Statistical post processing of extreme weather forecasts

Velthoen, J.J.

DOI
10.4233/uuid:f6a6096d-9eb1-4a77-b376-34e01b817011

Publication date
2022

Document Version
Final published version

Citation (APA)
Velthoen, J. J. (2022). *Statistical post processing of extreme weather forecasts* (1 ed.). [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:f6a6096d-9eb1-4a77-b376-34e01b817011

STATISTICAL POST PROCESSING OF EXTREME WEATHER FORECASTS

STATISTICAL POST PROCESSING OF EXTREME WEATHER FORECASTS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 14 september 2022 om 15:00 uur

door

Jasper Jonathan VELTHOEN

Master of Science in Applied Mathematics,
Technische Universiteit Delft, Nederland,
geboren te Driebergen-Rijsenburg, Nederland.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. G. Jongbloed
copromotor: Dr. J.J. Cai

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. G. Jongbloed,	Technische Universiteit Delft
Dr. J.J. Cai,	VU Amsterdam
Onafhankelijke leden:	
Prof. dr. A.J. Cabo,	Technische Universiteit Delft
Prof. dr. V. Chavez-Demoulin	Université de Lausanne
Dr. P. Naveau,	Laboratoire des Sciences du Climat et de l'Environnement
Prof. dr. C. Zhou,	Erasmus Universiteit Rotterdam
Prof. dr. ir. A.W. Heemink	Technische Universiteit Delft, reservelid
Overige leden:	
Dr. M. Schmeits	Koninklijk Nederlands Meteorologisch Instituut

Printed by: Print Service Ede

Front & Back: Jasper Velthoen

ISBN 978-90-832727-2-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

Summary vii

Samenvatting ix

1 Introduction 1
- 1.1 Statistical Post-Processing for weather forecast . . . 1
  - 1.1.1 Numerical Weather Prediction Models . . . 2
  - 1.1.2 Ensemble Prediction Systems . . . 2
  - 1.1.3 Statistical post processing . . . 3
- 1.2 Contribution of this thesis . . . 5
  - 1.2.1 Forecasting Extreme events . . . 5
  - 1.2.2 Variable selection for statistical post processing . . . 7
  - 1.2.3 Outline . . . 7

2 Improving extreme precipitation forecast using extreme quantile regression 9
- 2.1 Introduction . . . 9
- 2.2 Model and Estimation . . . 11
- 2.3 Asymptotic Properties . . . 13
- 2.4 Bandwidth selection . . . 15
- 2.5 Simulation . . . 16
- 2.6 Post-processing extreme precipitation . . . 18
- 2.7 Discussion . . . 23
- 2.A Proofs . . . 23
  - 2.A.1 Proof of Theorem 1 . . . 23
  - 2.A.2 Proof of Theorem 2 . . . 25
  - 2.A.3 Proof of Theorem 4 . . . 28
  - 2.A.4 Proof of Theorem 3 . . . 30

3 Gradient Boosting Extremes 31
- 3.1 Introduction . . . 31
- 3.2 Extreme quantile regression with gradient boosting . . . 33
  - 3.2.1 Background on extreme quantile estimation . . . 33
  - 3.2.2 Setup for extreme quantile regression . . . 34
  - 3.2.3 GPD modeling with gradient boosting . . . 35
  - 3.2.4 Extreme quantile regression . . . 37
- 3.3 Parameter tuning and interpretation . . . 38
  - 3.3.1 Parameter tuning . . . 38
  - 3.3.2 Tools for model interpretation . . . 40

3.4 Simulation studies . . . 42
3.4.1 Tuning parameters and cross validation . . . 43
3.4.2 Comparison with different methods . . . 44
3.4.3 Diagnostic plots . . . 44
3.5 Application to precipitation forecast . . . 47
3.5.1 Precipitation Data . . . 47
3.5.2 Model fitting . . . 48
3.5.3 Results . . . 49
3.6 Summary and discussion . . . 50
3.A Likelihood derivatives . . . 51

4 Interpretable random forest models through forward variable selection 53
4.1 Introduction . . . 53
4.2 Forward selection . . . 55
4.3 Forward selection using random forests . . . 58
4.3.1 Intermezzo: Random Forests . . . 58
4.3.2 Estimation of predictive loss . . . 59
4.3.3 One step forward . . . 60
4.3.4 Stopping on time . . . 61
4.4 Comparison based on simulation . . . 63
4.5 Post-Processing maximum temperature forecasts . . . 66
4.6 Summary and discussion . . . 71
4.A CRPS calculations . . . 72
4.B Calibration of forecasts for lead time 60 and station De Bilt . . . 72

5 Conclusion 75

Acknowledgements 79
References . . . 81

Curriculum Vitæ 87

List of Publications 89

SUMMARY

This thesis develops new statistical methodologies that are applied in the context of statistical post-processing. Statistical post-processing is the practice that improves the physics based weather forecast, coming from numerical prediction (NWP) models, by investigating the relationship between historical forecasts and observations. Through this process the inherent biases in the NWP forecast can be calibrated and corrected. The uncertainty within these NWP model forecast can also be quantified.

Within this thesis we address two different problems in the domain of statistical post-processing. First we consider post-processing for extreme events. Extreme weather events are by definition rare events and there forecasts are subject to high uncertainties. We develop, in the first two chapters, methods to quantify these uncertainties by predicting the tail of the forecast distribution. Secondly, we explore the problem of variable selection in post-processing. The large number of potential predictors coming from NWP models is extremely large and (strongly) correlated with each other. Reducing the number of predictors adds to the interpretability of the statistical model. Within our proposed method we take into account the high uncertainties within weather forecasts by specifically selecting features that help improve the uncertainty quantification.

In Chapter 2 we propose a method to estimate the high conditional quantiles. The method fits an intermediate quantile using local linear quantile regression. The exceedances of this intermediate quantile are then used to extrapolate to the high conditional quantiles by fitting a generalized pareto distribution to them using a adjusted Hill estimator. The post-processing technique is applied to a precipitation dataset for the warm half of the year and is shown to improve predictive performance compared to the upper ensemble member of the ECMWF ensemble.

In Chapter 3 we propose a different method for estimating the high conditional quantiles. This method fits an intermediate quantile using quantile regression forests. Then we use gradient boosting to fit a generalized Pareto distribution to the exceedances of this intermediate quantile. The gradient boosting approach fits two sequences of trees for both the scale and the shape parameter of the generalized Pareto distribution allowing them both the depend on predictors. The tree based methods for both the threshold and the extrapolation step allow us to include several predictors within the model without it quickly over fitting. The entire method is the first that estimates both the shape and scale parameter of the generalized Pareto distribution within a high dimensional predictor space.

The final Chapter 4 is concerned with variables selection. We propose a stepwise method for selecting predictors that improve the accuracy as well as the uncertainty quantification of the weather forecast. The selection procedure is build on quantile random forests which allow interaction between the predictors to be represented in the model. For the stepwise methodology we develop an early stopping strategy that allows

the algorithm to stop as soon as predictive performance is not further improved. The method is shown to behave well, even in situations where predictors are correlated.

SAMENVATTING

Dit proefschrift ontwikkelt nieuwe statistische methoden die worden toegepast in de context van statistische nabewerking. Met statistische nabewerking worden op fysica gebaseerde weersvoorspelling verbeterd, afkomstig van numerieke voorspellingsmodellen (NWP), door de relatie tussen historische voorspellingen en waarnemingen te onderzoeken en gebruiken. Door dit proces kunnen de inherente onzuiverheden in de NWP voorspelling worden gekalibreerd en gecorrigeerd. Ook de onzekerheid binnen deze NWP-model voorspelling kan worden gekwantificeerd.

In dit proefschrift behandelen we twee verschillende problemen op het gebied van statistische nabewerking. Ten eerste beschouwen we nabewerking voor extreme gebeurtenissen. Extreme weersomstandigheden zijn per definitie zeldzame gebeurtenissen en hun voorspellingen zijn in hoge mate onzeker. In de eerste twee hoofdstukken ontwikkelen we methoden om deze onzekerheden te kwantificeren door de staart van de voorspellingsverdeling te voorspellen. Ten tweede onderzoeken we het probleem van variabeleselectie in nabewerking. Het aantal potentiële covariaten afkomstig van NWP-modellen is extreem groot en deze covariaten zijn (sterk) met elkaar gecorreleerd. Het verminderen van het aantal covariaten draagt bij aan de interpreteerbaarheid van het statistische model. Binnen onze voorgestelde methode houden we rekening met de hoge onzekerheden binnen weersvoorspellingen door specifiek covariaten te selecteren die de onzekerheidskwantificering helpen verbeteren.

In Hoofdstuk 2 stellen we een methode voor om de extreme conditionele kwantielen te schatten. De methode schat eerst een hoog kwantiel met behulp van lokale lineaire kwantielregressie. De overschrijdingen van dit hoge kwantiel worden vervolgens gebruikt om te extrapoleren naar extreme contionele kwantielen. De parameter van de staartverdeling, een generaliseerde Pareto verdeling, schatten we met een aangepaste Hill-schatter. De nabewerkingstechniek wordt toegepast op een neerslag dataset voor de warme helft van het jaar. We tonen aan dat de methode de voorspellende kwaliteit verbetert in vergelijking met het bovenste ensemble member van het ECMWF-ensemble, een globaal probabilistisch weersvoorspellings model van de (European Centre for Medium range Weather Forecasts).

In Hoofdstuk 3 stellen we een andere methode voor om de extreme conditionele kwantielen te schatten. Deze methode schat eerst een hoog kwantiel met behulp van kwantiel random forests. De overschrijdingen van het hoge kwantiel worden dan tevens gebruikt om de parameters van een gegeneraliseerde Pareto verdeling the schatten. Hiervoor gebruiken we een gradient boosting methode. De gradient boosting methode schat aparte regressiebomen voor zowel de schaal- als de vorm parameter van de gegeneraliseerde Pareto-verdeling. Hierdoor kunnen beide parameters afhangen van covariaten. Doordat beide stappen in deze methode gebaseerd zijn op regressiebomen is het mogelijk om een hoog aantal covariaten mee te nemen zonder dat het model overfit. De

hele methode is de eerste die zowel de vorm- als de schaalparameter van de gegeneraliseerde Pareto-verdeling schat binnen een hoogdimensionale voorspellerruimte.

