regression technique with machine learning algorithms can produce probabilistic forecasts and have yielded promising results in other areas with volatile movements, see, e.g. Verbois et al. (2018) that predict day-ahead solar irradiance, and Nowotarski and Weron (2018) who review the advances in probabilistic forecasting of electricity prices. Several machine learning algorithms can be used for the purpose of probabilistic forecasting, for which some models are more suitable than others. Because the ERP is a continuous real number, the group of supervised machine learning algorithms for regression is a natural starting point for model selection. These techniques model the relationship between the dependent variable and the predictors. In the category of supervised algorithms, there is a vast amount of simple to more complex models to consider, such as models based on extensions to linear regression, neural nets, tree-based models, support vector machines, and k-nearest neighbours (see, e.g. Hastie et al. (2017) and James et al. (2021)).

The different models mentioned above are, to varying degrees, optimized for different use cases. We have decided to apply tree-based machine learning algorithms, i.e. gradient boosting and random forests, chosen based on several factors. Firstly, they are transparent and easily interpretable, pointed out by, e.g. Gu et al. (2020). Of their applied models, they find the most improved stock return prediction using neural networks and trees. They find neural networks slightly better than trees but acknowledge their shortcomings of transparency and interpretability. Secondly, the tree-based models are non-parametric and are thus flexible in that we do not need to make assumptions about the functional form of the relationship between the ERP and the predictors ex-ante. A further advantage of trees is that they easily handle both continuous and categorical variables (see, e.g. James et al. (2021)).

Lastly, the above-mentioned study by Verbois et al. (2018) uses quantile gradient boosting and Vaysse and Lagacherie (2017) use quantile regression forest, both with good results. A disadvantage of many machine learning models is that they are prone to overfitting (see, e.g. Gu et al. (2020)), which is why we use the regularized algorithms random forests and gradient boosting.

With our previously mentioned contributions, we distinguish from previous research. While several previous studies do, to some extent, predict for longer horizons, most concentrate on 1-month or 1-year predictions. We advocate for additionally considering the long-term level due to its relevance for practical use. Therefore, while introducing probabilistic forecasting to ERP prediction, we add to the existing literature by also predicting the long-term level of the ERP. Further, the evaluation of probabilistic forecasts is dependent on appropriate evaluation metrics, which we introduce in the context of ERP prediction. Finally, we seek to elaborate on the application of machine learning models, as to our knowledge, quantile regression forest and quantile gradient boosting are not being applied for ERP prediction in current literature. In that context, we assess the training and validation of these models for predicting the premium with respect to its time series characteristics.

We proceed as follows. The quantile machine learning models need predictive variables to condition the ERP on, and in section 2 we present the retrieval and preprocessing of our data, as well as its characteristics. In section 3 we describe the models used to generate prediction intervals, the empirical procedure for how to do it, and the metrics for evaluating forecasts produced by the models. Section 4 covers the results and discussion. Finally, we conclude and provide suggestions for further work in section 5.

2 Data

In this section, we present the data, i.e. the time series of monthly ERP and the set of selected predictive variables to condition the ERP on. We begin by explaining the process of retrieving and preprocessing the data before presenting relevant descriptive statistics.

2.1 Retrieval and preprocessing

As stated by Gu et al. (2020), there is a vast amount of available variables to condition the ERP on. Our data is primarily based on a well-known and commonly used dataset for the prediction of ERP, provided by Goyal and Welch (2008). This dataset is hereafter denoted GW1 and is generally comprised of macroeconomic variables. Additional variables investigated in Goyal et al. (2021), which provides an updated examination of the original variables as well as additional variables suggested in various literature after their initial article, are included in a second dataset. This dataset is denoted GW2 and includes more macroeconomic variables, as well as sentiment, cross-sectional stock information, and technical indicators. The data needed to calculate the variables from GW2 are gathered from several sources, but mainly from the Federal Reserve Bank of St. Louis, hereafter denoted MV. For details about the data sources for this dataset, see the appendix made available by Goyal et al. (2021). Finally, we further extend the dataset with seven macroeconomic variables retrieved from the macroeconomic database of the Federal Reserve Bank of St. Louis. This selection is chosen based on the determinants of the ERP presented in Damodaran (2021), where he, among others, covers economic risk, inflation and interest rates, and monetary policy. All variables included in the dataset are seen in Table 1.

The variables in GW1 and GW2 are calculated according to the procedure described in Goyal and Welch (2008) and Goyal et al. (2021) respectively. The variables in MV are calculated according to the appendix of St. Fred. Descriptions of the variable transformations can be found in Table 1. The ERP is calculated as the total return of the stock market, proxied by the S&P 500, minus the monthly short-term interest rate *tbl*, as presented by Goyal and Welch (2008). In addition to predicting the 1-month ERP, we also predict the 1-year and 5-year average monthly ERP. To calculate these to be used as target variables, we use the 1 and 5-year return of the S&P500, minus the return if invested in the t-bill over the same period. We calculate the average monthly ERP (note that this is not annualised, but presented as a monthly rate) over the period to find its long-term monthly level.

The variables in GW1 and GW2 mainly consist of monthly frequencies. However, a few are quarterly or annual. In addition, the variables start at different times. As we want to maintain accuracy by avoiding to deal with missing data, we remove incomplete data. Thus, we select only the monthly variables with start time in January 1960 or earlier. The variable *tchi* from GW2 is originally a principal component of several technical indicators. For the purpose of analysing the importance of its constituents, *tchi* is split into three parts: moving average, momentum, and volume. The final dataset consists of 23 variables from GW1 and GW2.

Variable		Description
Panel A: GW1 - Variables from Goyal & Welch (2008)		
dp	Dividend-price ratio	Difference between log of div and log of prices.
dy	Dividend yield	Difference between log of div and log of lagged prices.
ep	Earnings-price ratio	Difference between log of earnings and log of prices.
de	Div-payout ratio	Difference between log of div and log of earnings.
bm	Book-to-market ratio	B/M-ratio for the Dow Jones Industrial Average.
svar	Stock variance	Sum of squared daily returns on S&P500.
ntis	Net equity expansion	12-month moving sum of net issues by NYSE listed stocks divided by total market cap.
tbl	T-bill	3-month U.S. Treasury Bill rate.
lty	Long-term yield	Long-term U.S. Government Bond yields.
ltr	Long-term rate of return	Long-term U.S. Government Bond returns.
tms	Term spread	Difference between lty and tbl.
dfy	Default yield spread	Difference between BAA- and AAA-rated corporate bond yields.
dfr	Default return spread	Difference between corporate bond returns and ltr.
corpr	High quality corporate bond rate	Representing the high quality corporate bond market, i.e. bonds rated AAA, AA, or A
infl	Inflation	U.S. Consumer Price Index.
Panel B: GW2 - Variables from Goyal and Welch (2021)		
ogap	Output gap	Regressing log of industrial production y_t on time t with error v_t being the gap, $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + v_t$.
dtoy	Nearness to Dow 52 week high	Difference between current value and the 52-week high of the Dow Jones Index.
dtoat	Nearness to Dow all-time high	Difference between the current value and the all-time high of the Dow Jones Index.
ygap	Stock-bond yield gap	Difference between aggregated dividends and the 10-year U.S. T-bond yield ("Fed-model").
rdsp	Stock return dispersion	Cross-sectional standard deviation on the set of 100 size and book-to-market portfolios.
gip	Growth in industrial production	Year-end economic growth based on industrial production.
tchi_ma	Moving average (M.a.)	M.a. rule with buy signal when current price of S&P500 is greater or equal than m.a. of past nine months.
tchi_vol	Volume (Vol.)	Vol-rule with buy signal when current on-balance vol. (OBV) is greater or equal than m.a. of OBV past nine months.
tchi_mom	Momentum (Mom.)	Mom-rule with buy signal when current price of S&P500 is greater or equal than the price nine month past.
Panel C: MV - Macroeconomic variables St. Louis Fed Economic Database		
clf16ov	Labor force	First log differences with lagged values.
unrate	Unemployment rate	First differences with lagged values.
rpi	Real personal income	First log differences with lagged values.
uempmean	Avg. duration of unemployment	First differences with lagged values. The variable is aggregated per week.
houst	Housing starts	Log of total new privately-owned housing units started.
bogmbase	Adjusted monetary base	Second log differences with lagged values.
gdp	Gross domestic product	Normalised around zero.

Table 1: Included variables.

2.2 Descriptive statistics

The final dataset includes the ERP and its lagged values, as well as 31 predictive variables spanning from 1960:01-2020:01. Descriptive statistics of the ERP are presented in Table 2 and the variables in Table 5 in Appendix A. Some important features of the ERP should be noted. First, the monthly ERP has a standard deviation of 4.31% and is even more volatile than the monthly returns on S&P 500. The mean of the ERP is 0.43% while the median is 0.87%, indicating more extreme low values than high ones, which the slight negative skew confirms. In addition, the distribution is leptokurtic, indicating that the ERP has more extreme values than the normal distribution. *tbl_m* denote the monthly rate of *tbl*, which we use to calculate the monthly ERP.

	ERP	**IndexReturn**	**tbl_m**
Mean	0,43 %	0,81 %	0,37 %
Standard error	0,16 %	0,16 %	0,01 %
Median	0,87 %	1,20 %	0,38 %
Std	4,31 %	4,30 %	0,27 %
Kurt	2,40	2,46	0,88
Skew	-0,67	-0,66	0,75
Min	-24,77 %	-24,25 %	0,00 %
Max	14,89 %	15,51 %	1,37 %
25th	-1,90 %	-1,56 %	0,18 %
75th	3,25 %	3,61 %	0,51 %

Table 2: Descriptive statistics of monthly realized ERP.

Several economic events in U.S. history have influenced the ERP. In previous research, such as Goyal and Welch (2008) and Goyal et al. (2021), many predictors are dismissed due to their exceptional performance during these events, e.g. the oil crisis in 1973-74 (see French (1997)). Thus, we find it important to highlight two economic events of particular interest for our study: (i) the I.T. bubble, where euphoria over the emerging I.T. market led the S&P500 to a total growth of 320% from the end of 1994 to the end of 1999, with the subsequent collapse (see, e.g. Goodnight and Green (2010)). (ii) The 2008 financial crisis resulted in the second-worst market crash in U.S. history, thus resulting in the worst decline of the ERP in our dataset. Because of the need for consistent data on our variables, we end our dataset in January 2020, thus not covering the ERP through the COVID-19 pandemic starting March 2020.

3 Models and empirical procedure

In this section, we present the predictive models and the empirical procedure. As claimed in Nowotarski and Weron (2018), the most common way to construct probabilistic forecasts is through constructing prediction intervals. We do so by combining the quantile regression technique with machine learning. In subsection 3.1 we introduce the tree-based machine learning algorithms quantile regression forest (QRF) and quantile gradient boosting (QGB). These models have hyperparameters that need to be selected, and in subsection 3.2 we describe the learning procedure for tuning these parameters, as well as the empirical procedure of generating the OOS quantile predictions. For comparison, we benchmark the machine learning models against the historical unconditional model (historical model) and the lasso quantile regression model (QR), described in subsection 3.3. Finally, in subsection 3.4 we describe the evaluation metrics and statistical tests to assess the probabilistic forecasts.

3.1 Tree-based machine learning models

Tree-based machine learning algorithms are based on decision trees, introduced by Quinlan (1986), and further described in Hastie et al. (2017), and James et al. (2021). Decision trees can be used for both classification and regression, depending on the use case. As we aim to predict ERP as a number, our use case dictates the need for regression. The simplest regression tree algorithm creates one single tree based on all available data and recursively partitions the predictor space into several distinct and non-overlapping rectangular regions. The final set of partitions is often described by and visualized in a tree structure by a set of nodes: root node, internal nodes, and leaf nodes. The root node and the internal nodes split the data on a certain predictive variable based on a loss function. To yield a prediction of some unknown response variable, we start at the root node and follow a path through internal nodes until we reach a leaf node. The leaf node then returns the mean of its contained response values. The approach of growing one single tree is known for its high variance and infamous overfitting. We implement two ensemble learning algorithms to reduce the variance, selected due to their ability to predict the conditional quantiles and not only the conditional mean of the response.

3.1.1 Quantile regression forest

Let X denote the predictor matrix and Y the response vector. The general random forest (RF), introduced by Breiman (2001), uses n independent observations $(Y_i, X_i),\ i = 1, \ldots, n$ to grow a large ensemble of trees, where various subsets of the data are used to build each tree. The variable subset considered for splitting a node is determined at random. These features make the algorithm less prone to overfitting, thus making RF suitable for high-dimensional regression (see, e.g. Hastie et al. (2017)). As formulated in Meinshausen (2006) and Vaysse and Lagacherie (2017), given new data $X = x$, each tree returns a prediction in the form of an estimate $\hat{\mu}(x)$ of the conditional mean $\mathbb{E}(Y|X = x)$. The weight $w_i(x, \theta)$ described in Equation 2 is positive if the observation (Y_i, X_i) is part of the same leaf

$l(x, \theta)$ of the tree built from the random vector of variables θ in which x falls within, and zero otherwise.

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta) Y_i \tag{1}$$

$$w_i(x, \theta) = \frac{\mathbb{1}(X_i \in R_{l(x,\theta)})}{\#[j : X_j \in R_{l(x,\theta)})]} \tag{2}$$

Further, $\mathbb{1}(.)$ is the indicator function and $R_{l(x,\theta)}$ is the rectangular subspace defined by the leaf $l(x, \theta)$ of the tree built from θ. The final result is an averaged prediction of K single tree outputs (see Equation 3) and is the approximation of the conditional mean of the response variable. $w_i(x)$ is formulated as in Equation 4.

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{3}$$

$$w_i(x) = K^{-1} \sum_{k=1}^{K} w_i(x, \theta_k) \tag{4}$$

We implement QRF introduced by Meinshausen (2006). QRF is quite similar to RF, but differs in that the weighted observations are used to provide approximations of the whole conditional distribution of the response, and not only the conditional mean. QRF estimates the conditional distribution function $F(y|X = x) = P(Y \leq y|X = x)$ by Equation 5.

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) \mathbb{1}(Y_i \leq y) \tag{5}$$

QRF returns one model per time step which can be used to predict all quantiles. This makes the QRF consistent in its prediction of the conditional quantiles, i.e. predictions of the ERP at lower quantiles are smaller than predictions of greater quantiles, see Meinshausen (2006). We implement QRF using the Python package *RandomForestQuantileRegressor* from Skgarden. The algorithm in this package is based on QRF as described by Meinshausen (2006).

As we will evaluate the importance of each variable when predicting the ERP, we further describe the calculations performed at each partition to find the optimal variable to split on. Following the mathematical description from the documentation of the Python package Sklearn, at node m, the data available from the random subset is represented by Q_m with n_m samples. The node splits the data into two subsets, as described in Equation 6 and Equation 7, based on a candidate split $\Psi(j, t_m)$ for predictor j and threshold t_m.

$$Q_m^{left}(\Psi) = \{(Y_i, X_i) \in Q_m | x_{ij} \in Q_m \leq t_m\} \tag{6}$$

$$Q_m^{right}(\Psi) = Q_m \setminus Q_m^{left} \tag{7}$$

The quality of a split is measured by G, which is the weighted sum of the loss for partitioning the observations based on the candidate split. For regression tasks the loss function H, when using mean squared error, is described by Equation 9.

$$G(Q_m, \Psi) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\Psi)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\Psi)) \tag{8}$$

$$H(Q_m) = \frac{1}{n_m} \sum_{Y_i \in Q_m}^{n} (Y_i - \bar{Y}_m)^2 \tag{9}$$

The parameters best suited to minimise G are found by Equation 10. The above-mentioned procedure is performed recursively on both subsets until the algorithm reaches the limit set by the hyperparameters (described in subsection 3.2), depending on which threshold it hits first.

$$\Psi^* = argmin_\Psi G(Q_m, \Psi) \tag{10}$$

3.1.2 Quantile gradient boosting

While QRF grows multiple trees simultaneously, the gradient boosting algorithm grows $k = 1, \ldots, K$ trees sequentially. The algorithm initialises by fitting the first tree f_0 to minimise a loss function. Each successive tree f_k is then fitted to the negative gradient of the loss function evaluated based on the estimate of the subsequent tree f_{k-1}. The final output is then the estimate $\hat{f}(x) = f_K(x)$. The gradient is defined in Equation 11, where the loss function of using tree f to predict Y is defined in Equation 12. To get QGB, we fit the gradient boosting tree on quantiles, using the quantile or pinball loss function described in Equation 13, where τ denotes a quantile, e.g. the 0.5 quantile.

$$\frac{\partial L(Y_i, f(X_i))}{\partial f(X_i)} \tag{11}$$

$$L(f) = \sum_{i=1}^{n} L(Y_i, f(X_i)) \tag{12}$$

$$L_\tau(Y_i, f(X_i)) = \begin{cases} \tau(Y_i - f(X_i)) & Y_i - f(X_i) \geq 0 \\ (\tau - 1)(Y_i - f(X_i)) & otherwise \end{cases} \tag{13}$$

The mathematical equations are formulated as in Hastie et al. (2017). Otherwise, in regard to choosing the feature to split on at each node, gradient boosting works in a similar fashion to QRF. Differing from QRF, QGB might be inconsistent in its quantile predictions because it creates one model per quantile.