In het laatste hoofdstuk 4 presenteren we een methode voor het selecteren van variabelen. We stellen een stapsgewijze methode voor om variabelen te selecteren die zowel de nauwkeurigheid als de onzekerheidskwantificering van de weersvoorspelling verbetert. De selectieprocedure is gebaseerd op kwantiel random forests, waardoor interactie effecten ook worden meegenomen in het model. Voor de stapsgewijze methode ontwikkelen we een techniek om het algoritme vroeg te stoppen, op het moment dat de voorspellende prestaties niet verder worden verbeterd. We tonen aan dat de methode zich goed gedraagt, zelfs in situaties waarin variabelen gecorreleerd zijn.

1

INTRODUCTION

Extreme weather events can have catastrophic impacts on both the economy and society. In warming climates, heat waves are expected to be warmer and more frequent. In The Netherlands, for example, we observed record breaking temperatures exceeding 40 degrees Celcius in July of 2019. Extremes as consequence of changes to the climate are not only visible in the occurrence of more heat waves. Also extreme precipitation events are expected to become more frequent. A good example of such extreme precipitation happened over Germany on July 13 and 14, 2021. This caused massive flooding events in Germany, Belgium and The Netherlands as water discharge made its way through rivers towards the sea.

As occurrence and magnitude of these extreme weather events start to change, there is a strong need to be better able to understand and predict these events. This thesis addresses problems in statistical post-processing for (extreme) weather forecasting, in particular, post-processing for extreme events and variable selection for post-processing.

In this chapter we will provide the necessary background in statistical post-processing of weather forecasts. We will start by describing the definition and importance of statistical post-processing for weather forecast. The subsequent sections will be devoted to further discussing the two challenging topics, forecast for extreme events and variable selection. The chapter ends with an overview of the structure of the whole thesis.

1.1. STATISTICAL POST-PROCESSING FOR WEATHER FORECAST

The idea behind statistical post-processing is simple. Given a model that generates weather forecasts, statistical post-processing is the process of improving this weather forecast by correcting the systematic biases therein. By comparing the historical weather forecasts and the corresponding observations, systematic biases can be estimated. In future forecasts, the correction of these systematic biases can be applied as an additional second step following the forecast generation of the weather model.

In order to understand why such a two step approach is necessary, it is helpful to understand the mechanism of a weather forecasting model. Below we describe how weather forecasts are created and where potential biases might appear.

1.1.1. NUMERICAL WEATHER PREDICTION MODELS

From here on we refer to numerical weather prediction models as weather prediction models or NWP models. An NWP model generates forecasts based on non-linear differential equations that physically describe the flow and the heat transfer within the atmosphere. The solution to these differential equations is approximated by the NWP model on a four dimensional grid, i.e., latitude, longitude, height and time. Within the NWP model there are two main sources that contribute to the uncertainty of the forecast.

The first source is the estimated initial state of the atmosphere, commonly referred to as the analysis. The analysis is obtained by calibrating the most recent forecast of the NWP-model with an extensive world wide weather observations network, which includes station observations, weather balloons, radar and satellites. An accurate analysis is essential for accurate weather forecasts. As processes in the atmosphere are highly non-linear, the atmosphere is generally described as a chaotic system. This is commonly illustrated with the butterfly effect, where a small action (a butterfly flapping its wings) can cause enormous ripple effects (for example a hurricane on the other side of the ocean). What this essentially means, in the context of weather prediction, is that slight changes in the analysis can lead to completely different weather forecasts for a few days later.

Another source of uncertainty appears in the forecast for parameters such as precipitation, wind gust and cloud cover. These parameters cannot be explicitly computed from the solutions to the physical model as they depend on additional physical processes that happen on a sub grid scale. Instead, they are inferred using formulas known as parametrizations. The parametrizations are generally oversimplified and are susceptible to uncertainty. A potential reduction of the uncertainty can be achieved by computing the entire NWP model on a higher resolution grid. This comes at a high computational cost and therefore is infeasible to compute for the entire globe. Instead, these higher resolution models are computed in smaller regions and use the lower resolution global model as boundary conditions.

1.1.2. ENSEMBLE PREDICTION SYSTEMS

Given a weather prediction model, the uncertainty that the model contributes to the forecasts is quantified by the Ensemble Prediction System (EPS). An EPS contains an ensemble of forecasts using the same NWP forecasting system, but where the input of the analysis and/or the parametrizations are "randomized" in the directions where the analysis and/or parametrizations have the highest uncertainty. These are known as ensemble forecasts. Simply put, an EPS attempts to capture the uncertainty in the forecasts that comes from the propagated uncertainty within the analysis and the parametrizations.

Note that while such a quantification is highly desirable, an EPS clearly requires much more computation compared to a single NWP forecast. In practice, an EPS model generally is run on a coarser grid compared to the situtation where only a single deterministic NWP forecast is required. This means that although an EPS is a very good initial approach in order to quantify the uncertainty within a forecast, there are different types of uncertainty that are not accounted for in an EPS.

1.1.3. STATISTICAL POST PROCESSING

Let us first consider two classical examples of post-processing. Suppose that we are interested in the forecast for the temperature at 2 metres above the surface in Delft, to be predicted 36 hours ahead of time. Given corresponding observations, either a single NWP forecast or an ensemble of NWP forecasts are available to us. Our goal is to improve the forecast based on the historical forecasts and their corresponding observations. The standard and widely accepted method for post-processing NWP deterministic forecasts is called Model Output Statistics (MOS). The adaptation to EPS forecast is called Ensemble Model Output Statistics (EMOS).

EXAMPLE: MODEL OUTPUT STATISTICS

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote pairs of deterministic forecasts and corresponding observations for temperatures, where $X_i, Y_i \in \mathbf{R}$ for $i = 1, \ldots, n$.

Model Output Statistics relates the observed temperature with the forecast temperature via a linear relation, [36], [29].

$$Y_i = \alpha + \beta X_i + \epsilon_i \text{ for } i = 1, \ldots, n. \tag{1.1}$$

Here α can be seen as correcting for systematic biases that appear within the forecast. The parameters are then estimated by least squares estimation.

EXAMPLE : ENSEMBLE MODEL OUTPUT STATISTICS

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote pairs of ensemble forecasts and observations for temperature, where $\mathbf{X}_i = (X_{i1}, \ldots X_{iM})^T$ and $Y_i, X_{ij} \in \mathbb{R}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, M$. Define the ensemble mean by $\overline{\mathbf{X}}_i = \frac{1}{M}\sum_{j=1}^{M} X_{ij}$ and the ensemble variance by $\sigma^2(\mathbf{X}_i) = \frac{1}{M}\sum_{j=1}^{M}(X_{ij} - \mathbf{X}_i)^2$.

The Ensemble Model Output Statistics, [31], provide a forecast distribution whose moments are related to those of the ensemble forecasts. For example, we may forecast Y_i to follow a normal distribution,

$$Y_i | X = \mathbf{X}_i \sim N(\hat{\mu}_i, \hat{\sigma_i}^2) \tag{1.2}$$

where $\hat{\mu}_i = \alpha_1 + \beta_1 \overline{\mathbf{X}_i}$ and $\hat{\sigma}_i^2 = \alpha_2 + \beta_2 \sigma^2(\mathbf{X}_i)$. The parameters α_1 and β_1 have a similar interpretation as for the Model Output Statistics. The parameter α_2 and β_2 correct systematic over- and under-dispersion errors. The parameters can be estimated by maximum likelihood, assuming observations are conditionally independent given the forecast.

The MOS and EMOS are two standards for post-processing that are widely applied. Their extensions are much explored in the literature. We will now discuss a few of these extensions to set the stage for the research done in this thesis. The possible extensions include: relaxing the normality assumption, expanding the predictor space to contain more than just the forecast, and relaxing the strict linearity assumption to improve the predictive performance.

POST-PROCESSING BEYOND THE NORMAL DISTRIBUTION

The normal distribution has been shown to work well for post-processing mean temperature, but not for many other weather phenomena such as wind speed. The initially proposed distribution for wind speeds was a normal distribution truncated at zero [57]. This works only for low wind speed data since it was observed that high wind speeds conditional on the forecast do not follow a normal distribution. Therefore [42] proposes to post-process low wind speeds with a truncated normal and high wind speeds with an extreme value distribution. In [3] a log normal distribution was proposed for the higher wind speed events.

Other weather variables, such as precipitation, are better modelled with a mixutre discrete-continuous distribution to model dry and wet days accurately. As a result, these events are generally modelled using a continuous distribution with an additional point mass at zero. The choices for such continuous distributions include gamma distributions [64] and generalized extreme value distributions [51]. More recently an extended Pareto distribution was proposed [55], which behaves as a gamma distribution for low values and as a Pareto distribution for high values.

MACHINE LEARNING WITH POST-PROCESSING

All above models focus on proposing parametric models for both the distribution of the forecast and the linear dependence between forecast and observations. However, these parametric assumptions are often too rigid.

For example, separate EPS members can forecast the position of large scale atmospheric patterns in a different way. As a result, some EPS members will forecast a storm at location A, where the others forecast the storm to be at location B. The ensemble forecast at location A then forms a bimodal distribution with mass for both high and low wind speeds. In these cases, unimodal distributions are not sufficient any more and a more flexible way of quantifying the uncertainty is necessary.

In post-processing of the mean temperature it can be shown that a simple linear regression performs rather well. The reason is simple, mean temperature forecast is rather easy to forecast and therefore the relation between observations and forecasts is approximately linear. But in many cases this is not necessarily the case. Especially when including additional predictors, on top of the ensemble forecast, to provide the statistical model with more information about the state of the atmosphere.

In [64] additional covariates of atmospheric instability are used to improve precipitation forecasts for extreme precipitation amounts. Though it can be observed by means of scatter plots that the relation is not linear at all.

The two above examples motivate the use of machine learning methods to relax the parametric assumptions made in MOS and EMOS. An additional motivation of choosing machine learning approaches over classical non-parametric statistical methods is their stability in the presence of high dimensional predictor spaces.

For these reasons, recent expansions of the EMOS methodology have focused on combining post-processing with machine learning. An example of a method that requires little tuning and has strong predictive power is Random Forest [56], [55], [64]. Also approaches using neural networks are gaining traction [48], [58].

1.2. CONTRIBUTION OF THIS THESIS

In this thesis we focus on expanding machine learning methodology for post-processing in two directions: post-processing with a special focus on extreme events, and variable selection in case of large predictor spaces.