To use QGB, we apply the Python package LightGBM. LightGBM adds two techniques on top of the gradient boosting machine: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). Roughly summarized, GOSS randomly drops small gradients that are under a given threshold, and EFB reduces the feature space by bundling together sparse features that are almost mutually exclusive. Both methods reduce computational time remarkably while preserving almost the same accuracy. For a mathematical elaboration, we refer to Ke et al. (2017).

3.2 Learning procedure

From the total period spanning from 1960:01 to 2020:01, we start training the models using an initial hold out period from 1960:01 to 1989:12 (1960:01 to 1989:01 for 1-year-ahead predictions and 1960:01 to 1985:01 for 5-year-ahead predictions). For all the different time horizons, the first time step we predict is 1990:01. From that on, we train the model at each time step with new data being available consecutively. We predict 19 quantiles in the following range $\{0.05, 0.1, \ldots, 0.9, 0.95\}$, and at each time step we train and validate the model by tuning the hyperparameters of the model to minimise the error on the validation set, using a cross-validation procedure. Because we are working with time series data, using regular cross-validation, i.e. splitting the data randomly in training and validation data, would result in unintended "peeking" into the future when tuning the hyperparameters. For this reason, we follow the walk-forward cross-validation procedure to train the model for OOS prediction. Figure 1 shows the walk-forward cross-validation procedure at time t. In walk-forward validation, we split the available data at time t into five parts. The first part is the first 1/5 of the data, the second part is the first 2/5 of the data, and so on. Each part is then divided into a training set, visualised by the light blue rectangles, and a validation set, visualised by the dark blue rectangles. The validation set has the same size for all five validation runs. At each run, the most recent data is used for validation and the data prior to that for training.

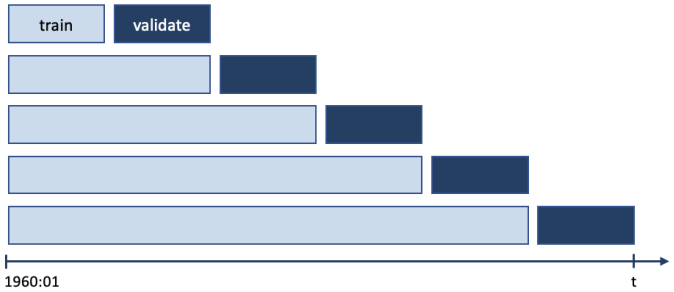


Figure 1: Visualisation of the walk-forward cross-validation procedure for selecting the hyperparameter values.

As stated by Gu et al. (2020) the tuning of parameters has little guidance in theory. Therefore, we optimize the hyperparameters based on a set of selected values, see Table 3. Learning the optimal hyperparameters is performed through a grid search on the different hyperparameter combinations, following the walk-forward cross-validation procedure. That effectively means that adding new hyperparameters has an exponential computational cost. Additionally, by simulating an environment where the models are reset for each time period, this procedure is repeated for every single time step, further increasing the computational cost. For this reason, most parameters are set to their default standards in the packages by Skgarden and LightGBM. More details on the hyperparameters are as follows:

1. Number of estimators: This refers to the number of trees used in the fitting procedure. For both QRF and QGB, this is the number of trees grown. However, the two models behave differently when adjusting this hyperparameter. The original paper by Breiman (2001) on the random forest algorithm states that increasing the number of estimators does not lead to overfitting, although this has been contested later by Segal (2003). At some point, the improvement of adding more trees is minimal, and it is subject to a trade-off between increased accuracy and computational time. In our case, the dataset is sufficiently small to have a stable performance when the number of trees is set at 100. On the other hand, QGB has a trade-off between the number of trees and the learning rate. Generally, a higher number of stages leads to a lower optimal learning rate Friedman (2001). This range is set between 100 and 500 with intervals in hundreds.

2. Learning rate: The learning rate lowers the contribution of each tree and is a property only seen in gradient boosting. During empirical testing, we find that lower learning rates lead to less overfitting and better generalization errors, which aligns with findings from previous studies (see Friedman (2001)).

3. Minimum number of samples per split and per leaf: The minimum sample size required to create an internal split in a node or the minimum number of samples required to create a leaf node. By nature, the two dictate each other, as the creation of a leaf node is dependent on whether a node has been split. If the minimum sample split is lower than the minimum sample leaf, the latter will override the former, and the split will not be created. Empirically, the difference in the validation error is insignificant when tuning both simultaneously. For this reason, only the leaves are tuned to avoid high computational costs.

4. Max depth: This is the maximum depth of the tree. Greater depths increase the bias and reduce the variance. In our training, high max depths seem to increase the bias further for broader quantiles than for the median, making the intervals narrow. When the minimum number of samples in leaf nodes is set to a relatively high number, this artificially lowers the maximum possible depth, as maximum depth is controlled by the number of available samples.

All the tuned hyperparameters with their corresponding values can be found in Table 3. As an end note, to reproduce the same results of our training, the parameter *random_state* should be set to 0.

	QRF	QGB
Nr. of estimators	$\in [100]$	$\in [100, 200, 300, 400, 500]$
Learning rate	-	$\in [0.001, 0.005, 0.01, 0.05, 0.1]$
Min samples per leaf	$\in [2, 5, 10, 20, 30, 50]$	$\in [2, 5, 10, 20, 30, 50]$
Max depth	$\in [2, 5, 10, 15, 20]$	$\in [2, 5, 10, 15, 20]$

Table 3: Sets of selected hyperparameter values.

3.3 Benchmark models

We implement two benchmark models for comparison and verification of the results: historical unconditional quantiles and lasso quantile regression. Thus, at each time step, we calculate the median and the 18 other quantiles corresponding to those predicted by the tree-based machine learning models described above. The historical model is based on the empirical quantiles calculated using all past ERP data at each time step.

We further implement a multivariate quantile regression model with L1-penalization, also called lasso quantile regression (see e.g. Koenker and Basset (1978) and Li and Zhu (2008)). The advantage of adding the L1-penalty is that it automatically performs variable selection by shrinking the fitted coefficient toward zero. The regularization term adds a hyperparameter λ. For each quantile τ, we fit the regression model formulated in Equation 14 which utilises the loss function formulated in Equation 13. Here, $f(X_i)$ is $\beta^\tau X_i^T$. We use the QR implementation provided by the Python package Sklearn, and follow the same walk-forward cross-validation approach as described in subsection 3.2 to find the optimal λ. QR can, as QGB, possibly be inconsistent in its predictions because it creates one model per quantile. We impose a rule such that the benchmark models achieve consistent quantiles to ensure that they are built on the same set of foundations as QRF.

$$\min_{\beta} \sum_{i=1}^{n} L_\tau(Y_i, \beta^\tau X_i^T) + \lambda \|\beta\|_{L_1} \tag{14}$$

3.4 Evaluation metrics

When evaluating the probabilistic forecast, we study the common properties for deterministic forecasts, namely measurement error and bias, and the additional properties specific to probabilistic forecasts, namely reliability and sharpness. Compared to deterministic forecasts, probabilistic forecasts have a larger number of properties, thus comparing different models is more complicated. Scoring rules that account for prediction accuracy, reliability and sharpness are introduced to help rank the different probabilistic forecasts. This section introduces the different metrics used to assess the properties of the forecasts, relevant scoring rules and methods to measure statistical significance.

3.4.1 Evaluation metrics for point estimates

Denote $\hat{Y}_{i,\tau}$ as the quantile forecast. The median, $\tau = 0.5$, is used as our point forecast. As a measure of the prediction accuracy for each model, the mean absolute error (MAE) is calculated at each time step, see Equation 15. We use MAE instead of MSE because the latter punishes outliers more harshly than MAE, making MAE slightly more desirable in our setting, as we do not want extraordinary events like economic recession or stock market bubbles to carry too much weight in the aggregated errors. In addition, as ERP is an economic variable, MAE is a more precise representation of the deviation seen from a practical view.

$$MAE = n^{-1} \sum_{i=1}^{n} |Y_i - \hat{Y}_{i,0.5}| \tag{15}$$

In addition to MAE, we evaluate the point estimates by prediction bias and correlation. Prediction bias is the average difference between the forecasts and the true values and can be measured by the mean bias error (MBE). The MBE can be positive or negative, revealing whether the model suffers from systematic overprediction or underprediction. We further evaluate the correlation between the point forecasts and the realised ERP using the Pearson correlation coefficient.

3.4.2 Evaluation metrics for probabilistic forecasts

The probabilistic forecasts are represented by prediction intervals (PI) based on the corresponding predicted quantile values. Equation 16 is the central $(1-\alpha) \times 100\%$ PI, where l_i is the predicted value, $\hat{Y}_{i,\tau}$, for the lower quantile $\tau = \frac{\alpha}{2}$, and u_i the predicted value for the the upper quantile $\tau = 1 - \frac{\alpha}{2}$.

$$PI_{i,\alpha} = [l_i, u_i] \tag{16}$$

When evaluating probabilistic forecasts, we cannot compare the predicted distribution to the true distribution because the true distribution is non-observable. We need suitable metrics to evaluate the probabilistic forecasts, and literature on the evaluation of probabilistic forecasts suggests reliability and sharpness (see, e.g. Gneiting and Raftery (2007)). Reliability refers to the statistical consistency of the PIs. A PI is considered reliable if its empirical coverage matches the nominal coverage, e.g. the 90% PI should cover 90% of the observations. Sharpness is a measure of the concentration of the distribution, and it does not include the actual observations. Thus, an arbitrarily sharp forecast can easily be created, and comparing the sharpness of two models is only of relevance if their reliability is equal. Reliability and sharpness are, therefore, closely related metrics, well described in previous literature where the probabilistic forecast is desired to maximize sharpness subject to reliability (see, e.g. Gneiting and Raftery (2007) and Nowotarski and Weron (2018)).

As a measure of the reliability of a forecast, the prediction interval coverage probability (PICP) with nominal coverage rate $(1-\alpha)$ can be used. The $PICP^{\alpha}$ is calculated as the coverage rate of the central prediction interval, $PI_{i,\alpha}$, and is described in Equation 17 where a "hit", I_i is described in Equation 18. If $PICP^{\alpha} \sim (1-\alpha)$ for any α, then the model produces a reliable forecast. As a measure of the sharpness of the probabilistic forecasts, we use the prediction interval width $PIW^{\alpha} = u_i - l_i$ and prediction interval average width (PIAW) for different nominal coverage rates.

$$PICP^{\alpha} = n^{-1} \sum_{i=1}^{n} I_i \tag{17}$$

$$I_i = \begin{cases} 1 \ if \ Y_i \in [l_i, u_i] \rightarrow \text{'hit'} \\ 0 \ if \ Y_i \notin [l_i, u_i] \rightarrow \text{'miss'} \end{cases} \tag{18}$$

To test whether the PIs constructed by a model are significantly reliable, we use both the Kupiec test and the Christoffersen test. The Kupiec test developed by Kupiec (1995) is a likelihood ratio test (LR) to evaluate whether a model provides the correct unconditional coverage. The test rejects the null hypothesis of an accurate $(1-\alpha) \times 100\%$ PI if the fraction of "misses" is statistically different from α. The statistic is defined in Equation 19, where $c = (1-\alpha)$, $\pi = n_1/(n_0+n_1)$, and n_0 and n_1 are the number of "misses" and "hits" respectively. The statistic is distributed asymptotically as $\chi^2(1)$.

$$LR_{UC} = -2ln\left\{ \frac{(1-c)^{n_0} c^{n_1}}{(1-\pi)^{n_0} \pi^{n_1}} \right\} \tag{19}$$

The Kupiec test has been claimed to be unsuitable for time series in that it only indicates whether a forecast is significantly reliable on average, not considering the clusters of outliers or violations of the PI. Therefore, Christoffersen (1998) has developed a conditional coverage test, which is a joint test for unconditional coverage and independence. The independence test is presented in Equation 20 and is distributed as $\chi^2(1)$.

$$LR_{Ind} = -2ln\left\{ \frac{(1-\pi_2)^{n_{00}+n_{10}} \pi_2^{n_{01}+n_{11}})}{(1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}}} \right\} \tag{20}$$

where $\pi_2 = (n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$, n_{ij} is the number of observations with value i followed by j and $\pi_{ij} = \mathbb{P}(I_t = j | I_{t-1} = i)$. When evaluating the first order condition, LR_{CC} is the sum of the unconditional coverage test and and an independence test $LR_{CC} = LR_{UC} + LR_{Ind}$, and is distributed asymptotically as $\chi^2(2)$. Both the Kupiec test and the Christoffersen test the null hypothesis of an accurate coverage rate. Thus, for both tests a lower test statistic is better, i.e. we don't want to be able to reject the null. Both tests are further model free, in that they do not test the models, only the forecasts produced by the models.

3.4.3 Scoring rules

When comparing two probabilistic forecasts, a compromise has to be made between reliability and sharpness. Scoring rules are proposed to find an optimal trade-off between the properties and help rank the different models. It is important that the scoring rule is proper, a characteristic described in, e.g. Gneiting and Raftery (2007), which simply means that the average score of the forecast has to be less or equal to the average score of the true distribution, see, e.g. Nowotarski and Weron (2018). Wrinkler (1972) introduced a proper scoring rule for evaluating PIs that assesses reliability and sharpness jointly, as presented in Equation 21. Narrow prediction intervals are rewarded, while "misses" incur a penalty whose size depends on α.

$$S_{\alpha,i}(l_i, u_i; Y_i) = (u_i - l_i) + \frac{2}{\alpha}(l_i - Y_i)\mathbb{1}[Y_i < l_i] + \frac{2}{\alpha}(Y_i - u_i)\mathbb{1}[Y_i > u_i] \tag{21}$$

Further, we use the pinball loss as a proper measure of fit per quantile, previously presented in subsubsection 3.1.2. We can average the score both per quantile and over all quantiles to evaluate the quantile predictions.