1.2.1. FORECASTING EXTREME EVENTS

Extreme weather events, such as extreme precipitation and extreme wind storms, have high impacts on agriculture, infrastructure and our daily lives. The heat waves of 2018 and 2019 in the Netherlands, for example, caused severe draught with large impacts on water management and agriculture. The heavy precipitation of July 2021 across Europe, as mentioned in the beginning of this introduction, caused severe flooding in the south east of The Netherlands.

Therefore it is of particular interest to forecast these events accurately. A way of improving forecast quality can be achieved through statistical post-processing. In practice, however, forecast of extremes poses additional challenges. First, the estimation efficiency of a statistical model increases with the number of data points available. Since the extreme events are infrequent, the performance of standard statistical post-processing is often unsatisfactory for extreme forecasts. Secondly, extreme events are subject to much higher intrinsic uncertainty. For example, extreme precipitation is generally localized and therefore especially hard to forecast.

Few models in the current statistical post-processing literature are able to accommodate the forecasting of extreme events. This is because most statistical models used in statistical post-processing deal explicitly with the tail of the distribution. Additionally, an ensemble forecast is constructed from a limited number of members and therefore has a harder time capturing low probability events. EMOS methods, for example, assume a parametric model for the entire predictive distribution, which often has similar tail characteristics that do not allow flexible modelling of the extremes. Finally, non-parametric methods and machine learning methods focus on fitting the bulk of the data therefore extreme events are not modelled effectively.

With these challenges in mind, we are motivated to move beyond the current machine learning and non-parametric statistical post-processing methodology and incorporate extreme value theory to improve the forecasts for the extremes. Extreme value theory is the field of statistics that concerns modelling large values, i.e., the tail of the distribution, especially beyond the range of observations.

Classical extreme value theory starts by inspecting the sample maximum M_n, i.e. the maximum number within a sample $\{X_i\}_{i=1}^n$. The foundational theorem within extreme value theory shows that the suitably normalized sample maximum $(M_n - a_n)/b_n$ converges to a limiting distribution called the generalized extreme value distribution, which is parametrized by a parameter γ known as the extreme value index. In order to estimate the γ, the block maxima method is applied by taking the maximum of coherent blocks. In environmental science, such a block generally consists of a year, which means the γ can be estimated by the yearly maxima.

Another possibility for modelling extremes is to look at the exceedances above a high threshold. The peaks-over-threshold theorem shows that these exceedances converge to

a different limiting distribution, called the generalized Pareto distribution,

$$G_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{x\gamma}{\sigma}\right)^{-1/\gamma}. \tag{1.3}$$

Similar to the generalized extreme value distribution, the generalized Pareto distribution is parametrized by the extreme value index γ and an additional scale parameter σ. Estimation is done by choosing a high enough threshold and fitting the parameters to the exceedances of this threshold.

The extreme value index γ plays an important role in extreme value theory and depending on its value, tail behaviour can be categorized in three different categories. A negative γ indicates short tail behaviour, i.e. the probability of observing anything beyond a finite point $x^* < \infty$ is zero. A zero γ indicates that the tails have exponential decaying behaviour. This means that the probability of an observation above a threshold decays exponentially in the threshold. Finally, a positive γ indicates the power law decay of the probability of exceedance.

Within post-processing and this thesis, it is natural to define extremes as the exceedances above a high threshold. In the rest of this thesis we will therefore use the generalized Pareto distribution to model extremal behaviour.

The methods that we develop fall within the area of extreme quantile regression, where extreme quantiles are estimated as a function of covariates. These methodologies have a two step estimation approach. First, an intermediate quantile is estimated using the existing data points. Second, the generalized Pareto distribution is used to extrapolate to the extreme tail. These two steps enable us to extract enough information from the data, and enough flexibility is given to modelling the tail.

In this thesis, we propose two methods for extreme quantile regression to be used in statistical post-processing. In Chapter 2, we propose a simple model that combines the estimation of threshold estimated by local linear quantile regression with an estimator for the extreme value index to estimate the extreme quantiles. We prove the uniform consistency of the non-parametric estimated threshold. This is a necessary condition such that the estimation error of the threshold does not influence the estimated extremal behaviour asymptotically.

However, local linear quantile regression does not work in high dimensional covariate spaces due to the curse of dimensionality. Therefore in Chapter 3, we extend the methodology of the common shaped tail estimator. We propose the GBEX method, gradient boosting for extremes. In GBEX we incorporate machine learning methodologies that rely on decision trees. These are constructed by recursively making splitting the data on the covariates. This means that there is inherent variable selection in these methods, which make them applicable for high dimensional covariate spaces.

Within GBEX, quantile random forests are used to fit the intermediate threshold. The reason for this is that it is a very robust method, which needs very little tuning. On the exceedances of this threshold we construct a gradient boosting procedure connecting two parallel sequences of trees to estimate both parameters γ and σ of the generalized Pareto distribution as functions on a high dimensional covariate space. In Chapter 3 we show in an example how our estimator can be used for post-processing extreme precipitation. We use the GBEX method to post-process extreme precipitation, taking into account both seasonality and spatial dependence.

1.2.2. VARIABLE SELECTION FOR STATISTICAL POST PROCESSING

For the statistical post-processing of the forecast of a weather event, the most straightforward choice of potential predictors comes from the other variables that are being forecast by the same NWP model, the number of which can be extremely large. This provides a large pool of predictors to choose from for statistical post-processing. In order to make accurate decisions on the weather forecasts and its uncertainty, they need to be interpretable. This make it crucial to address the variable selection problem here as the predictive powers of each predictor differs in various circumstances. For example, when forecasting precipitation, forecasts with lead times of more than a few days generally have no predictive skill any more. Temperature forecasts, on the other hand, keep having predictive skill for more than a week.

In general, the challenges in variable selection within statistical post-processing are quite specific and can be divided in three distinct aspects. First, there exists high correlation between predictors. For example a forecast for the maximum temperature is highly correlated with the forecast for minimum temperature. This does not mean that one should always be selected in favour of the other. Second, the model should remain interpretable. This means it should be clear to understand how each predictor occurs in the final statistical post-processing model. Finally, due to high uncertainties it is important to have accurate uncertainty quantification. Therefore the predictors should be selected in order to improve the entire forecast distribution instead of the expectation of the forecast distribution as in the mean regression model.

In Chapter 4, we introduce a stepwise variable selection process for random forests that deals with all three challenges. By incrementally computing the predictive performance of the model, a predictor is only added if the predictive performance is strictly increasing. The predictive performance is measured by the continuous ranked probability score, a score to measure the accuracy of a predictive distribution.

In an extensive simulation study, it is shown that in the presence of high correlations between predictor variables, the model is able to distinguish the variables from the simulated model much more reliably than competing methods, while at the same time not selecting additional predictors. In an application of post-processing the maximum temperature forecast, it can be observed that the proposed method selects much fewer features than competing methods, while keeping the same predictive power. Additionally, the method is selects predictors more consistently within a cross validation set-up

1.2.3. OUTLINE

The rest of the thesis is organized as follows. Chapter 2 introduces the extreme quantile regression with the common shape tail estimator. In Chapter 3 this methodology is extended using random forest and gradient boosting. The variable selection methodology for statistical post-processing is discussed in Chapter 4. This thesis concludes with a discussion in Chapter 5.

2

IMPROVING EXTREME PRECIPITATION FORECAST USING EXTREME QUANTILE REGRESSION

Aiming to estimate extreme precipitation forecast quantiles, we propose a nonparametric regression model that features a constant extreme value index. Using local linear quantile regression and an extrapolation technique from extreme value theory, we develop an estimator for conditional quantiles corresponding to extreme high probability levels. We establish uniform consistency and asymptotic normality of the estimators. In a simulation study, we examine the performance of our estimator on finite samples in comparison with a method assuming linear quantiles. On a precipitation data set in the Netherlands, these estimators have greater predictive skill compared to the upper member of ensemble forecasts provided by a numerical weather prediction model.

2.1. INTRODUCTION

Extreme precipitation events can cause large economic losses, when large amounts of water cannot be properly drained. For example, water boards in the Netherlands, responsible for water management, need to take preventive action in the case of large amounts of precipitation to prevent flooding. Accurate predictions are therefore vital for taking preventive measures such as pumping the water out of the system.

Weather forecasting relies on deterministic forecasts obtained by numerical weather prediction (NWP) models [35]. These models are based on non-linear differential equations from physics describing the flow in the atmosphere. Starting from an initial condition of the atmosphere and using so-called physical parametrizations to account for unresolved physical processes, the NWP models are used to forecast precipitation, among other weather quantities.

Parts of this chapter have been published in Extremes **22**, 599 (2019).

The uncertainty in these types of forecasts is attributed to uncertainty in the initial condition and in the physical parametrizations in the model itself. An ensemble prediction system quantifies the uncertainty due to these two factors by applying small perturbations to the original quantities and running the NWP model multiple times subsequently. An ensemble forecast is to be viewed as a sample from the distribution of the predicted variable, where uncertainties in initial condition and model parametrizations are taken into account. Therefore, it is natural to consider the empirical distribution function of the ensemble forecast as an estimator of the distribution of the predicted variable, in this chapter precipitation.

While the NWP ensemble prediction systems are rather skilful in forecasting precipitation for relatively short lead times, skill quickly decreases as lead time increases. Using upper ensemble members for forecasting extreme precipitation appears to be most challenging, due to the large spatial and temporal uncertainties of precipitation forecasts. Most methods that have been proposed to post-process forecasts are instead focussed on the bulk of the conditional distribution, see [65].

For the upper ensemble members there are two serious problems. First, the upper ensemble members tend to be not well calibrated, i.e. not reliable [6], especially for large amounts of precipitation, this is shown in [5]. Second, the highest probability level of the extreme precipitation forecast is limited by the number of ensemble members, which is typically not large due to computational costs. In the ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF), which we consider in our case study, the system generates 51 ensemble members. Thus, the largest probability level is given by $\frac{51}{52}$.

In this chapter, we aim to develop a post-processing approach for predicting extreme precipitation quantiles. More precisely, we focus on the problem of estimating the tail of the conditional distribution $F_{Y|X}$, with X a precipitation forecast by the NWP model and Y the observed precipitation. We are interested in the function $x \mapsto Q_{Y|X}(\tau|x)$ for τ close to one, where $Q_{Y|X}$ denotes the conditional quantile function.

Several estimators have already been proposed to estimate extreme conditional quantiles. All these estimators have a similar structure consisting of two steps. First, the quantile function $Q_{Y|X}$ is estimated for moderately high probability levels τ. In the second step, these estimated quantiles are used to extrapolate to obtain estimators of extreme conditional quantiles.

For the first step, general quantile estimation techniques are used. Examples are linear quantile regression in [63] and [62], a local polynomial approximation to the quantile function [4], a k-nearest neightbour approach in [25] and inverse of empirical conditional distribution functions smoothed in the covariates in [14], and [13]. For the second step two 'types' of approaches can be distinguished. First, a local approach, where an extreme quantile estimator is applied to a sequence of estimated quantiles for moderately high probability levels attained from the first step. This method is used in [63], [62], [14], [13], [25], [32] and [27]. The second type, where the exceedances above a threshold estimated in the first step are used to fit a generalized Pareto distribution, was introduced in [16]. An application of the result of [16] to precipitation data is discussed in [5], where a generalized Pareto distribution is fitted to the exceedances above an estimated linear quantile. They showed skilful short-range forecasts of extreme quantiles.

Most methods allow for a varying extreme value index depending on the covariates. The estimators of extreme value indices in such models are generally subject to high variability. In the context of weather forecasting, this may lead to inconsistent forecasts over the covariates. After carefully considering the trade-off between the generality of the model and the efficiency of the estimation, we propose an additive model with a constant extreme value index for all covariates cf. (2.1) and (2.2). Moreover, we assume that the extreme value index is positive. This assumption is supported by the result of our empirical study on summer rainfall in the Netherlands as well as the existing literature on precipitation data including [12], [9] and [24]. Apart from this, our model assumes that the conditional quantile of Y is a non-parametric function of the covariate, thus no parametric structure is required. In our two step procedure, we first estimate a non-stationary threshold, namely the non-parametric quantile function by local linear quantile regression and then extrapolate to extreme quantiles based on the exceedances of this threshold.

The scientific contribution of this chapter is fourfold. First, we propose a model that achieves a good balance between generality and estimation efficiency and it fits the feature of post-processing data sets. Second, we derive asymptotic properties of the estimators, by first showing uniform consistency of local linear quantile regression, using a uniform Bahadur representation for the quantile estimator. Moreover, we establish asymptotic normality of the estimators of the extreme value index as well as the extreme conditional quantiles. Third, we address the issues such as selection of the bandwidth and tuning parameters, which is highly relevant from the application point of view. Fourth, our procedure yields skilful prediction outperforming the upper ensemble member and showing similar skill to the linear estimator [63] based on cross-validation. Besides, our procedure can extrapolate to an extreme probability level that goes beyond the empirical quantile associated with the upper ensemble member.

The outline of the chapter is as follows: Section 2.2 we present our proposed model and develop the estimating procedures. The asymptotic properties of the estimator are studied in Section 2.3. In Section 2.4 we propose a data driven approach for bandwidth selection. We show with a detailed simulation study in Section 2.5 the finite sample performance of our estimator and compare it with an existing method. In Section 2.6 we apply our estimator to a dataset of precipitation observations and ensemble forecasts in the Netherlands. Finally, in Section 2.7 we discuss future research directions. The proofs of the theoretical results are provided in the appendix.

2.2. MODEL AND ESTIMATION

We aim to estimate the conditional tail quantiles of Y given X, namely $Q_{Y|X}(\tau|\cdot)$ for τ close to one. To this end, we assume that there exists a $\tau_c \in (0,1)$ such that

$$Q_{Y|X}(\tau|x) = r(x) + Q_\epsilon(\tau) \text{ if } \tau \geq \tau_c, \tag{2.1}$$

where r is a smooth continuous function and Q_ϵ denotes the quantile function of an error variable ϵ, which is independent of X. In order to make the model identifiable, it is assumed that $Q_\epsilon(\tau_c) = 0$. As a result, $Q_{Y|X}(\tau_c|x) = r(x)$. Moreover, we assume that the

distribution of ϵ has a heavy right tail, that is there exists $\gamma > 0$ such that,

$$\lim_{t\to\infty} \frac{Q_\epsilon\left(1-\frac{1}{tx}\right)}{Q_\epsilon\left(1-\frac{1}{t}\right)} = x^\gamma \ , \ x > 0, \tag{2.2}$$

where γ is the extreme value index of ϵ. Note that (2.2) implies that the conditional distribution of Y given that $X = x$ also has a heavy right tail with the same extreme value index γ.

It is important to note that this additive structure is only assumed for probability levels τ exceeding τ_c, which allows us to model the tail of the conditional distribution without assuming structure for $\tau < \tau_c$. On one hand, the quantile curve $x \to Q_{Y|X}(\tau|x)$ for any $\tau \geq \tau_c$ has the same shape as r. On the other hand, the distance between the two quantile curves, that is $Q_{Y|X}(\tau_1|x) - Q_{Y|X}(\tau_2|x)$ for any $\tau_1 > \tau_2 \geq \tau_c$, is determined by Q_ϵ only and thus does not depend on x. We will refer to our model as the Common Shape Tail (CST) model.

We remark that various types of additive structures have been proposed in recent studies on modeling extremes with covariates. In [63], a linear structure is assumed for r, where two scenarios are considered: the slope of the linear function is a nonparametric function of τ or it is constant. The latter scenario is a special case of our model. In [62], a linear structure is assumed for the conditional quantile function after the power transformation. In both papers, r is estimated by linear quantile regression. In [44], a nonparametric location-scale representation is assumed and local linear mean regression is used to estimate the conditional quantile called α-CVaR in that paper, where the existence of the fourth moment of the error variable is required. This requirement implies an upper bound on the extreme value index: $\gamma < \frac{1}{4}$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote i.i.d. paired observations satisfying (2.1). Based on this random sample, we construct a two step estimation procedure for $Q_{Y|X}(\tau_n|\cdot)$, where for asymptotics, $\tau_n \to 1$ as $n \to \infty$. We shall estimate r and $Q_\epsilon(\tau_n)$ respectively in each of the two steps.

First, for the estimation of r we choose to follow the local linear quantile regression approach studied in [67]. An obvious advantage of the quantile regression approach is that it does not impose a constraint on the moments of the conditional distribution. Let $h = h_n$ denote the bandwidth. In a window of size $2h$ around a fixed point x, we approximate the function linearly:

$$r(\tilde{x}) \approx r(x) + r'(x)(\tilde{x} - x) =: \alpha + \beta(\tilde{x} - x), \qquad \tilde{x} \in [x - h, x + h].$$

The function r and its derivative are estimated by the solution of the following minimization problem:

$$(\hat{r}_n(x), \hat{r}'_n(x)) = \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{n} \rho_{\tau_c}(Y_i - \alpha - \beta(X_i - x)) K\left(\frac{X_i - x}{h}\right), \tag{2.3}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile check function, cf. [37] and K a symmetric probability density function with $[-1, 1]$ as support.

Second, for the estimation of $Q_\epsilon(\tau_n)$, we consider the residuals defined by $e_i = Y_i - \hat{r}_n(X_i)$, $i = 1, \ldots, n$. Using the representation of $Y_i = Q_{Y|X}(U_i|X_i)$, with $\{U_i, i = 1, \ldots, n\}$

i.i.d. uniform random variables, and the model assumption (2.1), the residuals permit a more practical expression as below.

$$e_i = \begin{cases} Q_\epsilon(U_i) + (r(X_i) - \hat{r}(X_i)) & \text{if } U_i \geq \tau_c \\ Q_{Y|X}(U_i|X_i) - \hat{r}(X_i) & \text{otherwise .} \end{cases} \tag{2.4}$$

Denote the order statistics of the residuals by $e_{1,n} \leq \ldots \leq e_{n,n}$. Let k_n be an intermediate sequence depending on n such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then a Hill estimator of the extreme value index is given by

$$\hat{\gamma}_n = \frac{1}{k_n} \sum_{i=1}^{k_n} \log \frac{e_{n-i+1,n}}{e_{n-k_n,n}}.$$

The intuitive argument behind this estimator is that $\{e_{n-i,n}, i = 0, \ldots, k_n\}$ are asymptotically equivalent to the upper order statistics of a random sample from the distribution of ϵ, i.e. for some $\delta > 0$,

$$\max_{i=0,\ldots,k_n} |e_{n-i,n} - Q_\epsilon(U_{n-i,n})| = o_p(n^{-\delta});$$

see the proof of Theorem 2 in the Appendix. For the same reason, we use the well known Weissman estimator of $Q_\epsilon(\tau_n)$ based on the upper residuals:

$$\hat{Q}_\epsilon(\tau_n) = e_{n-k_n,n} \left(\frac{k_n}{n(1-\tau_n)} \right)^{\hat{\gamma}_n}. \tag{2.5}$$

Combining the estimator of $r(x)$ given by (2.3) and the estimator of $Q_\epsilon(\tau_n)$ given by (2.5), we obtain the estimator of the conditional tail quantile:

$$\hat{Q}_{Y|X}(\tau_n|x) = \hat{r}(x) + \hat{Q}_\epsilon(\tau_n). \tag{2.6}$$

By construction, this estimator of the conditional tail quantile is continuous in x. We shall refer to our estimator as CST-estimator.

2.3. ASYMPTOTIC PROPERTIES

In this section, we present the asymptotic properties of the estimators obtained in Section 2.2. We begin with uniform consistency of $\hat{r}_n$ in (2.3). We first state the assumptions with respect to our model (2.1). Let g denote the density of X, $f_{Y|X}(\cdot|x)$ denote the conditional density of Y given $X = x$ and c denote an arbitrary finite constant.

A1 The support of g is given by $[a, b]$ and $\sup_{x\in[a,b]} |g'(x)| \leq c$.

A2 The third derivative of r is bounded, i.e. $\sup_{x\in[a,b]} |r'''(x)| \leq c$.

A3 The function $x \to f_{Y|X}(r(x)|x)$ is Lipschitz continuous and $f_{Y|X}(r(x)|x) > 0$ for all $x \in [a, b]$.