3.4.4 Test of equal predictive performance

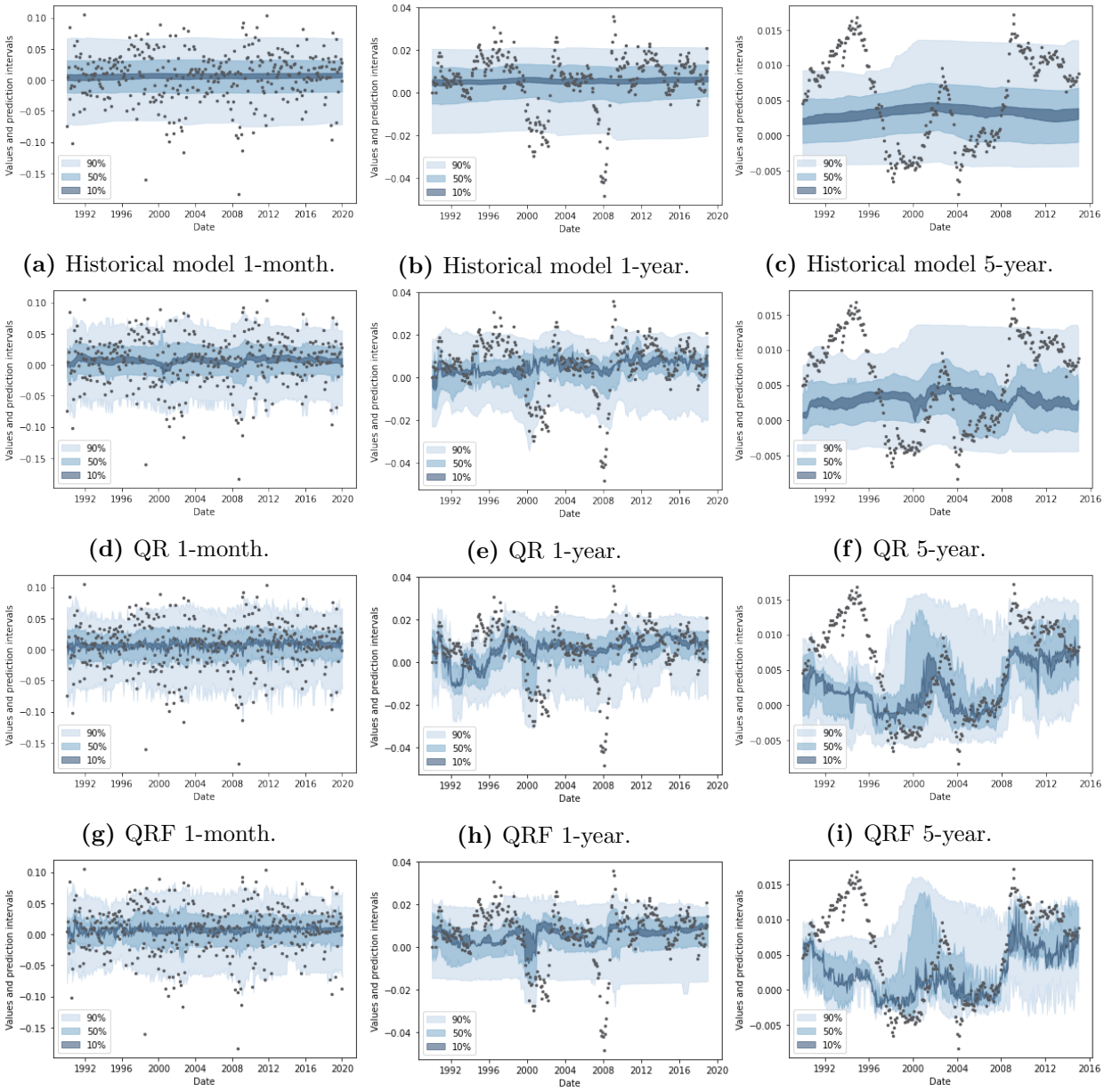
We use the Diebold-Mariano test (DM test) created by Diebold and Mariano (1995) to determine whether forecasts generated by one model are significantly better than those generated by a different model. The test is model-free in that no models are needed, only two forecast error series, alternatively, two scoring result series for evaluating probabilistic forecasts. Thus, the DM test can both be applied to test for equal predictive performance of point and interval forecasts. Denote $d_t = loss(e_{1t}) - loss(e_{2t})$ as the difference in loss (e.g. a quadratic or absolute loss function) between the two forecasts compared, and $\overline{d} = \frac{1}{T}\sum_{i=1}^{T} d_t$ denotes the sample mean of d_t. Under the null hypothesis, the test statistic tends towards the normal distribution, see Equation 22, and states that there is no difference in the predictive accuracy between the two forecast series ($H0 : E[d_t] = 0$). The two-sided alternative hypothesis states that the forecasts are not equally accurate ($H1 : E[d_t] \neq 0$). The null can be rejected if $|DM|$ is greater than a critical value. In practice, it is normal to run the test one-sided, where the alternative hypothesis is that one forecast is less accurate than the other. We will perform the test in such a manner. The DM test requires the loss differentials to be covariance stationary, referred to as the Assumption DM by Diebold (2012) (see Equation 23).

$$DM = \frac{\overline{d}}{\hat{\sigma}_{d_t}} \sim \mathcal{N}(0, 1) \tag{22}$$

$$AssumptionDM : \begin{cases} E[d_t] = \mu \\ cov(d_t, d_{t-\tau}) = \gamma(\tau) \\ 0 < \sigma_{d_t}^2 < \infty \end{cases} \tag{23}$$

4 Results and discussion

The probabilistic forecasts from all the models are presented in Figure 2, which show the 10%, 50% and 90% PIs of the 1-month, 1-year and 5-year ERP. The points are the realised ERP. As seen, the monthly realised ERP is volatile, seemingly acting like a white noise process, which partly explains why the 1-month risk premium is difficult to forecast. For the 1-year and 5-year horizons, the realised ERP shows signs of emerging patterns.



(a) Historical model 1-month. **(b)** Historical model 1-year. **(c)** Historical model 5-year.

(d) QR 1-month. **(e)** QR 1-year. **(f)** QR 5-year.

(g) QRF 1-month. **(h)** QRF 1-year. **(i)** QRF 5-year.

(j) QGB 1-month. **(k)** QGB 1-year. **(l)** QGB 5-year.

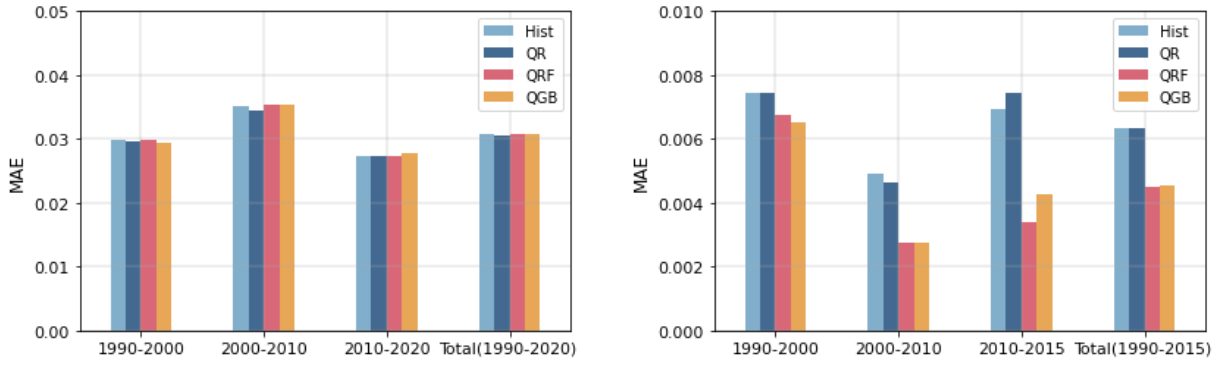
Figure 2: Plot of 10%, 50%, and 90% prediction intervals generated by the models.

The figures illustrate some of the differences between the models. For the historical model, the PIs show little adaptation through time, while QR, QRF, and QGB are more adaptive. When predicting the 1-month ERP, all the models are approximately static, likely because there is no clear pattern for the machine learning models to learn. Conversely, for the 1-year ERP forecasts, the machine learning models show signs of adaptation, and even more on the 5-year horizon. From around the year 2000 and forward, QGB and QRF look reasonably accurate, but they fail to predict the IT bubble spike in the 1990s. This could be because the models had not seen such extreme ERP values in the past or simply due to the incrementally increasing training sets. Regarding the 2008 financial crisis, the 5-year predictions of QRF and QGB are quite accurate, whereas the 1-year predictions are more impacted by this event. The figures further illustrate the differences in sharpness between the models, see, e.g. the far sharper PIs generated by QGB 5-year in Figure 2l relative to QR 5-year in Figure 2f.

In the following, we conduct a systematic analysis of the performance of the models' abilities to produce probabilistic forecasts of the ERP. The point forecasts are first evaluated in subsection 4.1 and the probabilistic properties of the models are then evaluated in subsection 4.2. We further evaluate the importance of each variable in predicting the ERP subsection 4.3. We focus on the results of the 1-month and 5-year predicted ERP because the models for the 1-month and 1-year predictions yield similar outcomes. All analyses of the 1-year predictions are presented in Appendix A.

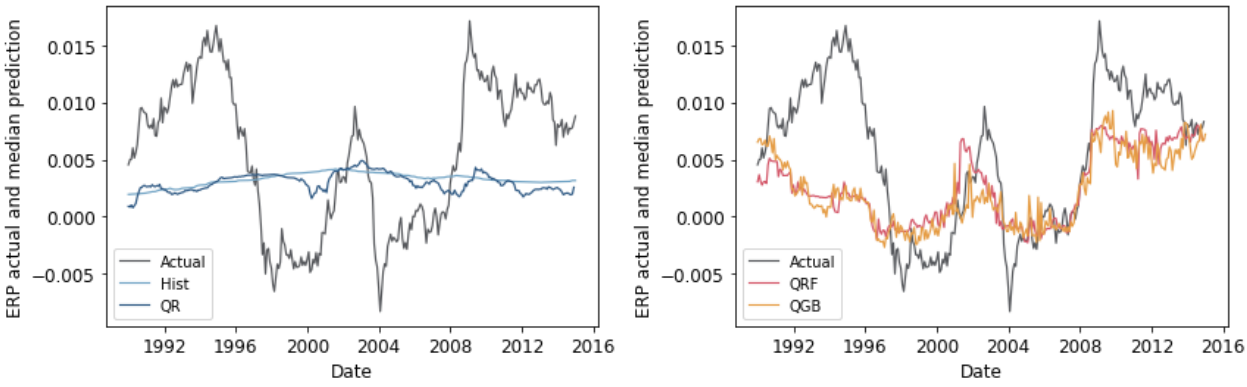
4.1 Evaluation of point estimates

The MAEs of the models are presented in Figure 3, both for three separate decades from 1990 to 2020, as well as the total average error over the whole prediction period. For the 1-month predictions, all the models have almost equal predictive performances, and the DM test confirms that the forecasts produced by QRF and QGB are not significantly better than the forecasts produced by the benchmark models, see Table 4. As we could have expected QRF and QGB to generate better point estimates since they are more complex models, we believe this further substantiates the results of Goyal and Welch (2008) and Goyal et al. (2021), illustrating the difficulty of predicting the 1-month ERP.



(a) 1-month. **(b)** 5-year.

Figure 3: MAE bar charts of 1-month and 5-year point estimates.



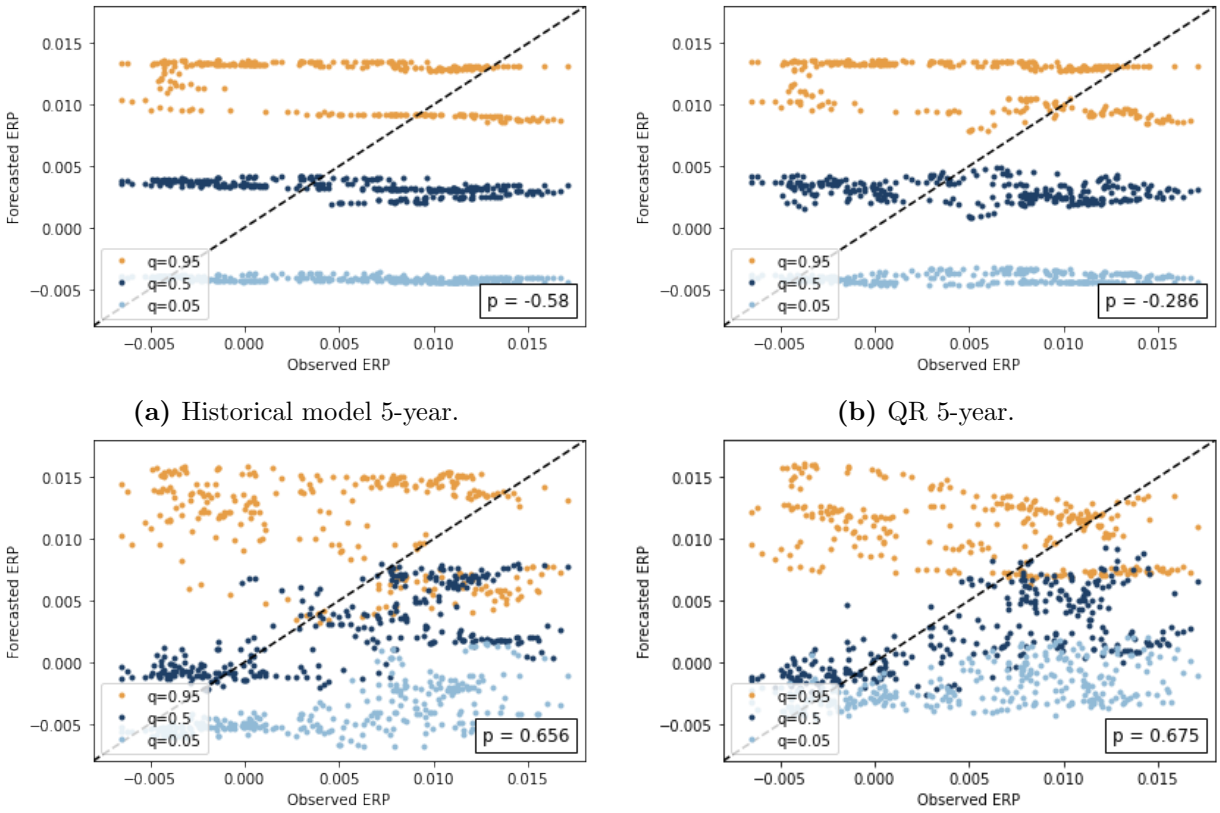
(a) Historical model and QR. **(b)** QRF and QGB.

Figure 4: Median predictions against the realised 5-year average ERP.

For the 5-year ERP, QRF and QGB show considerably lower MAE than the benchmark models from the year 2000 and forward. In the decade between 1990 and 2000, the QRF and QGB do not have considerably lower errors. This is partly due to the effect of the IT bubble. Another reason why QRF and QGB get lower MAE in comparison to the benchmark models through time is because of the incrementally larger amount of data available for training. Their performances against the historical model are the greatest between 2010–2015, with QRF and QGB having 62% and 51% lower MAE, respectively. Over the total period, QRF and QGB have 29% and 27% lower MAE. The DM test supports these findings, i.e. the forecasts produced by QRF and QGB are significantly better than the forecasts produced by the benchmark models. The considerable lower MAE of QRF and QGB compared to the benchmark models proves the machine learning models to be valuable in predicting the long-term level of the ERP. The MAE performances are further illustrated in Figure 4, which shows the 5-year predicted median for the models against the realised ERP. The benchmark models generate somewhat static predictions, while from the year 2000 and forward, QRF and QGB are clearly able to predict the realised values.

	1m		**1yr**		**5yr**	
Model	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**
Panel A: Hist as benchmark						
QR	1.38	0.08	-1.11	0.87	0.87	0.19
QRF	-0.43	0.67	-2.70	1.00	12.10	0.00**
QGB	-0.19	0.57	-0.51	0.69	10.95	0.00**
Panel B: QR as benchmark						
Hist	-1.38	0.92	1.11	0.13	-0.87	0.66
QRF	-1.60	0.95	-2.29	0.99	11.43	0.00**
QGB	-1.20	0.88	0.45	0.33	10.64	0.00**

Table 4: Diebold-Mariano test where H0: Benchmark and model x have the same accuracy, and H1: Benchmark is less accurate than model. (*) denotes significance at the 5% level and (**) denotes significance at the 1% level.



(a) Historical model 5-year.

(b) QR 5-year.

(c) QRF 5-year.

(d) QGB 5-year.

Figure 5: Scatter plot of predicted against realised ERP, with their correlation.