Theorem 1. *Let $\hat{r}_n$ be the estimator defined in (2.3). Choose K a symmetric Lipschitz continuous probability density function supported on $[-1,1]$ and $h_n = O(n^{-\delta_h})$, with $\delta_h \in \left(\frac{1}{5}, \frac{1}{2}\right)$. Under Assumptions* A1-A3*, there exists a $\delta \in (0, \frac{1}{2} - \delta_h)$ such that as $n \to \infty$,*

$$\sup_{x\in[a,b]} |\hat{r}_n(x) - r(x)| = o_p(n^{-\delta}).$$

This theorem quantifies the direct estimation error made in the first step of our procedure. Note that the "error" made in the first step is transmitted to the second step by the definition of the residuals. Thus, the uniform consistency of $\hat{r}$ is important for deriving the asymptotic property of $\hat{Q}_{Y|X}(\tau_n|\cdot)$ not only because $\hat{r}$ is a constructing part of $\hat{Q}_{Y|X}(\tau_n|\cdot)$, but it also influences the asymptotic behavior of $\hat{Q}_\epsilon(\tau_n)$.

Remark 1. *Although many studies have been devoted to the non-parametric quantile regression, to the best of our knowledge, there is no existing result on the uniform consistency for $\hat{r}_n$ for an additive model. In [39], a general uniform Bahadur representation is obtained for local polynomial estimators of M-regression for a multivariate additive model. A local linear quantile regression is one of the M-regression and thus is included in the estimators considered in that paper. Corollary 1 in [39] is our starting point for deriving the uniform consistency of $\hat{r}_n$.*

For the asymptotic normality of $\hat{\gamma}_n$, we assume that Q_ϵ satisfies the following condition, which is a second order strengthening of (2.2).

A4 There exist $\gamma > 0$, $\varrho < 0$ and an eventually positive or negative function $A(t)$ with $\lim_{t\to\infty} A(t) = 0$ such that for all $x > 0$,

$$\lim_{t\to\infty} \frac{\frac{Q_\epsilon(1-\frac{1}{xt})}{Q_\epsilon(1-\frac{1}{t})} - x^\gamma}{A(t)} = x^\gamma \frac{x^\varrho - 1}{\varrho}. \tag{2.7}$$

As a consequence, $|A(t)|$ is regularly varying with index ϱ.

Theorem 2. *Let the conditions of Theorem 1 and* A4 *be satisfied. Let $k_n \to \infty$ and $k_n/n \to 0$, $\sqrt{k_n}A(n/k_n) \to \lambda \in \mathbf{R}$ and $k_n^{\gamma+1} n^{-(\delta+\gamma)} \to 0$ as $n \to \infty$, with δ from Theorem 1. Then*

$$\sqrt{k_n}(\hat{\gamma}_n - \gamma) \xrightarrow{d} N\left(\frac{\lambda}{1-\varrho}, \gamma^2\right) \text{ as } n \to \infty.$$

Remark 2. *When deriving asymptotic properties for extreme statistics, it typically requires some regular conditions on k_n, the number of tail observations used in the estimation when the sample size is n. For the original Hill estimator, which is based on i.i.d. observations, the asymptotic normality is proved under Assumption* A4 *and $\sqrt{k_n}A(n/k_n) \to \lambda \in \mathbf{R}$. The assumption of $\varrho < 0$ is a technical condition, which is common for heavy tailed data and it allows us to choose $k_n = n^\alpha$ for $\alpha > 0$.*

The condition $\lim_{n\to\infty} k_n^{\gamma+1} n^{-(\delta+\gamma)} = 0$ is used to make sure that the upper order residuals behave similarly to the upper order statistics of a random sample from the distribution of ϵ. Suppose one chooses $k_n = n^\alpha$ for $0 < \alpha < \min\left(\frac{2\varrho}{2\varrho-1}, \frac{\delta+\gamma}{\gamma+1}\right)$, it satisfies all the conditions on k_n. So in theory, there exists a wide range of choices for a proper k_n. In practice, it is challenging to choose a k_n. In Section 2.5 we propose to use a fixed choice of k_n that worked well in several simulation studies.

The asymptotic normality of $\hat{Q}_{Y|X}(\tau_n|x)$ defined in (2.6) is now given below. To simplify notation, we denote with $p_n = 1-\tau_n$.

Theorem 3. *Let the conditions of Theorem 2 be satisfied. Assume* $np_n = o(k_n)$, $|\log(np_n)| = o(\sqrt{k_n})$ *and* $\frac{\sqrt{k_n p_n^\gamma}}{n^\delta \log\left(\frac{k_n}{np_n}\right)} \to 0$, *then as* $n \to \infty$,

$$\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)}\left(\frac{\hat{Q}_{Y|X}(\tau_n|x)}{Q_{Y|X}(\tau_n|x)} - 1\right) \xrightarrow{d} N\left(\frac{\lambda}{1-\varrho}, \gamma^2\right).$$

Remark 3. *The condition* $np_n = o(k_n)$ *guarantees that the conditional quantile is an extreme one. It gives the upper bound for* p_n. *And the condition* $|\log(np_n)| = o(\sqrt{k_n})$ *gives the lower bound on* p_n, *which limits the range of extrapolation. Clearly* $p_n = O(n^{-1})$ *satisfies both conditions. The asymptotic normality holds even for some* $p_n < \frac{1}{n}$, *which means it is beyond the range of the available data. In the weather forecast context, predicting the amount of precipitation so extreme that it never occurred during the observed period is also feasible. The assumption* $\lim_{n\to\infty} \frac{\sqrt{k_n p_n^\gamma}}{n^\delta \log\left(\frac{k_n}{np_n}\right)} = 0$ *is a technical condition we use to guarantee that the error made in the first step does not contribute to the limit distribution.*

The proofs for Theorems 1, 2 and 3 are provided in the Appendix.

2.4. BANDWIDTH SELECTION

The selection of the bandwidth is a crucial step in local linear quantile regression cf. (2.3). The bandwidth controls the trade-off between the bias and variance of the estimator. Increasing the bandwidth h decreases the variance, but tends to increase the bias due to larger approximation errors in the local linear expansion.

In [67], the authors propose to estimate the optimal bandwidth for quantile regression by rescaling the optimal bandwidth for mean regression. There is a rich literature on bandwidth selection for mean regression. However, in our setting this approach is not satisfactory because the scaling factor is difficult to estimate and it also assumes the existence of the first moment, i.e. it limits us to the case $\gamma < 1$.

Instead we adopt a bootstrap approach, similar to the one proposed in [4] to estimate the global optimal bandwidth with respect to the mean integrated squared error (MISE), i.e.,

$$h_{opt} = \arg\min_h \mathbb{E}\left[\int_a^b \left(Q_{Y|X}(\tau_c|x) - \hat{Q}^h_{Y|X}(\tau_c|x)\right)^2 \mathrm{d}x\right] =: \arg\min_h S(h),$$

where $\hat{Q}^h_{Y|X}(\tau_c|x)$ denotes the τ_c quantile estimated by (2.3) with bandwidth h.

Let B denote the number of bootstrap samples. The bootstrap samples $(X_1^j, Y_1^j), \ldots, (X_n^j, Y_n^j)$ for $j = 1, \ldots, B$ are sampled with replacement from the original n data pairs. The optimal bandwidth is estimated by minimizing the bootstrap estimator $\hat{S}(h)$ of $S(h)$, which is given by the objective function in (2.8).

$$\hat{h} = \arg\min_h \frac{1}{B}\sum_{j=1}^{B}\int_a^b \left(\hat{Q}^{h_0}_{Y|X}(\tau_c|x) - \hat{Q}^{h,j}_{Y|X}(\tau_c|x)\right)^2 \mathrm{d}x, \tag{2.8}$$

where h_0 is an initial bandwidth chosen by visual inspection and $\hat{Q}^{h,j}_{Y|X}(\tau_c|x)$ denotes the estimate of the conditional quantile function based on the j-th bootstrap sample. In practice, the integral is approximated using numerical integration.

Two alternative approaches were attempted. First, a bootstrap approach, fixing the covariates X and sampling for each covariate level an uniform random variable U. For values of $U \geq \tau_c$ a positive residual e is sampled and the bootstrap sample is $Y^b = \hat{Q}^{h_0}_{Y|X}(X) + e$. In the case $U < \tau_c$ a local linear quantile estimate is obtained at the covariate level X with bandwidth h_0 at probability level U. The bandwidth is then estimated by the solution of the minimization in (2.8). Second, a leave-one-out cross validation approach that minimizes the quantile loss function is used to obtain the estimator of the optimal bandwidth:

$$\hat{h} = \arg\min_h \hat{S}(h) = \arg\min_h \sum_{i=1}^{n} \rho_{\tau_c}(Y_i - \hat{Q}^{h,-i}_{Y|X}(\tau_c|X_i)),$$

where $\hat{Q}^{h,-i}_{Y|X}$ denotes the conditional quantile estimate with bandwidth h and leaving out the ith observation. Intuitively, the cross validation approach is attractive as it is much faster compared to the bootstrap approach and it is based on the idea of scoring the quantile curve with the same scoring function used for estimation. Yet, based on a simulation study, the direct bootstrap procedure performed significantly better compared to these alternative approaches. This is in accordance with the conclusions drawn in [4].

2.5. SIMULATION

In this section, the finite sample performance of the CST-estimator is assessed using a detailed simulation study. A comparison is made with the estimator proposed in [63], where also a two step procedure is used. The first step consists of estimating a sequence of linear quantile curves for moderately high probability levels, using quantile regression. And the second step then uses a Hill estimator for the extreme value index based on the estimated quantiles. Extrapolation to the extreme quantiles is done by a Weissman type estimator, similar to the one in (2.5).

Define the simulation model from which the data is drawn by,

$$Y = r(X) + \sigma(X)\epsilon. \tag{2.9}$$

We choose X uniformly distributed in $[-1,1]$ and independently, ϵ follows from a generalized Pareto distribution with $\gamma = 0.25$, or a Student t_1 distribution. For the function σ, we consider two cases: $\sigma(x) = 1$ and $\sigma(x) = \frac{4+x}{4}$. Note that for $\sigma(x) = 1$, our model assumption (2.1) is satisfied with $\tau_c = 0$. For $\sigma(x) = \frac{4+x}{4}$, our model assumption is not satisfied since the distribution of the additive noise depends on x, which allows us to study the robustness of the model assumptions.

We consider three choices for the function r: linear, nonlinear monotone and a more wiggly function,

$$r_1(x) = x,\ r_2(x) = \exp(x),\ r_3(x) = \sin(2\pi x)(1 - \exp(x))$$

Performance is compared for two sample sizes : $n = 500$ and $n = 2500$.

The estimation of the quantile curves $x \mapsto Q_{Y|X}(\tau|x)$ with $\tau = 0.99$ and $\tau = 0.995$ is assessed with an empirical estimator of the mean integrated squared error: $\frac{1}{m}\sum_{i=1}^{m}\int_{-1}^{1}(\hat{Q}^{(i)}_{Y|X}(\tau|x) - Q_{Y|X}(\tau|x))^2 dx$, where $m = 500$ and $\hat{Q}^{(i)}_{Y|X}(\tau|x)$ denotes the estimate based on the i-th sample. The integral is approximated by numerical integration. Tables 2.1 and 2.2 report the estimated MISE for different models and different methods.

For the CST estimator, we choose $\tau_c = 0.5$ while the model holds for any $\tau_c \geq 0$. Simulations show that the results are not sensitive to the level of τ_c that is chosen. The value of k is typically chosen by inspection at the point where the Hill plot, i.e $(k, \hat{\gamma}(k))$, becomes stable. In the simulation study it is not possible to choose the stable point for every simulation. Therefore, we choose a fixed $k = [4n^{1/4}]$, where [.] denotes the integer part. From simulations we see that the estimate becomes stable around this value of k.

For the estimator in [63], it is proposed to choose $k = [4.5n^{1/3}]$. Additionally, the probability sequence for which the linear quantile curves are estimated is given by, $\frac{n-k}{n},\ldots,\frac{n-3}{n}$, trimming of the most extreme quantiles, $\frac{n-2}{n},\ldots,\frac{n}{n}$. This is needed in order to obtain a Bahadur expression for the regression quantiles. In [63] it is suggested to trim off $[n^\eta]$ observations, with $\eta \in (0, 0.2)$. In our simulation trimming off the three most extreme probabilities gave the best performance. The estimator allows for varying extreme value indices as well as a constant extreme value index. A constant extreme value index is used as this is assumed in our setting. We refer to this estimator as the linear estimator. The model assumption for this method is satisfied only when $r = r_1$, the linear case.

For generalized Pareto errors, the mean integrated squared errors are shown in Table 2.1. For the case $\sigma(x) = 1$, the CST estimator performs best, as expected, since the data follow the model assumption (2.1). For the case $\sigma(x) = \frac{4+x}{4}$ a similar conclusion can be drawn for $n = 500$. Though, for a sample size of 2500 the linear estimator does slightly better. The deviation from the model assumption clearly affects the behaviour of the CST estimator, but not the linear estimator. The difference between the methods becomes visible for larger sample sizes as the bias for the CST estimator starts to play a bigger role in the MISE.

For Student t_1 errors, the results are shown in Table 2.2. For sample size $n = 500$, the CST estimator has smaller MISE for $\tau = 0.99$ and larger MISE for $\tau = 0.995$, in comparison with the linear estimator. For a larger sample size $n = 2500$, the CST estimator outperforms the linear method. For small sample size the r is subject to high variance locally, this leads to errors in the residuals and as a result in the extreme value index. This is shown in the extrapolation to the 0.995 quantile. When the sample size is larger this is not an issue, which leads to better performance of the CST estimator. The relative effect of the deviation from the model by choosing $\sigma(x) = \frac{4+x}{x}$ is lower now for a large $\gamma = 1$. As a result the CST estimator performs better sometimes for large sample size and $\sigma(x) = \frac{4+x}{x}$.

Remark 4. *The estimator that is proposed in [14] was also compared to the CST estimator and the linear estimator and was outperformed clearly in all instances by these methods, although it is the only method for which the model assumptions are satisfied for all settings. The procedure does not assume any structure in the data and it allows for varying extreme value indices, which requires to estimate the extreme value index locally by using a very limited amount of observations. As a consequence, the function $\hat{\gamma}(x)$ fluctuates*

Table 2.1: Mean integrated squared errors based on samples from (2.9), with errors GPD($\gamma = 0.25$).

		$\sigma(x) = 1$		$\sigma(x) = \frac{4+x}{4}$	
r	method	0.99	0.995	0.99	0.995
		$n = 500$			
r_1	CST	**2.62**	**9.16**	**5.28**	**14.42**
r_1	linear	9.04	18.53	7.75	15.66
r_2	CST	**2.78**	**9.51**	**5.64**	**15.04**
r_2	linear	8.69	18.92	8.57	18.47
r_3	CST	**2.66**	**8.01**	**5.27**	**13.95**
r_3	linear	9.05	18.83	8.98	18.55
		$n = 2500$			
r_1	CST	**0.64**	**1.59**	3.23	5.91
r_1	linear	2.04	6.14	**1.88**	**5.53**
r_2	CST	**0.71**	**1.70**	3.24	5.85
r_2	linear	1.95	6.09	**1.86**	**5.64**
r_3	CST	**0.75**	**1.56**	3.42	6.04
r_3	linear	2.15	5.98	**2.12**	**5.91**

heavily and it further creates large inaccuracies in the quantile extrapolation. From the simulation result, it is clear that this method suffers severely from lack of efficiency for the sample sizes considered here. Therefore, the results were left out to focus on the comparison between the CST and the linear method.

2.6. POST-PROCESSING EXTREME PRECIPITATION

Our dataset consists of observations and ECMWF ensemble forecasts of daily accumulated precipitation at eight meteorological stations spread across the Netherlands (de Bilt, De Kooy, Twente, Eelde, Leeuwarden, Beek, Schiphol and Vlissingen). The data in this study is for the warm half year, namely 15th of April until 15th of October, in the years 2011 till 2017. The lead time is defined as the time between initialization of the ensemble run and the end of the day at 00 UTC for which the forecast is valid. We consider lead times from 24 hours up till 240 hours with 12 hour increments. For each lead time and location the number of observations is about 1287.

For fixed lead time and location, an ensemble forecast consists of 51 exchangeable members, which can be seen as a sample from the distribution of precipitation, where the uncertainty in the initial condition and model parametrizations are accounted for. As a result, quantile estimates for probability levels $\frac{i}{52}$, for $1 \leq i \leq 51$, are given by the order statistics of the ensemble forecast. Note that the precipitation observations are not used by the ensemble forecast as standard the amount of precipitation is set to zero at initialization of the NWP model.

In practice, it is known that the upper ensemble member is not well calibrated in the sense that it leads to underestimation of the extremes, see [5]. This is partly caused by a representatively error, because the forecast is a grid-cell average and the observation

Table 2.2: Mean integrated squared errors $\times 10^{-2}$ based on samples from (2.9), with errors from Student t_1.

		$\sigma(x) = 1$		$\sigma(x) = \frac{4+x}{4}$	
r	method	0.99	0.995	0.99	0.995
		$n = 500$			
r_1	CST	**3.41**	31.69	**3.33**	28.83
r_1	linear	4.69	**26.66**	4.74	**26.82**
r_2	CST	**3.83**	38.35	**4.40**	43.25
r_2	linear	5.19	**30.44**	5.01	**29.81**
r_3	CST	**3.97**	40.56	**3.44**	**29.47**
r_3	linear	4.78	**27.62**	5.32	30.30
		$n = 2500$			
r_1	CST	**0.69**	**5.14**	**1.20**	**6.49**
r_1	linear	1.30	10.68	1.35	10.94
r_2	CST	**0.82**	**5.98**	1.26	**7.31**
r_2	linear	1.27	10.70	**1.24**	10.31
r_3	CST	**0.83**	**6.03**	**1.17**	**7.10**
r_3	linear	1.38	11.30	1.32	10.90

is a station point value. Statistical post-processing can correct this and other systematic errors [65]. For long lead times, a forecast, especially the upper ensemble member loses all predictive skill, [5]. We show that, by applying the CST estimator, we can calibrate the upper ensemble member and obtain more skilful forecasts for short and long lead times. To relate to the notation of Section 2.2, we denote the daily accumulated precipitation by Y and the upper ensemble member by X.

For each lead time we pool data from all eight locations. These locations are spread over the Netherlands and as most extreme events are caused by local deep convective showers, the observations can be considered approximately independent. We compare the performance of the ensemble method with the CST estimator as in (2.6) and the linear estimator as explained in Section 2.5.

As precipitation is often modelled using a point mass on 0 for the dry days, we model the point mass using a logistic regression with as covariate the number of ensemble members equal to zero. The distribution function is then given by:

$$F_{Y|X}(y|x) = p_0(x) + (1 - p_0(x))F_{Y|X,Y>0}(y|x) \tag{2.10}$$

Where the quantiles are given by:

$$Q_{Y|X}(\tau|x) = \begin{cases} 0 & \text{if } \tau \leq p_0(x) \\ Q_{Y|X,Y>0}\left(\frac{\tau - p_0(x)}{1 - p_0(x)}\right) & \text{if } \tau > p_0(x) \end{cases} \tag{2.11}$$

We then apply the CST estimator to estimate $Q_{Y|X,Y>0}$, where we choose $\tau_c = 0.95$. This choice is based on best validation score, as explained below, based on one year of data. The bandwidth h is determined using the bandwidth selection method described

in Section 2.4 and $k = [4n^{1/4}]$, the same as in the simulation study. Alternative to choosing X as the upper ensemble member we have also considered other ensemble members and trimmed means of the ensemble members. Among these choices the upper ensemble member showed best performance.

For the linear method we do not incorporate the point mass as the method already takes this into account as all quantiles are estimated globally instead of the CST estimator, which estimates the quantiles in a local manner. Incorporating the point mass led to severely worse results for the linear method. The same hyper parameters were chosen as in Section 2.5; changing these did not influence the results.

Note that for days that have a large point mass on 0 and the rescaled probability is not extreme, in these cases we just use a local linear quantile estimator as described in Equation 2.3 as the estimator of $Q_{Y|X,Y>0}\left(\frac{\tau-p_0(x)}{1-p_0(x)}\right)$.

The predictive performance of a quantile estimator $\hat{Q}_i(\tau)$ can be quantified by the quantile verification score and visualized by the quantile reliability diagram, which are discussed in detail in [6]. The quantile verification score is defined as,
$\mathrm{QVS}_\tau(\hat{Q}) = \sum_{i=1}^n \rho_\tau(Y_i - \hat{Q}_i(\tau))$, where ρ_τ is the quantile check function. The score is always positive, where low scores represent good performance and high scores bad performance. In [6] it is shown that the score can be decomposed in three components: uncertainty, reliability and resolution, where only the last two depend on the estimator itself. A reliable or calibrated forecast has the same distribution as the underlying distribution that is estimated.