Scatter plots of the 5-year forecasts against the realised ERP are shown in Figure 5, with the Pearson correlation between the forecasts and the realised ERP presented at the bottom right corner of each plot. Ideally, the median forecasts (in dark blue colour) would lay perfectly over the identity line, i.e. the stapled line in the figure. The median plots of the benchmark models are horizontally oriented and show that the medians almost consistently predict the same value, independently of the realised ERP. QRF and QGB are, to a greater extent, aligned with the identity line and show a greater correlation with the realised ERP. Damodaran (2021) evaluates the ERP predictions mainly on a model's correlation with realised ERP, and in regard to this, QRF and QGB yield good results.

The medians of QRF and QGB are mostly below the identity line, which indicates a prediction bias, i.e. the models tend to systematically predict lower ERP than what is realised. This is also supported by the MBE results, which show QRF and QGB to be more biased than the benchmark models. Nonetheless, this additional bias is almost completely incurred during the IT bubble. Plots of MBE are illustrated in Figure 14 in Appendix A. Interestingly, all models considerably reduce their bias between 2000–2010. For QRF and QGB, this is because the models fit the realised ERP better than before. In contrast, subsequent periods of overprediction and underprediction seem to cancel each other out for the benchmark models.

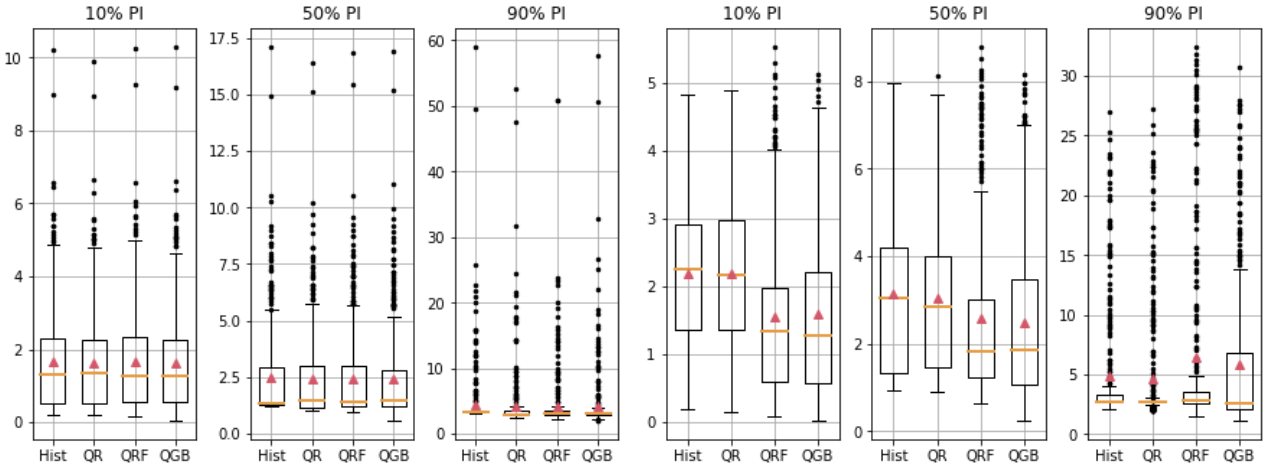
Along with the medians, the predicted 0.05 and 0.95 quantiles are plotted. A greater proportion of the predicted 0.95 quantile is below the stapled line for all models, while almost none of the predicted 0.05 quantile are above it. This observation indicates that the models generally are too low in their predictions of the upper quantiles, which again is a result of the IT bubble. The medians of QRF and QGB are mostly below the identity line, which indicates a prediction bias, in this case that the models tend to systematically predict lower ERP than what is realised. This is also supported by the MBE results, which show QRF and QGB to be more biased than the benchmark models. Nonetheless, this additional bias is almost completely incurred during the IT bubble.

4.2 Evaluation of probabilistic forecasts

We present the results of the probabilistic forecasts below. We begin by presenting the scores of the PIs, before presenting the results of the constituent parts of the score, namely reliability and sharpness.

4.2.1 Scoring of prediction intervals

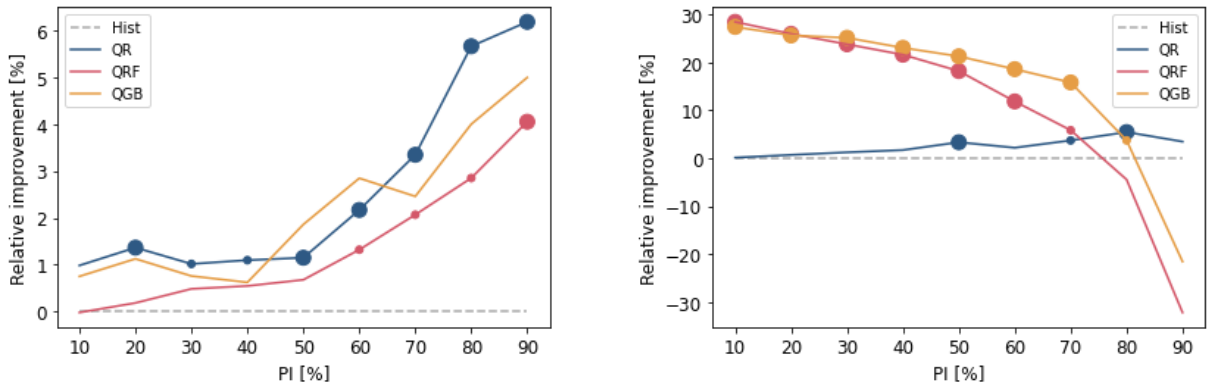
The results of the Wrinkler scoring rule are normalised by the mean and standard deviation of the realised ERP for the purpose of comparison. Figure 6 presents the normalised distribution of the interval score at each time step for the 10%, 50%, and 90% PIs for the 1-month and 5-year predictions. Generally, we observe that the upper tails of the box plots are longer than the lower. This implies that intervals often are far off when they first miss an actual value. Further, the interval scores for 1-month predictions are quite similar, and none of the models stands out, analogous to the results of the point estimates. Conversely, the average interval score of both QRF and QGB show superior performance against the benchmark models on a 5-year horizon, excluding the 90% PI.



(a) 1-month. **(b)** 5-year.

Figure 6: Interval score boxplots for the 1-month and 5-year predictions, with the median (orange line) and mean (red triangle). Lower interval scores are better, and 0 is considered a perfect score.

As the differences between the models in Figure 6 are hard to detect, we further inspect the interval scores by plotting their relative performance against the historical model as a baseline in Figure 7. This figure includes all of the PIs, as well as incorporating the results from the DM test. For 1-month predictions, the mean interval scores of QR, QRF and QGB are performing better than the historical model, as shown in Figure 7a. However, the DM test shows that only the predictions generated by QR and QRF are significantly better than the predictions generated by the historical model for most PIs. In fact, QR ends up as the best-performing model, implying that neither QRF nor QGB is more valuable than the benchmarks in predicting the probabilistic distribution of the 1-month ERP. Additional results from the DM test together with test statistics and p-values can be found in Table 6 in Appendix C.



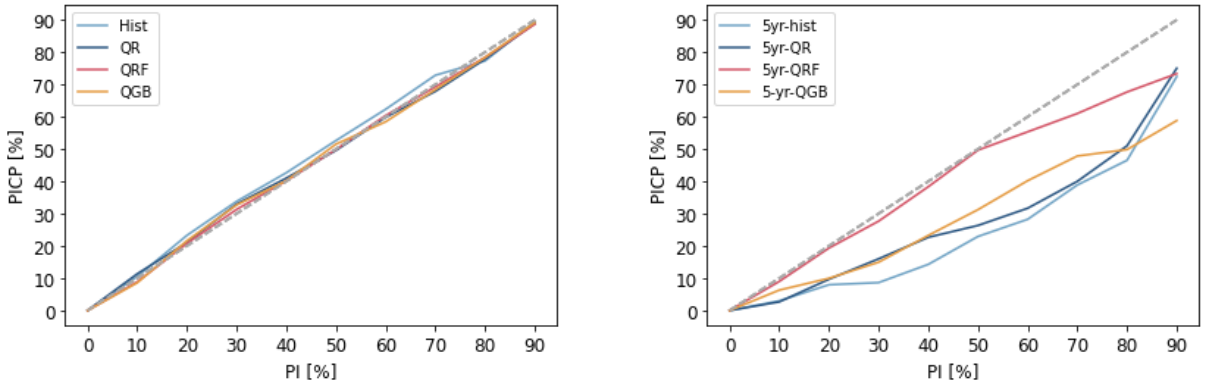
(a) Mean interval score for 1-month predicted ERP. **(b)** Mean interval score for 5-year predicted ERP.

Figure 7: Interval score compared to the historical benchmark model. Significantly better performance at the 1% (5%) significance level is denoted by the large (small) circles.

For 5-year predictions, the DM test confirms that forecasts produced by both QRF and QGB are significantly better than the benchmark models for most of the PIs. Notably, the relative mean interval scores of these models decrease as the interval widens. This reduction is due to the IT bubble disrupting the performance of the wider PIs. To elaborate, the 90% PIs of QRF and QGB are more adaptive in the years preceding and during the IT bubble in comparison to the historical model. At the same time, it is during this period that they incur their relative worse interval scores compared to the historical model. After the bubble, their 90% PIs visibly stabilise at a wide level, see Figure 2i and Figure 2l, doing this to be able to cover similar ERP values in the future. After the IT bubble, QRF and QGB score relatively similar to the benchmark models throughout the remainder of the sampling period. Summarised, the remarkable drop in performance of QRF and QGB compared to the historical model for the 90% PIs is due to a higher score incurred during the initial period. The more central PIs are not considerably affected by the IT bubble for their further performance, i.e. they are able to adjust to the movement of the long-term level of the ERP.

4.2.2 Reliability and sharpness of prediction intervals

In the following part of this section, we present the models' reliability and sharpness to analyse the constituents of the interval score. Figure 8 shows the PICP, which measures the reliability of the forecasting models. A model is perfectly reliable if the line perfectly covers the stapled grey line. All the models seem reliable in predicting the 1-month ERP, and the results from the Kupiec test in Figure 9a verify that they are significantly reliable. However, Figure 9c shows that the conditional coverage is not statistically significant for the greater PIs, i.e. the sequence of "hits" and "misses" for a PI is not random and comes in clusters (see Equation 18).

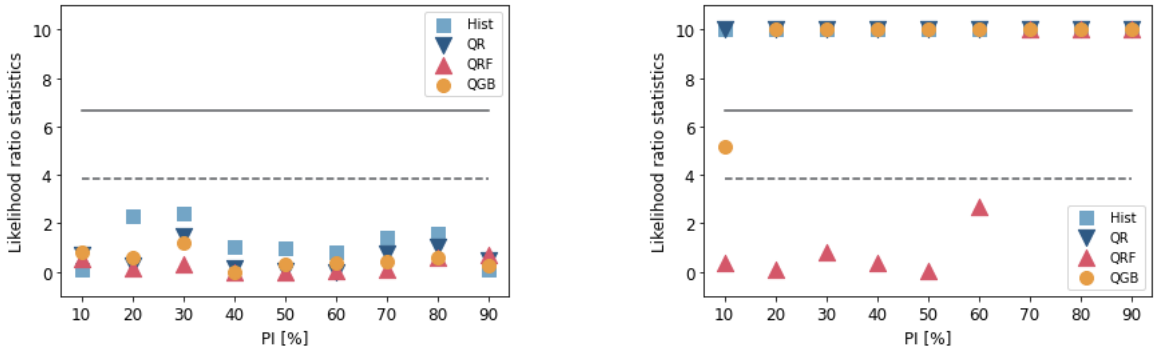


(a) 1-month. **(b)** 5-year.

Figure 8: Reliability plots of 1-month and 5-year predictions.

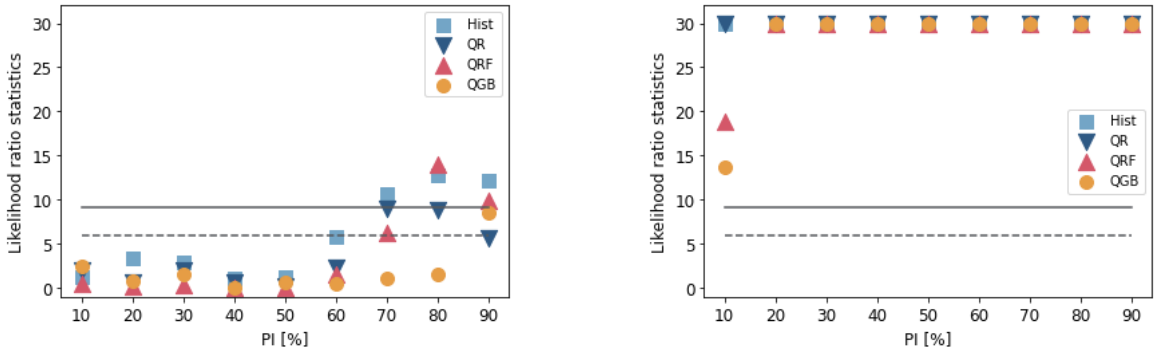
For the 5-year predictions, the models seem less reliable in Figure 8b, and the unconditional coverage test in Figure 9b confirms that only QRF is significantly reliable. The rest of the models are far below the 1:1 line and do not manage to capture the percent of values that the PIs should. When we further apply the Christoffersen test, QRF is no longer significantly reliable. The reason is that the model "misses" too many values sequentially, as previously seen in Figure 2.

Sharpness is conditional to reliability, i.e. to distinguish if a model's forecast is better than another based on sharpness, the models should be approximately equally reliable. The relative sharpness of QR, QRF and QGB compared to the historical benchmark model are presented in Figure 10. For the 1-month ERP predictions, the Christoffersen test shows that all models are significantly reliable for the lower PIs, and we can thus compare their sharpness to analyse the performance of the models. As seen in Figure 10a, QR, QRF and QGB are all consistently sharper than the historical model for the PIs above 10%. Since all models are equally reliable, QR, QRF and QGB outperform the interval score of the historical benchmark due to this slight improvement. For the 5-year predicted ERP, Figure 10b shows that QGB is performing approximately 25% better than the historical benchmark for all of the PIs. Interestingly, the rest of the models (historical, QR and QRF) achieve approximately the same sharpness. However, since QRF outperforms all of the models on reliability, the resulting interval score is greater than that of the benchmark models.



(a) Kupiec test, 1-month.

(b) Kupiec test, 5-year.

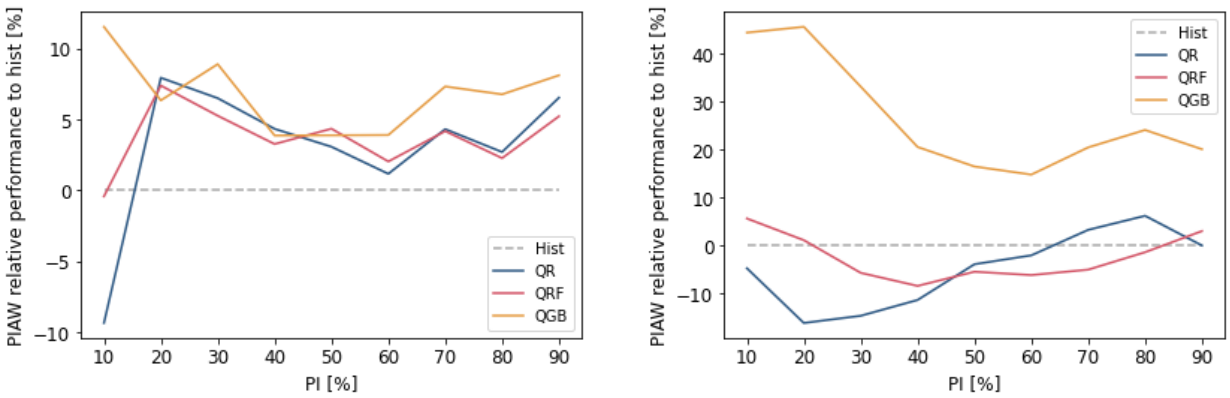


(c) Christoffersen test, 1-month.