The quantile reliability diagram visualizes the reliability of the forecast quantile by creating equally sized bins with respect to the forecast quantile and then graphing the empirical quantile of the corresponding observations in the bin against the mean forecast quantile in the bin. For the forecast to be reliable these points should lie on the line $y = x$.

It is natural to compare the predictive performance of a quantile estimator to some reference quantile estimator $\hat{Q}_{\mathrm{ref}}$. For this we take the climatological empirical quantiles as the reference method, i.e. the empirical quantiles of the sample Y_i, $1 \le i \le n$. Note that this is the simplest estimate we can obtain without making use of a numerical weather prediction model. The quantile verification skill score, given by $\mathrm{QVSS}_\tau(\hat{Q}) = 1 - \frac{\mathrm{QVS}_\tau(\hat{Q})}{\mathrm{QVS}_\tau\left(\hat{Q}_{\mathrm{ref}}\right)}$, is a relative measure of performance compared to the reference method, taking values in $(0,1]$ when $\hat{Q}$ improves on $\hat{Q}_{\mathrm{ref}}$ and values below zero when the opposite is true.

The validation is carried out using a seven-fold cross validation, where, in every iteration, one year is left out of the model estimation and used as the independent validation sample. In Figure 2.1 the QVSS is shown as a function of lead time. The bands are obtained by calculating the QVSS for each location separately. The graph on the left shows the performance of the CST estimator in red, the linear estimator in green and the ensemble in blue for $\tau = \frac{51}{52}$. It can be observed that the CST and the linear estimator improve upon the ensemble especially for short lead times and for very long lead times. On the right side of the figure the performance of the $\tau = 0.995$ quantile is shown for the CST and the linear estimators, showing that skilful quantile estimates are obtained up till 144 hours. The CST estimator seems to have slightly less spread in the scores than the

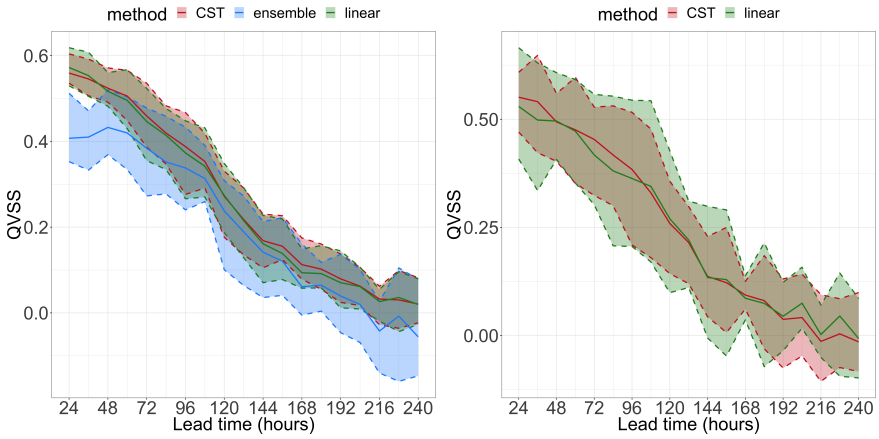


Figure 2.1: QVSS as a function of lead time for CST estimator in red, the ensemble in blue and the linear estimator in green, on the left for the $\frac{51}{52}$ quantile and on the right for the 0.995 quantile. The bands are obtained by validating for each location separately.

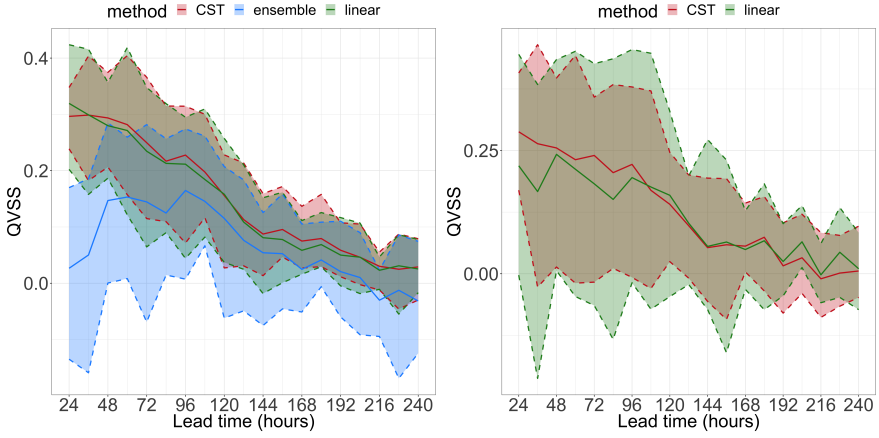


Figure 2.2: QVSS as a function of lead time conditioned on $X > 5$, for CST estimator in red, the ensemble in blue and the linear estimator in green, on the left for the $\frac{51}{52}$ quantile and on the right for the 0.995 quantile. The bands are obtained by validating for each location separately.

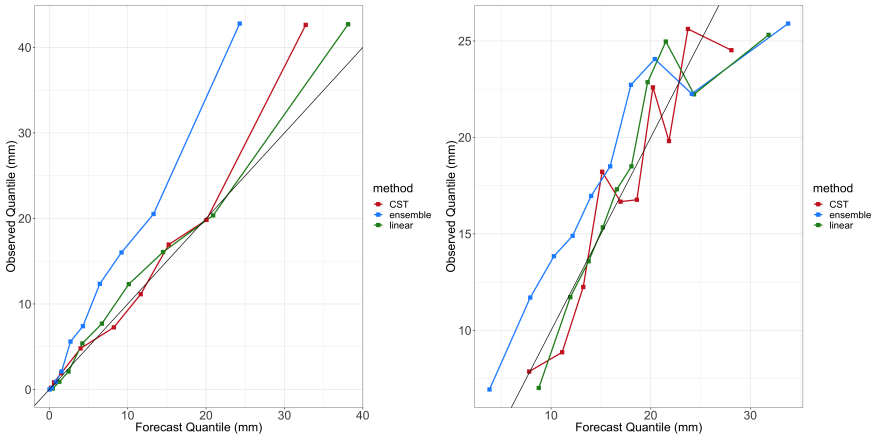


Figure 2.3: Quantile reliability diagrams for the CST estimator in red, the ensemble in blue and the linear estimator in green for the $\frac{51}{52}$ quantile; on the left side for 24 hours lead time and on the right side 192 hours lead time.

linear method.

In practice the quantile estimates are of interest when the ensemble is already high, i.e. $X > t$ for t large. In Figure 2.2 similar plots are shown as in Figure 2.1, but now the verification is done based on a subset of the data where we condition on $X > 5$, which is the 60 percent quantile for a lead time of 24 hours. Note that this means that also the reference climatological quantile has this conditioning. It can be seen in the left panel of Figure 2.2 that the ensemble method is outperformed by the CST and the linear estimator for shorter lead times. For the extrapolation to $\tau = 0.995$ in the right panel of Figure 2.2, the spread in skill of the different stations is much larger, but still showing skilful forecasts for most stations for short lead times. Also here the CST appears to have less spread than the linear estimator. In Figure 2.3 two quantile reliability diagrams are shown, for 24 hours lead time on the left and 192 hours lead time on the right, using all data without conditioning. The ensemble clearly underestimates the extremes generally for both lead times. The CST and the linear estimators improve calibration for 24 hour lead time. For a lead time of 192 hours the CST estimator looks a bit more unstable, though it remains close the the calibration line, where the ensemble is consistently underestimating the upper quantile.

From all plots it can be concluded that the CST and the linear estimator are very comparable, an assumption of linear quantiles is in this context also not strange. Even though the CST estimator has a more flexible assumption on the quantile curves, it does not influence the results.

To conclude, we have shown that the CST estimator is comparable to the linear estimator and has more skill than the upper ensemble member for both short and long lead times. Additionally, it is able to extrapolate further into the tail and obtains skilful estimates for higher quantiles than are available from the ensemble.

2.7. DISCUSSION

We have estimated the conditional tail quantile curves, $x \mapsto Q_{Y|X}$, using a two step procedure. First we use local linear quantile regression to estimate a non-stationary threshold and secondly, extrapolate to the tail using the exceedances of this threshold. The assumption that $\gamma > 0$ fits to the application of summer precipitation in the Netherlands, which is heavy tailed. There is a clear motivation for extending the model to the cases of light tailed, $\gamma = 0$ and short tailed data, $\gamma < 0$. This would enable also post-processing of extreme precipitation in winter periods, but also temperature, wind speed and gusts and other weather phenomena.

It is clear from the simulation that the linear method from [63] is better able to deal with heteroskedastic data. Extending the model to allow for non-homoskedastic errors would be a valuable addition, allowing it to model data from a wider range of classes.

Finally, in the application we now calibrate tail quantiles of the ensemble, using the statistical relation between the upper ensemble member and the observations. It would be of interest though, to consider a wider range of covariates from the NWP model. It would therefore be of value to extend the method to a multivariate covariates setting.

2.A. PROOFS

This section contains the proofs of Theorems 1-3 in Section 2.3. Throughout this section, $c, c_1, c_2, \ldots$ denote positive constants, which are not necessarily the same at each occurrence.

2.A.1. PROOF OF THEOREM 1

The uniform consistency of $\hat{r}$ relies heavily on the uniform Bahadur representation for $\hat{r}$. We make use of the Bahadur representation obtained in [39].

Let $\psi_\tau(u) = \tau - I(u < 0)$, that is the right derivative of ρ_τ at u. Then by Corollary 3.3 and Proposition 1 in [39], we have

$$\sup_{x\in[a,b]} \left| \hat{r}(x) - r(x) + h_n^2 c r''(x) - \frac{1}{nh_n} \sum_{i=1}^{n} \psi_{\tau_c}(\epsilon_i) C_{n,i}(x) K\left(\frac{X_i - x}{h_n}\right) \right|$$
$$= O_p\left(\left\{ \frac{\log n}{nh_n} \right\}^{3/4} \right) = O_p\left(\left\{ \frac{\log n}{n^{1-\delta_h}} \right\}^{3/4} \right),$$

where $C_{n,i}(x)$ is a Lipschitz continuous function and thus absolutely bounded in $[a,b]$. Define

$$\Delta_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \psi_{\tau_c}(\epsilon_i) C_{n,i}(x) K\left(\frac{X_i - x}{h_n}\right).$$

Then, the triangle inequality leads to

$$\sup_{x\in[a,b]} |\hat{r}(x) - r(x)| \le \sup_{x\in[a,b]} \left| h_n^2 c r''(x) \right| + \sup_{x\in[a,b]} |\Delta_n(x)| + O_p\left(\left\{ \frac{\log n}{n^{1-\delta_h}} \right\}^{3/4} \right)$$
$$= O(n^{-2\delta_h}) + \sup_{x\in[a,b]} |\Delta_n(x)| + O_p\left(\left\{ \frac{\log n}{n^{1-\delta_h}} \right\}^{3/4} \right). \tag{2.12}$$

The last equality follows from the fact that r'' is uniformly bounded by Assumption A1.