(d) Christoffersen test, 5-year.

Figure 9: Likelihood Ratio Statistics for 1% (solid line) and 5% (dashed line) significance level of reliability. LR values greater than 10 or 30 are set to 10 or 30, for the Kupiec and Christoffersen test respectively.

As mentioned previously in subsection 3.3, the quantiles of QR and QRF are always consistent. We emphasize that when we specify consistency amongst the quantiles, the models' predictions are less sharp, which can possibly lead to a lower interval score. We observe this case in Figure 10b, where there is an evident gap between the consistent (historical, QR & QRF) and the inconsistent (QGB) models.



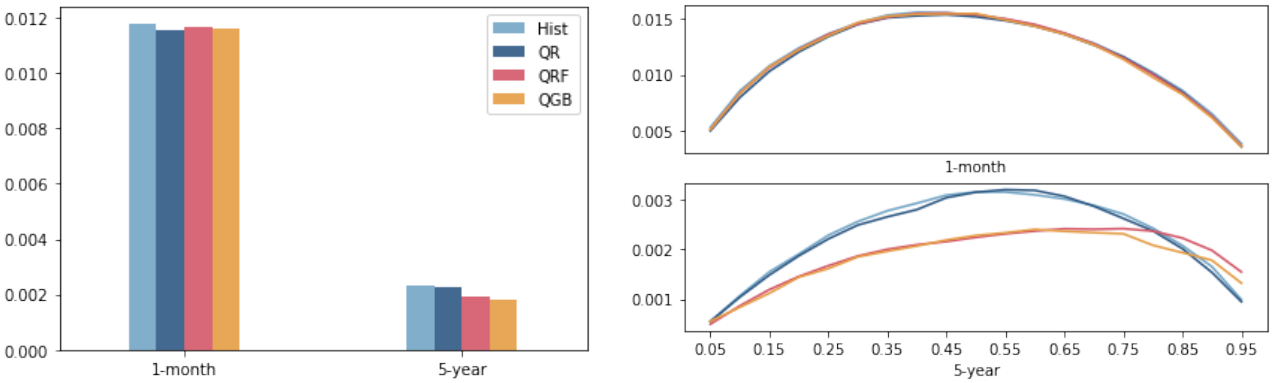
(a) 1-month.

(b) 5-year.

Figure 10: Sharpness of QR, QRF and QGB compared to the historical benchmark model.

4.2.3 Pinball loss for quantile predictions

Figure 11 shows both the average pinball loss per quantile and over all quantiles. For the 1-month predictions, the average loss over all quantiles is similar, while QRF and QGB have a lower average loss for the 5-year predictions. The losses differ for each quantile, where central quantile prediction losses generally contribute more to the total average losses than the outer quantile prediction losses. Since the quantiles of QRF are set to minimise the mean pinball loss of all quantiles, QRF is more reliant on good central estimates when choosing hyperparameters. Conversely, this is not a problem for QR and QGB, as they model each quantile by optimising its pinball loss.



(a) Average pinball loss over all quantiles.

(b) Average pinball loss per quantile.

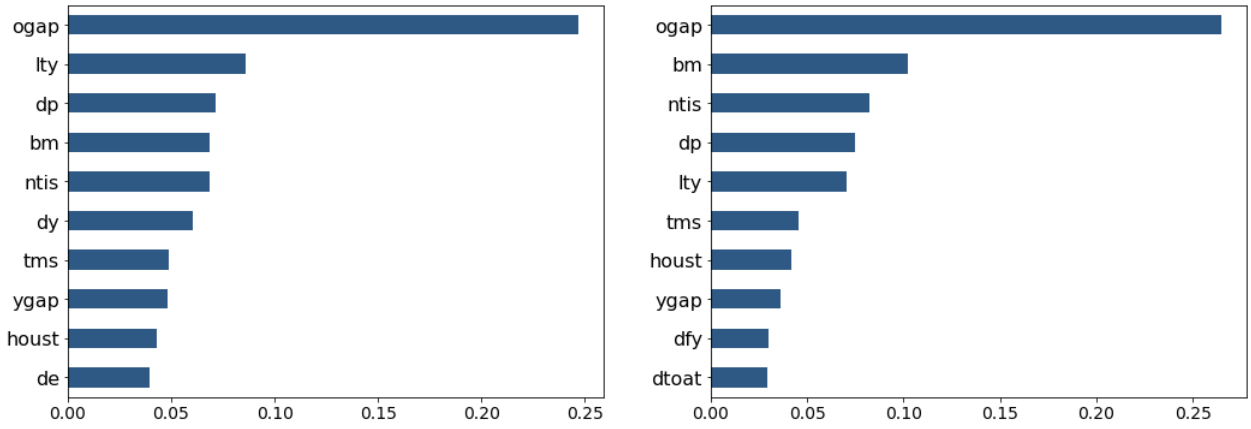
Figure 11: Pinball loss.

4.3 Evaluation of feature importance

QRF and QGB both report the variable or feature importance when growing the tree by averaging the impurity-based feature importance of each tree (see Equation 8). The feature importance provides information on which features from Table 1 the model considers important for predicting the ERP. As QRF and QGB grow trees differently, the feature importances returned by the models end up different. In the following subsections, we analyse the feature importances, focusing on the 5-year predictions as the models yield the best performances for this time horizon. All of the models' feature importances can be found in Appendix D.

4.3.1 The main contributing variables of the models

Figure 12 shows the most important features over the prediction period for the 5-year predictions. The output gap is undoubtedly given the most importance in both models, denoted *ogap* in the figure. This variable is a well-known macroeconomic measure of the difference between the actual output of an economy and the potential output of the economy. Cooper (2009) concludes that the variable has strong predictive power on U.S. stock returns on a monthly basis, but this was later dismissed by



(a) Feature importance in QRF.

(b) Feature importance in QGB.

Figure 12: Feature importance returned by QRF and QGB for the 5-year predictions.

Goyal et al. (2021), who argues that the variable has an insignificant IS coefficient and poor OOS performance. Our results suggest that the variable is a strong decision contributor, although in a multivariate non-linear model and on a longer time horizon.

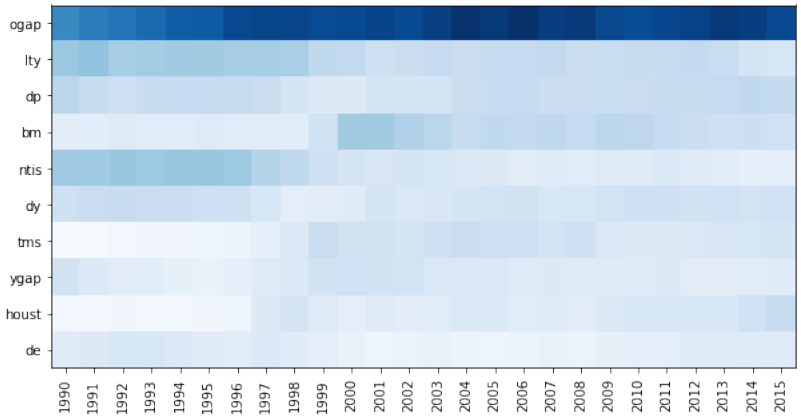
Notably, both models share almost the same ten best performing variables, although not in the exact same magnitude. Long-term yield (*lty*), dividend-price ratio (*dp*), book-to-market ratio (*bm*) and net equity expansion (*ntis*) were all given over 5% importance by both models. In Goyal et al. (2021) all mentioned variables have poor OOS-performance predicting five years ahead. In our case, the variables in GW1 generally have more importance than those in GW2 and MV. However, both gross domestic product (*gdp*) and total new privately-owned housing units (*houst*) from MV are given fairly high importance in the models. All other variables are deemed almost irrelevant.

The individual performance of variables can be partly credited to how they are engineered. Our dataset is mainly engineered for monthly prediction, and augmenting the horizon will favour the variables that capture more historical data. Most of the variables in MV are calculated as the difference between two months or the value at the current time point. The main contributor to the performance of *ogap* is likely explained by it being the only variable based on a regression on all of its previous values, thus capturing fluctuations over a longer time period. The capturing of a trend is also likely a convenience for variables that use 12-month moving sums or averages, such as *ntis* and *dp*. This property of the well-performing variables implies that our variables could benefit from feature engineering to better predict the long-term ERP.

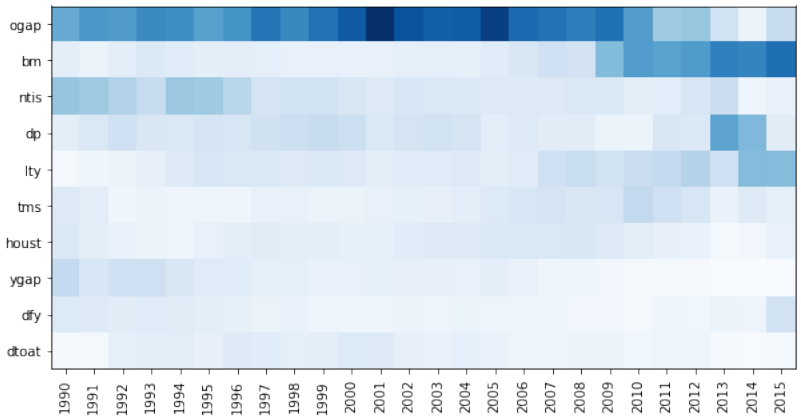
4.3.2 Evolution of the feature importances

A recurring point from previous literature on the prediction of the ERP, e.g. Goyal and Welch (2008) and Goyal et al. (2021), is that certain predictors only perform well due to specific economic events, such that good predictive performances are highly influenced by the starting or ending point of the data

used in the research. An advantage of using a multivariate model, in contrast to using a simple linear model, is the ability of the weights to change over time to capture new information. Figure 13 shows the evolution of the feature importance during the test period. The models have a clear distinction in that the feature importance of QRF is more stable over time in comparison to QGB, primarily due to the EFB attribute of LightGBM described in subsection 3.1.2. *ogap* is the most important variable for the most recent years in the dataset, except in QGB. Another interesting insight is that *ntis* has decreasing importance, implying that the variable has become less influential over the years. Conversely, we see the opposite development with *bm*. In QRF, *bm* start with 3.3% importance in 1990, then increases to its maximum of 10.4% in 2001 before ending as the most important feature with 7.9% in 2015. This development is even greater in QGB, where it finisishes as high as 32.8%.



(a) Evolution of the feature importance for QRF from 1990-2015.



(b) Evolution of the feature importance for QGB from 1990-2015.

Figure 13: The development of the feature importances during the test period. Darker colour yields greater importance. As we stop training five years before the prediction point, 2000 on the x-axis means the feature importance from the training data up to 1995.

4.3.3 Remarks on the interpretation of feature importance

As a final remark, the variables with high feature importance should not uncritically be regarded as having superior predictive power. Even though a variable is given high importance, it may have contributed to bad model performance. E.g. *ntis* is given high importance in the years leading up to the IT bubble, but in this period, the models perform poorly. There is simply no way to know if the high importance of *ntis* made the models prediction better or worse. Furthermore, both QRF and QGB are biased towards variables with higher cardinalities. When the tree-based models create their splits, the number of possible splits grows non-linearly with the cardinality. This can lead to variables with higher cardinality achieving higher feature importance (see Strobl et al. (2007) and Zhu (2020)). The same argument can be made for the variables with low cardinality, which are seldom picked as a split and are thus deemed insignificant in the models.

5 Conclusion and suggestions for further work

We evaluate quantile regression forest (QRF) and quantile gradient boosting (QGB) in the probabilistic forecasting of the ERP. Our findings show that neither QRF nor QGB achieves significantly better performance against the benchmark models in predicting the 1-month point estimate. This supports the findings in previous literature by Goyal and Welch (2008) and Goyal et al. (2021). When predicting the probabilistic distribution, QRF and QGB perform slightly better than the historical benchmark model when evaluated on the Wrinkler interval score. Although, they do not outperform the quantile regression benchmark model. The results regarding reliability, an essential constituent of the interval score, prove both the models to be suited for probabilistic forecasting as an extension to point estimates. Both models produce reliable predictions at a 1% significance level for all prediction intervals (PI) according to the Kupiec test and for the $PIs < 70\%$ according to the Christoffersen test. Nonetheless, as the benchmark models are significantly reliable as well, QRF and QGB are not more valuable for the purpose of probabilistic forecasting.

When predicting the 5-year ERP, we find that QRF and QGB produce significantly more accurate point estimates than the benchmark models, with 29% and 27% lower MAE, respectively. According to the Kupiec test, the results from the probabilistic forecasts show that only QRF produces significantly reliable predictions for the $PIs < 60\%$. When we apply the Christoffersen test, none of the models' predictions are reliable. This test punishes sequential violations, i.e. clusters of when PIs fail to cover the realised ERP, which is evident for all models during several time periods. However, when evaluating the probabilistic forecast by the Wrinkler score, the results show that QRF and QGB create significantly better forecasts than the benchmark models for $PIs < 60\%$, with an average of 22% and 24% lower average interval score, respectively. The conclusion is that QRF and QGB produce remarkable results in predicting the point estimates against the benchmark models on a long-term level and, in addition, are significantly better in forecasting most of the prediction intervals.

Lastly, the tree-based models report the feature importance of each variable for each time step. The variable output gap shows promising potential as a long-term predictor in our models and is by far regarded as the most important variable throughout the majority of our sample period. Nevertheless, further exploration of variables more suited for long-term prediction is needed to prove the predictive power of the output gap.

With our work, we want to emphasise that probabilistic forecasts are well-suited for forecasting the ERP. Following the evaluation of the quantile machine learning models, we find potential areas for further work that can contribute to increase the performance of the forecasts. In general, we advocate for continuing the research of forecasting the long-term ERP and highlight two areas for further work: (i) improving the dataset of predictor variables and (ii) exploring other ways of creating probabilistic forecasts.

For predicting the long-term level of the ERP, the variables should be selected or engineered to be suitable for this purpose. The prevalent issue of our variables is that most of them represent short-term movements, thus being more suitable for short-term predictions. We have excluded variables with a quarterly or annual frequency. However, their inclusion should be considered in more detail when building new datasets since predicting on a long-term time scale implies that the data granularity can be coarser. We also encourage trying other new variables that could be relevant for long-term ERP prediction. As an example, our models are not able to capture the effect of the IT bubble. When looking at our variables, there are generally few that could describe the phenomena of "hype". We believe that variables such as the growth of new investors in the equity market or data that could describe a bandwagon effect could be of interest to implement and analyse.

As for the quantile machine learning models, the technology seems to have some starting difficulties. The inconsistency of the quantiles in QGB, in particular, is problematic. Since the point estimation is relatively successful, experiments should be conducted on alternative ways to build the quantiles around the median, as the QRF does. Alternatively, other ways to create probabilistic forecasts could be explored. We especially advocate investigating quantile regression neural networks to evaluate their performance in ERP prediction, both in predicting power and stability.

References

Alessandrini, S., Davò, F., Sperati, S., M. Benini1, L. and Monache, D. (2014), 'Comparison of the economic impact of different wind power forecast systems for producers', *Advances in Science and Research* **11**(1), 49–53.

Ang, A. and Bekaert, G. (2007), 'Stock return predictability: is it there?', *Review of Financial Studies* **20**(2), 651–707.

Ballings, M., den Poel, D. V., Hespeels, N. and Gryp, R. (2015), 'Evaluating multiple classifiers for stock price direction prediction', *Expert Systems with Applications* **42**(20), 7046–7056.

Basak, S., Kar, S., Saha, S., Khaidem, L. and Dey, S. R. (2019), 'Predicting the direction of stock market prices using tree-based classifiers', *The North American Journal of Economics and Finance* **47**, 552–567.

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

Brotherson, W. T., Eades, K. M., Harris, R. S. and Higgins, R. C. (2015), '"Best practices" in estimating the cost of capital: an update', *Journal of Applied Finance* **23**(1).

Campbell, J. (1987), 'Stock returns and term structure', *Journal of Financial Economics* **18**(2), 373–399.

Campbell, J. Y. and Vuolteenaho, T. (2004), 'Inflation illusion and stock prices', *The American Economic Review* **94**(2), 19–23.

Christoffersen, P. F. (1998), 'Evaluating interval forecasts', *International Economic Review* **39**(4), 841–862.

Clemen, R. T. (1989), 'Combining forecasts: a review and annotated bibliography', *International Journal of Forecasting* **5**(4).

Cochrane, J. H. (2008), 'The dog that did not bark: a defense of return predictability', *The Review of Financial Studies* **21**(4), 1533–1575.

Cooper, I. (2009), 'Time-varying risk premiums and the output gap', *Review of Financial Studies* **22**(7), 2601–2633.

Damodaran, A. (2021), 'Equity risk premiums (ERP): determinants, estimation, and implications - the 2021 edition'.

Dawid, A. P. (1984), 'Statistical theory: the prequential approach (with discussion)', *Journal of the Royal Statistical Society* **147**(2), 278–290.

Diebold, F. X. (2012), 'Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests', *PIER Working Paper* **12**(35).

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–265.

Duarte, F. and Rosa, C. (2015), 'The equity risk premium: a review of models', *FRBNY Economic Policy Review* pp. 39–57.

Fama, E. and French, K. (1998), 'Value versus growth: the international evidence', *Journal of Finance* **53**(6), 1975–1999.

French, M. (1997), *U.S economic history since 1945*, Manchester University Press.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *The Annals of Statistics* **29**(5), 1189–1232.

Garrat, A., Lee, K., Pesaran, M. H. and Shin, Y. (2003), 'Forecast uncertainties in macroeconomic modeling: an application to the U.K. economy', *Journal of the American Statistical Association* **98**(464), 829–838.

Gneiting, T. (2008), 'Editorial: probabilistic forecasting', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **171**(2), 319–321.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007), 'Probabilistic forecasts, calibration and sharpness', *Journal of the Royal Statistical Society* **64**(2), 243–268.

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.

Goodnight, G. T. and Green, S. (2010), 'Rhetoric, risk, and markets: the dot-com bubble', *Quarterly Journal of Speech* **96**(2), 115–140.

Goyal, A. and Welch, I. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *The Review of Financial Studies* **21**(4).

Goyal, A., Welch, I. and Zafirov, A. (2021), A comprehensive look at the empirical performance of equity premium prediction II.

Gu, S., Kelly, B. and Xiu, D. (2020), 'Empirical asset pricing via machine learning', *The Review of Financial Studies* **33**(5), 2223–2273.

Guo, H. (2006), 'On the out-of-sample predictability of stock market returns', *The Journal of Business* **79**(2), 645–670.

Hastie, T., Tibshirani, R. and Friedman, J. (2017), *The elements of statistical learning*, 2 edn, Springer Verlag, New York, USA.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021), *Introduction to statistical learning*, 2 edn, Springer Verlag, New York, USA.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017), 'Lightgbm: a highly efficient gradient boosting decision tree'.

Koenker, R. and Basset, G. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

Kupiec, P. (1995), 'Techniques for verifying the accuracy of risk measurement models', *The Journal of Derivatives* **3**(2), 73–84.

Li, Y., Ng, D. and Swaminathan, B. (2013), 'Predicting market returns using aggregate implied cost of capital', *Journal of Financial Economics* **110**, 419–436.

Li, Y. and Zhu, J. (2008), 'L1-norm quantile regression', *Journal of Computational and Graphical Statistics* **17**(1), 163–185.

Meinshausen, N. (2006), 'Quantile regression forests', *Journal of Machine Learning Research* **7**, 983–999.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D. and Vrontos, S. D. (2014), 'A quantile regression approach to equity premium prediction', *The Journal of Forecasting* **33**(7), 558–576.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D. and Vrontos, S. D. (2019), 'Out-of-sample equity premium prediction: A complete subset quantile regression approach', *The European Journal of Finance* **27**(1), 110–135.

Neely, C. J., Rapach, D. E., Tu, J. and Zhou, G. (2014), 'Forecasting the equity risk premium: the role of technical indicators', *Management Science* **60**(7), 1772–1791.

Nowotarski, J. and Weron, R. (2018), 'Recent advances in electricity price forecasting: A review of probabilistic forecasting', *Renewable and Sustainable Energy Reviews* **81**, 1548–1568.

Onkal, D. and Muradoglu, G. (1994), 'Evaluating probabilistic forecasts of stock prices in a developing stock market', *European Journal of Operational Research* **74**, 350–358.

Quinlan, J. R. (1986), 'Decision trees', *Machine Learning* **1**, 81–106.

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.

Rapach, D. E. and Zhou, G. (2020), Time-series and cross-sectional stock return forecasting: new machine learning methods, *in* E. Jurczenko, ed., 'Machine Learning for Asset Management: New Developments and Financial Applications', Wiley, Hoboken, New Jersey, pp. 1–34.

Segal, M. (2003), 'Machine learning benchmarks and random forest regression', *Technical Report, Center for Bioinformatics Molecular Biostatistics, University of California, San Francisco* .

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007), 'Bias in random forest variable importance measures: Illustrations, sources and a solution." bmc bioinformatics, 8(1), 25', *BMC bioinformatics* **8**, 25.

Vaysse, K. and Lagacherie, P. (2017), 'Using quantile regression forest to estimate uncertainty of digital soil mapping products', *Geoderma* **291**, 55–64.

Verbois, H., Rusydi, A. and Thieryd, A. (2018), 'Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting', *Solar Energy* **173**, 313–327.

Vijh, M., Chandola, D., Tikkiwal, V. A. and Kumar, A. (2020), 'Stock closing price prediction using machine learning techniques', *Procedia Computer Science* **167**, 599–606.

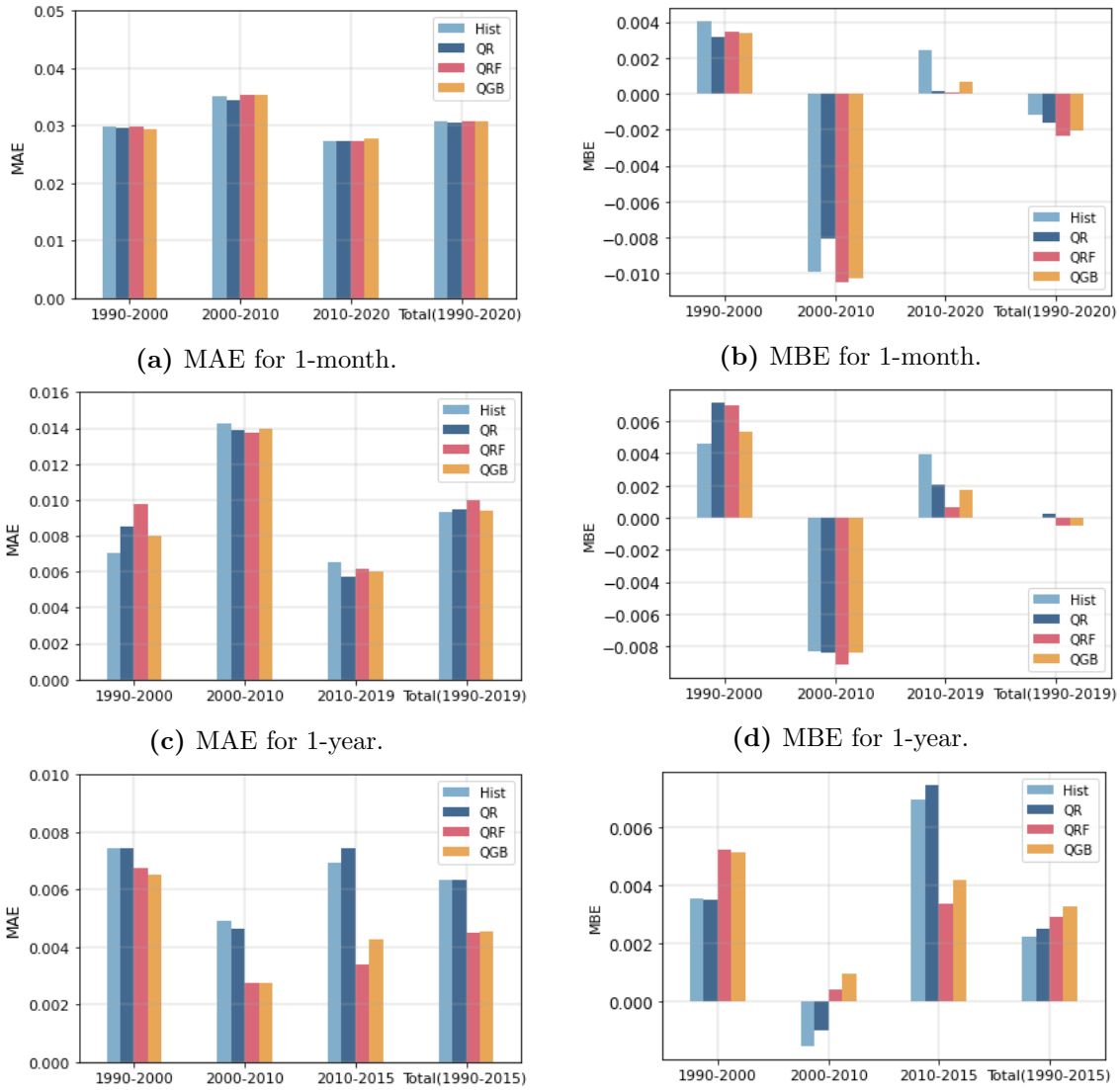
Wolff, D. and Neugebauer, U. (2019), 'Tree-based machine learning approaches for equity market predictions', *Journal of Asset Management* **20**, 273—-288.

Wrinkler, R. L. (1972), 'A decision-theoretic approach to interval estimation', *The Journal of Derivatives* **67**(337), 187–191.

Zhu, T. (2020), 'Analysis on the applicability of the random forest', *Journal of Physics: Conference Series* **1607**(1), 012123.

Appendices

A Additional Graphs



(a) MAE for 1-month.

(b) MBE for 1-month.

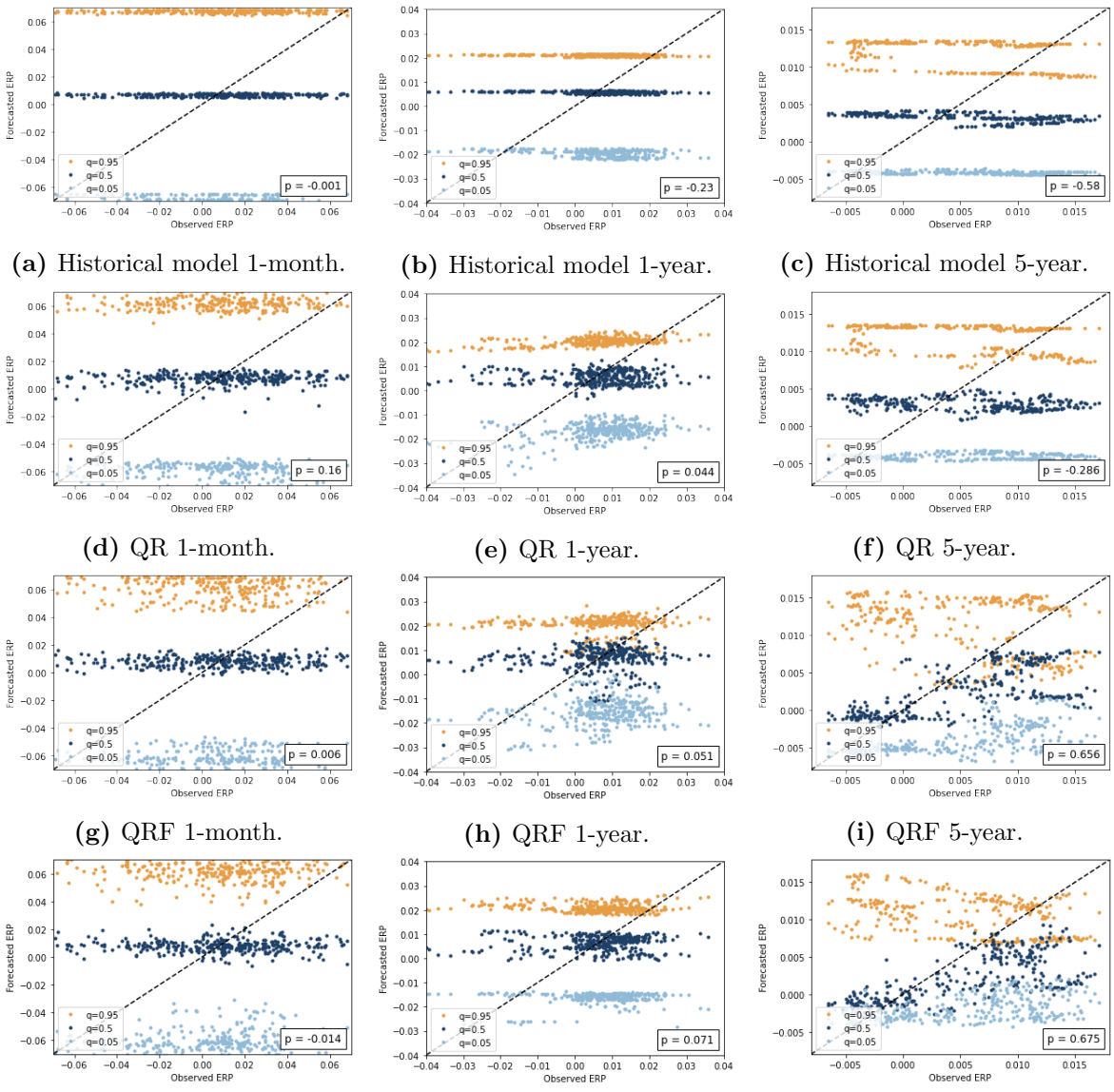
(c) MAE for 1-year.

(d) MBE for 1-year.

(e) MAE for 5-year.

(f) MBE for 5-year.

Figure 14: MAE and MBE bar charts of 1-month, 1-year, and 5-year point predictions.



(a) Historical model 1-month. **(b)** Historical model 1-year. **(c)** Historical model 5-year.

(d) QR 1-month. **(e)** QR 1-year. **(f)** QR 5-year.

(g) QRF 1-month. **(h)** QRF 1-year. **(i)** QRF 5-year.

(j) QGB 1-month. **(k)** QGB 1-year. **(l)** QGB 5-year

Figure 15: Model forecast correlation with realised ERP.

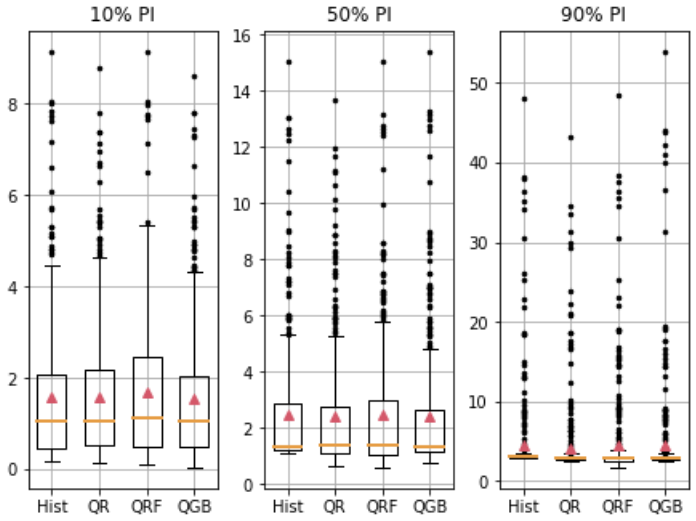
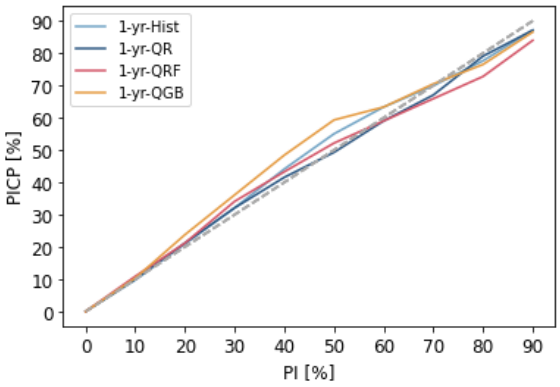
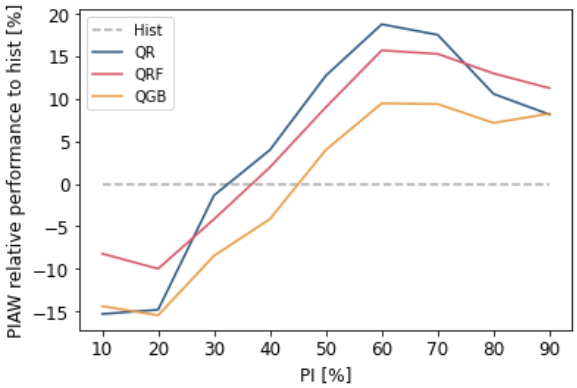


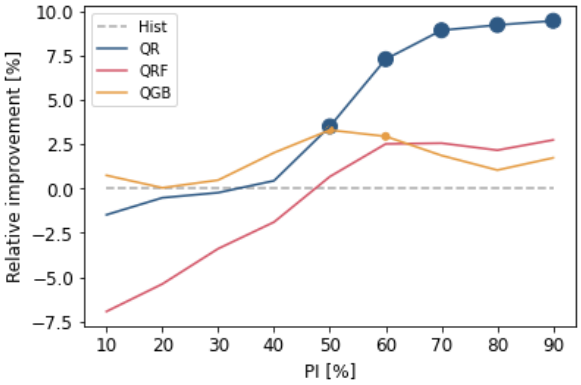
Figure 16: Interval score boxplots of the 1-year ERP with median (line) and mean (triangle).



(a) Reliability 1-year.

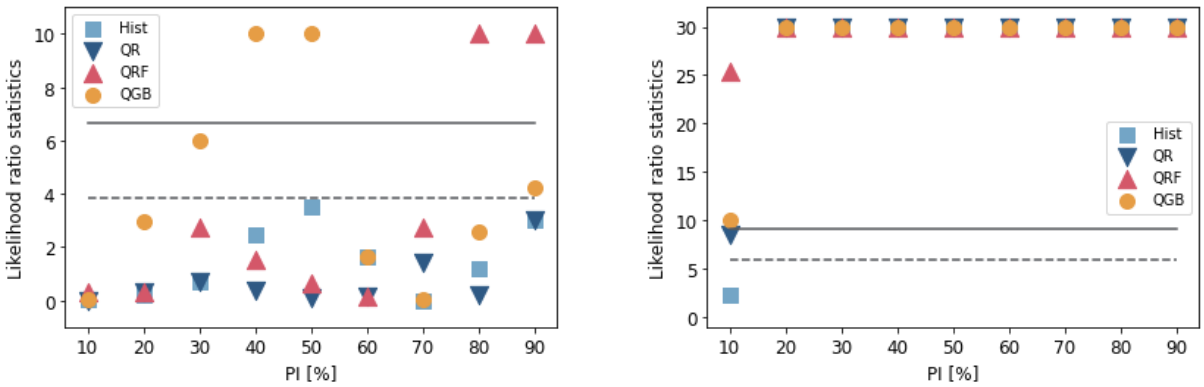


(b) Average sharpness score for the 1-year ERP.



(c) Average interval score for the 1-year ERP.

Figure 17: Results compared to the historic benchmark model.



(a) Kupiec 1-year.

(b) Christoffersen 1-year.

Figure 18: Likelihood Ratio Statistics for 1% (solid line) and 5% (dashed line) significance level of reliability. LR values greater than 10 or 30 are set to 10 or 30, for the Kupiec and Christoffersen test respectively.

B Additional tables

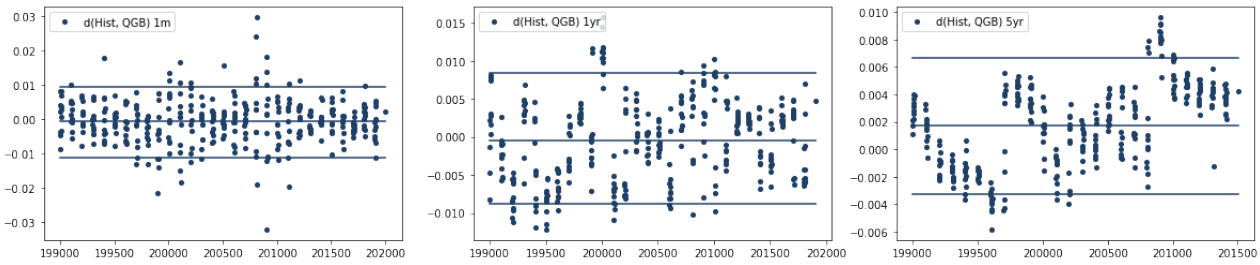
	Mean	Median	Min	Max	Std	25th	75th	Kurt	Skew
erp	0.004	0.009	-0.248	0.149	0.043	-0.019	0.033	2.400	-0.669
dp	-3.613	-3.536	-4.524	-2.753	0.393	-3.945	-3.349	-0.786	-0.095
dy	-3.607	-3.529	-4.531	-2.751	0.392	-3.937	-3.347	-0.757	-0.100
ep	-2.863	-2.884	-4.836	-1.899	0.425	-3.098	-2.675	2.820	-0.568
de	-0.750	-0.778	-1.244	1.380	0.303	-0.913	-0.598	15.867	2.771
bm	0.486	0.431	0.121	1.207	0.256	0.286	0.637	-0.159	0.839
svar	0.002	0.001	0.000	0.073	0.005	0.001	0.002	127.981	10.246
ntis	0.010	0.013	-0.056	0.051	0.020	-0.003	0.024	0.007	-0.582
tbl	0.045	0.046	0.000	0.163	0.032	0.022	0.061	0.857	0.740
lty	0.062	0.059	0.006	0.148	0.028	0.042	0.080	0.132	0.627
ltr	0.006	0.004	-0.112	0.152	0.029	-0.010	0.023	2.593	0.437
tms	0.018	0.017	-0.037	0.046	0.014	0.007	0.029	-0.221	-0.226
dfy	-0.010	-0.009	-0.034	-0.003	0.004	-0.012	-0.007	4.639	-1.793
dfr	0.000	-0.001	-0.074	0.098	0.015	-0.006	0.006	7.525	0.676
infl	0.003	0.003	-0.019	0.018	0.004	0.001	0.005	3.055	0.041
ogap	1.278	1.267	0.974	1.415	0.074	1.219	1.345	-0.443	-0.145
ygap	-4.711	-4.740	-6.360	-3.660	0.395	-4.940	-4.450	1.080	-0.239
rdsp	3.047	2.785	1.779	12.341	1.096	2.379	3.317	18.296	3.328
gip	0.002	0.003	-0.136	0.062	0.010	-0.002	0.006	63.063	-4.168
rpi	0.003	0.003	-0.051	0.123	0.007	0.001	0.005	102.196	5.049
clf16ov	0.001	0.001	-0.040	0.017	0.003	0.000	0.003	42.217	-2.890
unrate	0.002	0.000	-0.176	2.341	0.093	-0.020	0.018	556.588	22.062
uempmean	0.002	0.000	-0.574	0.458	0.048	-0.021	0.024	41.518	-0.518
houst	7.220	7.279	6.170	7.822	0.310	7.076	7.412	1.213	-0.996
bogmbase	0.000	0.001	-0.160	0.152	0.021	-0.007	0.008	16.807	0.156
gdp	-0.012	0.012	-8.451	2.891	1.267	-0.561	0.772	4.883	-1.124

Table 5: Descriptive statistics

Prediction interval	10		20		30		40		50		60		70		80		90	
	Stat	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**
Panel A: 1-month																		
Hist - QR	1.53	0.06	2.43	0.01	1.91	0.03*	2.25	0.01**	2.39	0.01**	3.37	0.00**	3.47	0.00**	4.28	0.00**	3.67	0.00**
Hist - QRF	-0.02	0.51	0.26	0.40	0.67	0.25	0.83	0.20	1.12	0.13	1.82	0.03*	2.17	0.02*	2.31	0.01**	2.37	0.01**
Hist - QGB	0.10	0.46	0.16	0.44	0.10	0.46	0.08	0.47	0.30	0.38	0.45	0.32	0.37	0.35	0.58	0.28	0.61	0.27
QR - Hist	-1.53	0.94	-2.43	0.99	-1.91	0.97	-2.25	0.99	-2.39	0.99	-3.37	1.00	-3.47	1.00	-4.28	1.00	-3.67	1.00
QR - QRF	-1.34	0.91	-1.70	0.95	-0.81	0.79	-0.80	0.79	-0.75	0.77	-1.20	0.88	-1.46	0.93	-2.30	0.99	-1.26	0.90
QR - QGB	-0.07	0.53	-0.07	0.53	-0.08	0.53	-0.12	0.55	0.09	0.46	0.09	0.47	-0.18	0.57	-0.28	0.61	-0.17	0.57
Panel B: 1-year																		
Hist - QR	-1.15	0.87	-0.40	0.65	-0.17	0.57	0.31	0.38	2.39	0.01**	4.20	0.00**	5.69	0.00**	5.59	0.00**	4.25	0.00**
Hist - QRF	-2.64	1.00	-2.14	0.98	-1.37	0.91	-0.80	0.79	0.30	0.38	1.15	0.13	1.18	0.12	0.94	0.17	0.89	0.19
Hist - QGB	0.48	0.32	0.03	0.49	0.31	0.38	1.24	0.11	1.99	0.02*	1.77	0.04	1.13	0.13	0.62	0.27	0.71	0.24
QR - Hist	1.15	0.13	0.40	0.35	0.17	0.43	-0.31	0.62	-2.39	0.99	-4.20	1.00	-5.69	1.00	-5.59	1.00	-4.25	1.00
QR - QRF	-2.27	0.99	-2.06	0.98	-1.42	0.92	-1.10	0.86	-1.36	0.91	-2.58	0.99	-3.68	1.00	-3.99	1.00	-2.58	0.99
QR - QGB	1.61	0.05*	0.41	0.34	0.53	0.30	1.08	0.14	-0.14	0.55	-2.91	1.00	-4.85	1.00	-5.10	1.00	-3.27	1.00
Panel C: 5-year																		
Hist - QR	0.09	0.47	0.54	0.30	0.97	0.17	1.34	0.09	2.67	0.00**	1.61	0.05*	2.30	0.01**	2.99	0.00**	1.62	0.05*
Hist - QRF	12.04	0.00**	11.08	0.00**	10.06	0.00**	9.15	0.00**	7.54	0.00**	4.59	0.00**	1.91	0.03*	-1.39	0.92	-7.25	1.00
Hist - QGB	11.03	0.00**	10.35	0.00**	10.53	0.00**	9.78	0.00**	9.07	0.00**	7.86	0.00*	6.25	0.00**	1.68	0.05*	-7.47	1.00
QR - Hist	-0.09	0.53	-0.54	0.70	-0.97	0.83	-1.34	0.91	-2.67	1.00	-1.61	0.95	-2.30	0.99	-2.99	1.00	-1.62	0.95
QR - QRF	10.81	0.00**	9.71	0.00**	8.84	0.00**	8.01	0.00**	6.04	0.00**	3.68	0.00**	0.75	0.23	-2.78	1.00	-6.87	1.00
QR - QGB	10.24	0.00**	9.35	0.00**	9.26	0.00**	8.47	0.00**	7.21	0.00**	6.47	0.00**	4.76	0.00**	-0.61	0.73	-6.58	1.00

Table 6: Diebold-Mariano test where H0: Benchmark and model x have the same accuracy and H1: Benchmark is less accurate than model. (*) denotes significance at the 5% level and (**) denotes significance at the 1% level.

C Evaluation of loss differentials



(a) 1 month **(b)** 1 year **(c)** 5 year

Figure 19: Absolute loss differentials for the historic model against QGB on predicting the 1-month, 1-year, and 5-year ERP. The lines show the mean and the 5% and 95% quantiles.

	1m		1yr		5yr	
	Stat	p-val	Stat	p-val	Stat	p-val
Panel A: Hist as benchmark						
d(Hist,QR)	-6.09	0.01	-4.95	0.01	-4.10	0.01
d(Hist,QRF)	-7.23	0.01	-5.11	0.01	-3.46	0.05
d(Hist,QGB)	-5.70	0.01	-5.49	0.01	-3.37	0.06
Panel B: QR as benchmark						
d(QR,Hist)	-6.09	0.01	-4.95	0.01	-4.10	0.01
d(QR,QRF)	-6.66	0.01	-4.18	0.01	-3.66	0.03
d(QR,QGB)	-5.48	0.01	-5.75	0.01	-3.54	0.04

Table 7: Augmented Dickey-Fuller for the absolute loss differentials for QRF and QGB against the historic model and QR. The table shows that for all the 1 month and 1 year differential-series based on the point estimates, we can reject the null of a unit root in the series with a significance at the 1% level. For the 5 year series, we can reject all but one loss differential series at the 5% significance level. The test is performed with lag order 7.

Prediction interval	10		20		30		40		50		60		70		80		90	
	Stat	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**	**Stat**	**p-val**
Panel A: 1-month																		
d(Hist,QR)	-5.64	0.01	-5.74	0.01	-6.01	0.01	-5.50	0.01	-5.43	0.01	-5.66	0.01	-5.43	0.01	-5.10	0.01	-6.06	0.01
d(Hist,QRF)	-7.08	0.01	-7.00	0.01	-7.02	0.01	-6.79	0.01	-6.78	0.01	-6.60	0.01	-5.74	0.01	-5.99	0.01	-5.12	0.01
d(Hist,QGB)	-8.73	0.01	-8.87	0.01	-8.73	0.01	-9.07	0.01	-9.09	0.01	-8.98	0.01	-9.08	0.01	-8.66	0.01	-8.44	0.01
d(QR,Hist)	-5.64	0.01	-5.74	0.01	-6.01	0.01	-5.50	0.01	-5.43	0.01	-5.66	0.01	-5.43	0.01	-5.10	0.01	-6.06	0.01
d(QR,QRF)	-6.98	0.01	-7.04	0.01	-6.63	0.01	-6.49	0.01	-6.54	0.01	-6.96	0.01	-6.48	0.01	-6.01	0.01	-5.99	0.01
d(QR,QGB)	-8.80	0.01	-8.91	0.01	-8.68	0.01	-8.86	0.01	-9.07	0.01	-9.14	0.01	-8.72	0.01	-8.47	0.01	-9.48	0.01
Panel B: 1-year																		
d(Hist,QR)	-4.75	0.01	-4.95	0.01	-4.85	0.01	-4.60	0.01	-4.46	0.01	-4.75	0.01	-5.27	0.01	-6.01	0.01	-5.34	0.01
d(Hist,QRF)	-5.05	0.01	-4.96	0.01	-4.61	0.01	-4.57	0.01	-4.61	0.01	-4.83	0.01	-4.96	0.01	-5.68	0.01	-6.30	0.01
d(Hist,QGB)	-5.53	0.01	-5.36	0.01	-5.32	0.01	-5.23	0.01	-5.35	0.01	-5.21	0.01	-6.17	0.01	-7.01	0.01	-6.25	0.01
d(QR,Hist)	-4.75	0.01	-4.95	0.01	-4.85	0.01	-4.60	0.01	-4.46	0.01	-4.75	0.01	-5.27	0.01	-6.01	0.01	-5.34	0.01
d(QR,QRF)	-4.12	0.01	-4.40	0.01	-4.08	0.01	-4.06	0.01	-3.81	0.02	-4.06	0.01	-4.42	0.01	-4.41	0.01	-5.21	0.01
d(QR,QGB)	-5.46	0.01	-5.84	0.01	-5.53	0.01	-5.49	0.01	-5.03	0.01	-4.65	0.01	-5.30	0.01	-5.04	0.01	-5.58	0.01
Panel B: 5-year																		
d(Hist,QR)	-4.97	0.01	-4.83	0.01	-5.04	0.01	-5.21	0.01	-5.58	0.01	-5.39	0.01	-4.53	0.01	-5.51	0.01	-5.72	0.01
d(Hist,QRF)	-3.49	0.04	-3.48	0.05	-3.36	0.06	-3.15	0.10	-3.07	0.13	-3.06	0.13	-3.10	0.11	-3.35	0.06	-2.84	0.22
d(Hist,QGB)	-3.52	0.04	-3.43	0.05	-3.65	0.03	-3.53	0.04	-3.16	0.10	-3.06	0.13	-2.99	0.16	-4.10	0.01	-4.02	0.01
d(QR,Hist)	-4.97	0.01	-4.83	0.01	-5.04	0.01	-5.21	0.01	-5.58	0.01	-5.39	0.01	-4.53	0.01	-5.51	0.01	-5.72	0.01
d(QR,QRF)	-3.89	0.01	-3.89	0.01	-3.69	0.02	-3.56	0.04	-3.53	0.04	-3.56	0.04	-3.53	0.04	-3.52	0.04	-2.97	0.17
d(QR,QGB)	-3.82	0.02	-3.83	0.02	-4.09	0.01	-3.94	0.01	-3.51	0.04	-3.42	0.05	-3.52	0.04	-3.81	0.02	-4.27	0.01

Table 8: Augmented Dickey-Fuller for the absolute loss differentials for QRF and QGB against the historic model and QR. The table shows that for almost all the differential-series based on the prediction intervals we can reject the null of a unit root in the series. with a significance at the 1% level.

D Additional feature importance plots.

This section of the appendix shows all feature importance plots for all the time horizons. The feature importance per quantile is only of relevance for the QGB models, but we solely include the 5-year QGB feature importance as it exemplifies all the other time horizons. For all plots, darker colour yields greater importance.

D.1 Feature importance for each quantile

In quantile gradient boosting the model needs to be trained for each quantile, therefore each quantile accordingly has a feature importance. As Figure 13b shows the average feature importance of all quantiles. We have changed the dimensions by calculating the feature importance of each quantile averaged over all the years in Figure 20. Despite the features having large differences inbetween them at a given time, e.g. *dp* ranges from 0.01 to 0.09 at 2015-01, the differences smooth out over the test period.

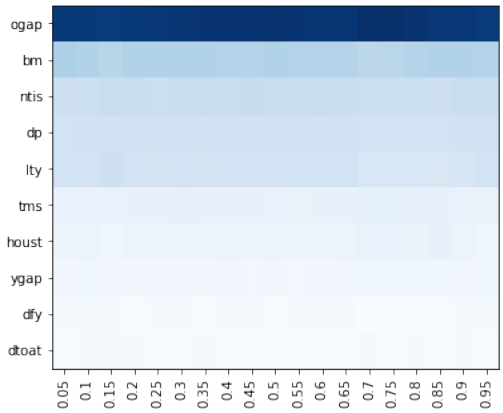
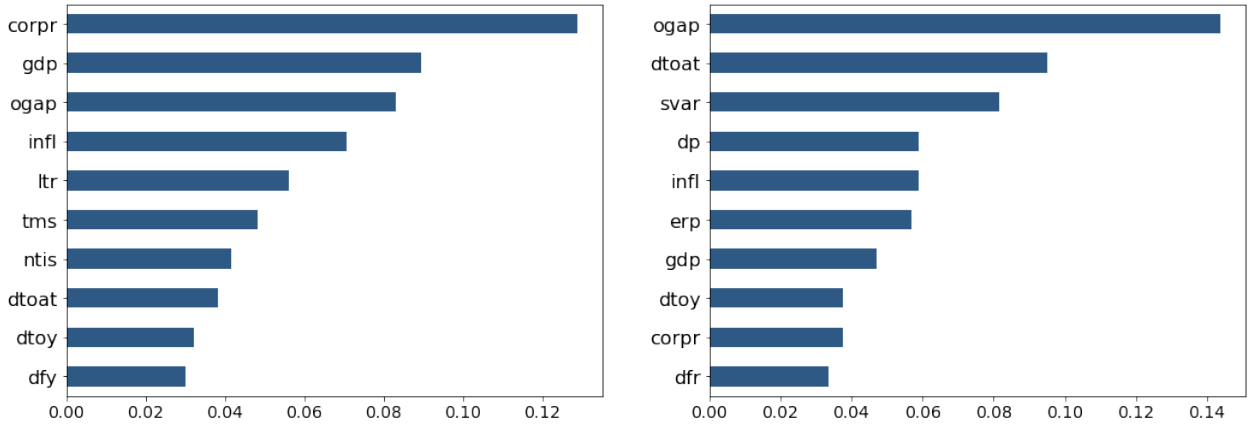


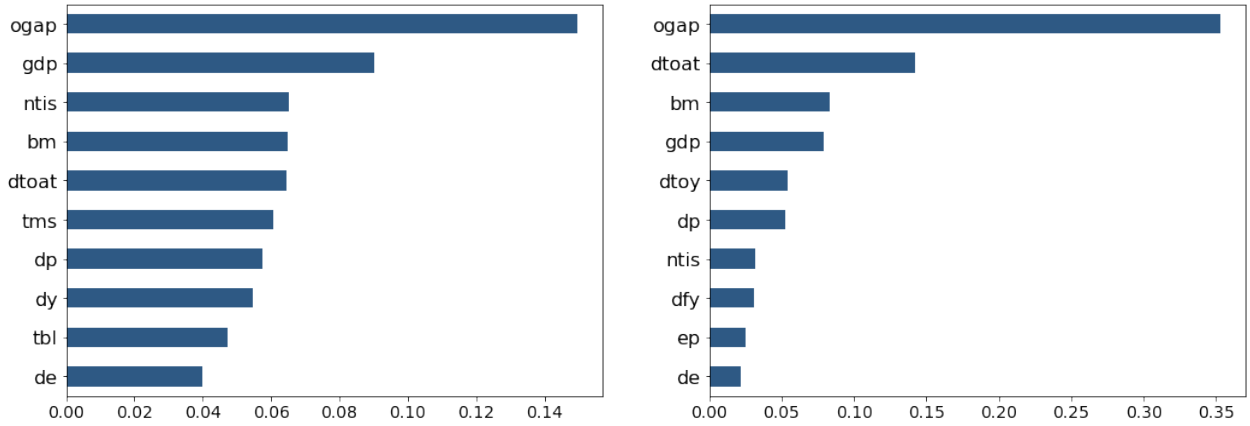
Figure 20: Feature importance for the different quantiles in quantile gradient boosting.

D.2 Additional plots for the best feature importances of 1-month and 1-year predictions.



(a) Feature importance for QRF. **(b)** Feature importance for QGB.

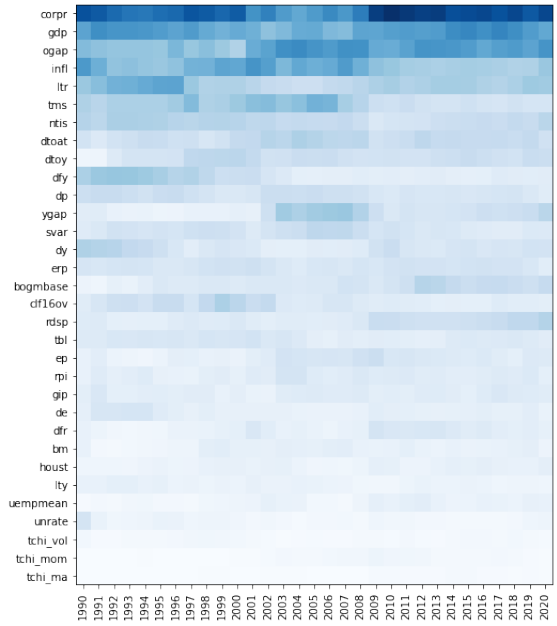
Figure 21: Top 10 feature importances returned by QRF and QGB for the 1-month predictions.



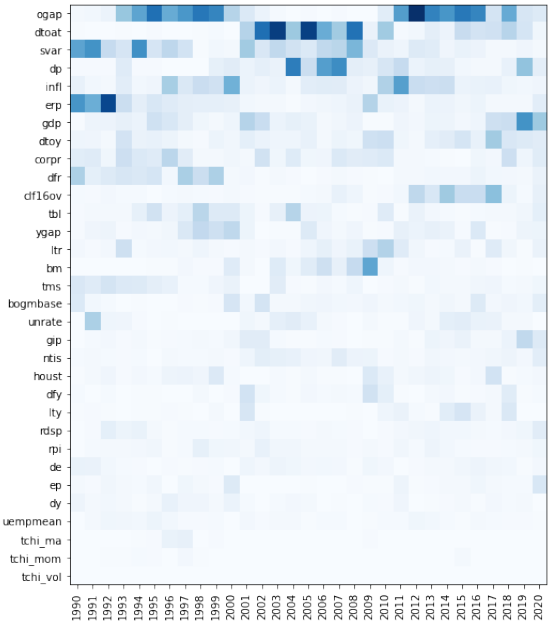
(a) Feature importance for QRF. **(b)** Feature importance for QGB.

Figure 22: Top 10 feature importances returned by QRF and QGB for the 1-year predictions.

D.3 Additional plots for the evolution of the feature importance.

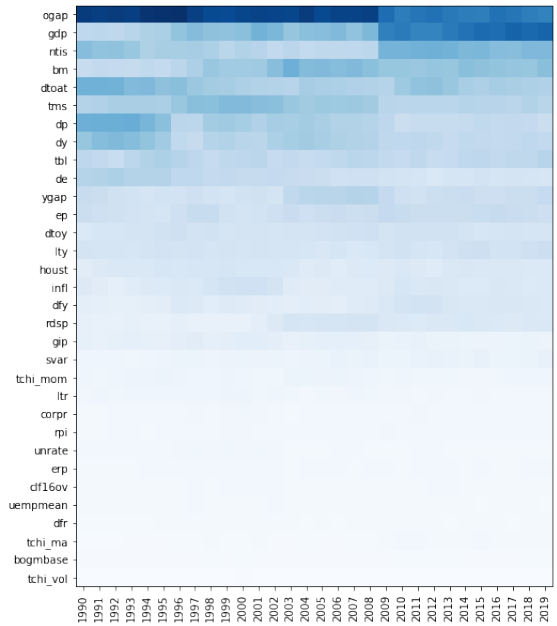


(a) Evolution of the feature importance for QRF.

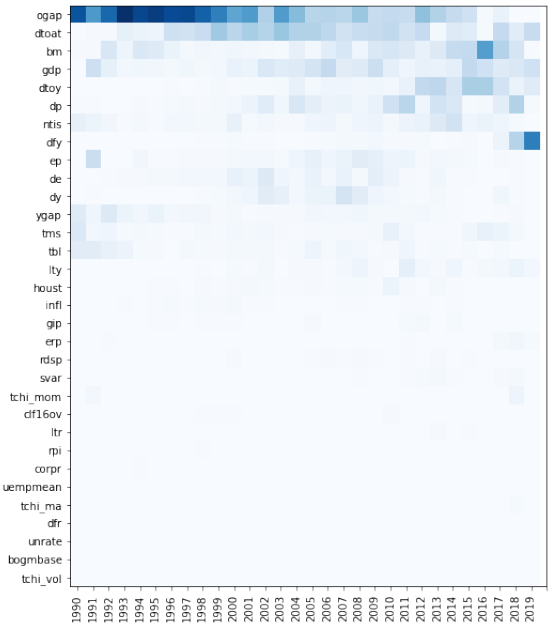


(b) Evolution of the feature importance for QGB.

Figure 23: The development of the feature importances when predicting on the 1-month ERP.



(a) Evolution of the feature importance for QRF.

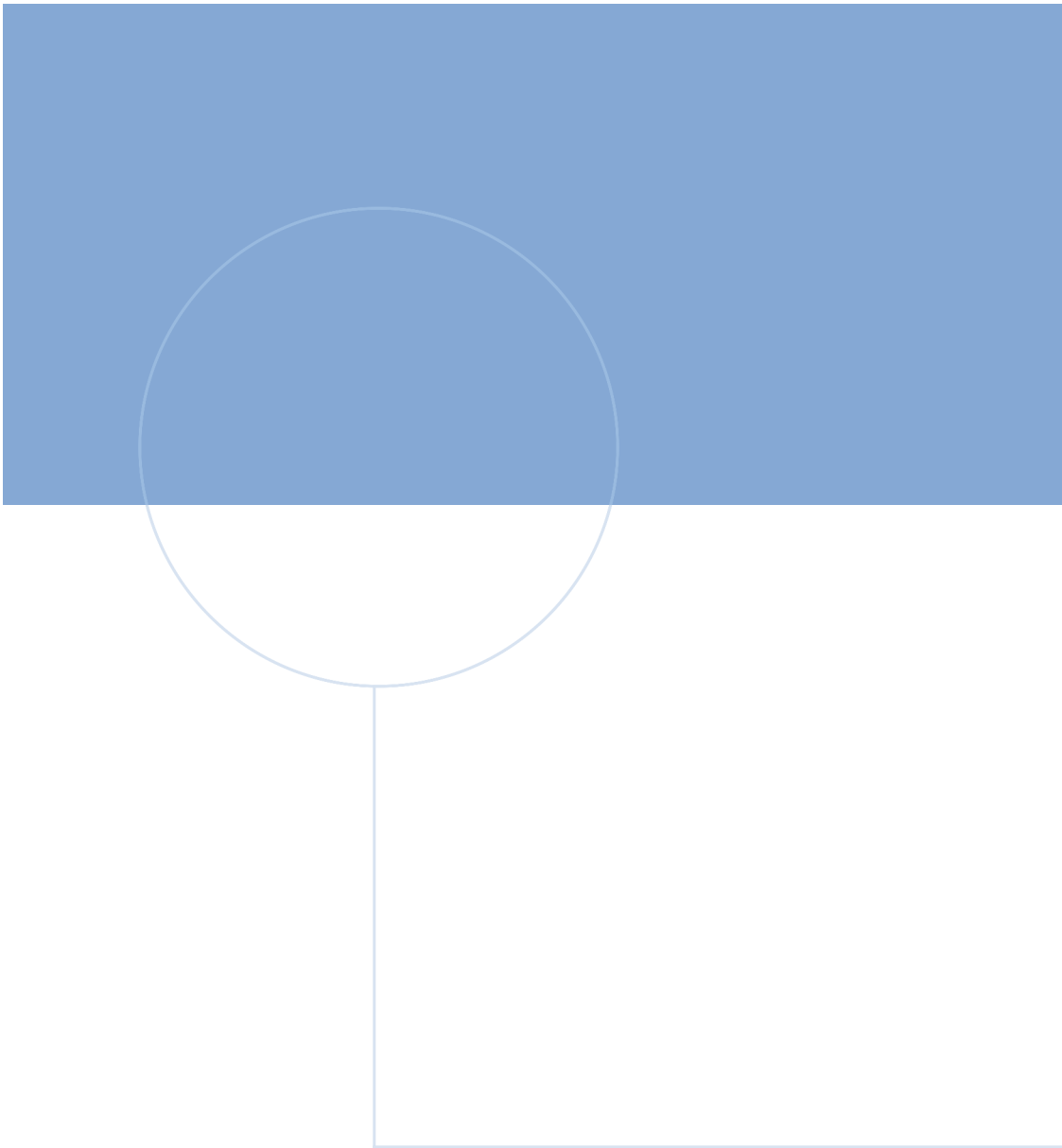


(b) Evolution of the feature importance for QGB.

Figure 24: The development of the feature importances when predicting on the 1-year ERP.

(a) Evolution of the feature importance for QRF. **(b)** Evolution of the feature importance for QGB.

Figure 25: The development of the feature importances when predicting on the 5-year ERP.



NTNU

Norwegian University of
Science and Technology