Next, we show that, there exists a $\delta_C \in (0, \frac{1}{2} - \delta_h)$ such that

$$\sup_{x\in[a,b]} |\Delta_n(x)| = o_p(n^{-\delta_C}). \tag{2.13}$$

2

Define $T_i(x) := h_n K\left(\frac{X_i - x}{h_n}\right) C_{n,i}(x)$. Then for any $x, y \in [a,b]$, by the triangle inequality and the Lipschitz continuity of K, we have

$$\begin{aligned}
&\left|T_i(x) - T_i(y)\right| = h_n \left| K\left(\frac{X_i - x}{h_n}\right) C_{n,i}(x) - K\left(\frac{X_i - y}{h_n}\right) C_{n,i}(y) \right| \\
\leq & h_n |C_{n,i}(x)| \left| K\left(\frac{X_i - x}{h_n}\right) - K\left(\frac{X_i - y}{h_n}\right) \right| + h_n K\left(\frac{X_i - y}{h_n}\right) |C_{n,i}(x) - C_{n,i}(y)| \\
\leq & c_1 \left|x - y\right| + c_2 h_n |x - y| \sup_{u\in[-1,1]} K(u) \\
\leq & c|x - y|.
\end{aligned}$$

Note that the constant c does not depend on i, that is, the Lipschitz continuity is uniform in i for all T_i's. Consequently, it follows from that $\left|\psi_\tau(u)\right| \leq 1$ that,

$$|\Delta_n(x) - \Delta_n(y)| = \frac{1}{nh_n^2} \left| \sum_{i=1}^n \psi_{\tau_c}(\epsilon_i)(T_i(x) - T_i(y)) \right| \leq c \frac{|x - y|}{h_n^2}.$$

Let $M_n = n^{\delta_C + 2\delta_h} \log n$ and $\{I_i = (t_i, t_{i+1}], i = 1, \ldots, M_n\}$ be a partition of $(a,b]$, where $t_{i+1} - t_i = \frac{b-a}{M_n}$. Then for $t \in I_i$,

$$|\Delta_n(t) - \Delta_n(t_i)| \leq \frac{c(b-a)}{M_n h_n^2},$$

or equivalently,

$$\Delta_n(t_i) - \frac{c(b-a)}{M_n h_n^2} \leq \Delta_n(t) \leq \Delta_n(t_i) + \frac{c(b-a)}{M_n h_n^2}.$$

Therefore, for n sufficiently large,

$$\begin{aligned}
&\mathrm{P}\left(\sup_{x\in[a,b]} |\Delta_n(x)| > n^{-\delta_C} \right) && = \mathrm{P}\left(\max_{1\leq i\leq M_n} \sup_{t\in I_i} |\Delta_n(t)| > n^{-\delta_C} \right) \\
&\leq \sum_{i=1}^{M_n} \mathrm{P}\left(\sup_{t\in I_i} |\Delta_n(t)| > n^{-\delta_C} \right) && \leq \sum_{i=1}^{M_n} \mathrm{P}\left(|\Delta_n(t_i)| > n^{-\delta_C} - \frac{c(b-a)}{M_n h_n^2} \right) \\
&\leq \sum_{i=1}^{M_n} \mathrm{P}\left(|\Delta_n(t_i)| > \frac{1}{2} n^{-\delta_C} \right) && = \sum_{i=1}^{M_n} \mathrm{P}\left(\left| \sum_{j=1}^n \frac{T_j(t_i)\psi_{\tau_c}(\epsilon_j)}{h_n} \right| > \frac{1}{2} h_n n^{1-\delta_C} \right) =: \sum_{i=1}^{M_n} P_i,
\end{aligned}$$

where the third inequality is due to that $\frac{c(b-a)}{M_n h_n^2} < \frac{1}{2} n^{-\delta_C}$ for n sufficiently large. Next, we apply Hoeffding's inequality to bound P_i. Define

$$W_{n,i,j} := \frac{T_j(t_i)\psi_{\tau_c}(\epsilon_j)}{h_n} = K\left(\frac{X_j - t_i}{h_n}\right) C_{n,j}(t_i)\psi_{\tau_c}(\epsilon_j).$$

For each i and n, $\{W_{n,i,j}, 1 \le j \le n\}$ is a sequence of i.i.d. random variables. And with probability one, $|W_{n,j,i}| \le \sup_{-1\le u\le 1} K(u) \sup_{a\le x\le b} C_{n,i}(x) =: c_3$. Moreover, $\mathbb{E}\left(W_{n,j,i}\right) = 0$ because $\mathbb{E}(\psi_{\tau_c}(\epsilon_j)) = 0$ and X_j and ϵ_j are independent. Thus, by Hoeffding's inequality,

$$P_i = \mathrm{P}\left(\left|\sum_{j=1}^{n} W_{n,i,j}\right| \ge \frac{1}{2} h_n n^{1-\delta_C}\right) \le 2\exp\left(-\frac{n^{1-2\delta_C} h_n^2}{8c_3^2}\right) = 2\exp\left(-cn^{1-2\delta_h-2\delta_C}\right).$$

Note that $1-2\delta_h-2\delta_C > 0$ by the choice of δ_C. Thus, for $n \to \infty$,

$$\mathrm{P}\left(\sup_{x\in[a,b]} |\Delta_n(x)| > n^{-\delta_C}\right) \le 2M_n \exp\left(-cn^{1-2\delta_h-2\delta_C}\right) \to 0.$$

Hence, (2.13) is proved. Now by choosing $\delta = \delta_C$, we obtain via (2.12) that,

$$n^{\delta} \sup_{x\in[a,b]} |\hat{r}_n(x) - r(x)| = O(n^{\delta_C-2\delta_h}) + o_p(1) + O_p\left(n^{-\frac{3}{4}+\frac{3}{4}\delta_h+\delta_C} (\log n)^{\frac{3}{4}}\right) = o_p(1),$$

due to that $\delta_h \in (\frac{1}{5}, \frac{1}{2})$ and $\delta_C < \frac{1}{2} - \delta_h$.

2.A.2. PROOF OF THEOREM 2

The proof follows a similar line of reasoning as that of Theorem 2.1 in [63]. The uniform consistency of $\hat{r}_n$ given in Theorem 1 plays a crucial role. Define $V_n := ||\hat{r}_n - r||_\infty = o_p\left(n^{-\delta}\right)$.

Let $U_i = F_{Y|X}(Y_i|X_i)$ for all $1 \le i \le n$. Then $\{U_i, i = 1, \ldots, n\}$ constitute i.i.d. random variables from a standard uniform distribution. Recall the definition of e_i:

$$e_i = Y_i - \hat{r}_n(X_i) = Q_{Y|X}(U_i|X_i) - \hat{r}_n(X_i).$$

Thus, the ordering of $\{e_i, i = 1, \ldots, n\}$ is not necessarily the same as the ordering of $\{U_i, i = 1, \ldots, n\}$. The main task of this proof is to show that the k_n largest e_i's correspond to the k_n largest U_i's; see (2.15). To this aim, we first prove that with probability tending to one, $e_{n-j,n}$ for $j = 0, \ldots, k_n$ can be decomposed as follows,

$$e_{n-j,n} = Q_\epsilon(U_{i(j)}) + r(X_{i(j)}) - \hat{r}_n(X_{i(j)}) \text{ for } j = 0, \ldots k_n, \tag{2.14}$$

where $i(j)$ is the index function defined as $e_{i(j)} = e_{n-j,n}$. In view of (2.4), it is sufficient to prove that with probability tending to one, $U_{i(j)} > \tau_c$ jointly for all $j = 0, \ldots, k_n$. Define

another index function, $\tilde{i}(j)$ by $U_{\tilde{i}(j)} = U_{n-j,n}$. Then it follows for n large enough,

$$\begin{aligned}
\mathrm{P}\left(\cup_{j=0}^{k_n}\{U_{i(j)} < \tau_c\}\right) &= \mathrm{P}\left(\cup_{j=0}^{k_n}\{Y_{i(j)} < Q_{Y|X}(\tau_c|X_{i(j)})\}\right) \\
&= \mathrm{P}\left(\min_{0\le j\le k_n}\left(Y_{i(j)} - r(X_{i(j)})\right) < 0\right) \\
&= \mathrm{P}\left(\min_{0\le j\le k_n}\left(Y_{i(j)} - \hat{r}_n(X_{i(j)}) - r(X_{i(j)}) + \hat{r}_n(X_{i(j)})\right) < 0\right) \\
&\le \mathrm{P}\left(\min_{0\le j\le k_n} e_{n-j,n} - \sup_{x\in[a,b]}|\hat{r}_n(x) - r(x)| < 0\right) \\
&= \mathrm{P}\left(e_{n-k_n,n} < V_n\right) = 1 - \mathrm{P}(e_{n-k_n,n} \ge V_n) \\
&\le 1 - \mathrm{P}\left(\cap_{j=0}^{k_n}\{e_{\tilde{i}(j)} \ge V_n\}\right) \\
&= 1 - \mathrm{P}\left(\cap_{j=0}^{k_n}\left\{Q_\epsilon(U_{n-j,n}) + r(X_{\tilde{i}(j)}) - \hat{r}_n(X_{\tilde{i}(j)}) \ge V_n\right\}\right) \\
&\le 1 - \mathrm{P}\left(Q_\epsilon(U_{n-k_n,n}) \ge 2V_n\right),
\end{aligned}$$

where the second equality follows from that $Q_{Y|X}(\tau_c|X_{i(j)}) = r(X_{i(j)})$ and the last equality follows from (2.4) and the fact that $U_{n-k_n,n} > \tau_c$ for n large enough. Then, $\lim_{n\to\infty}\mathrm{P}\left(\cup_{j=0}^{k_n}\{U_{i(j)} < \tau_c\}\right) = 0$ follows from $Q_\epsilon(U_{n-k_n,n}) \to \infty$ and $V_n = o_p(1)$ as $n \to \infty$. Hence, (2.14) is proved.

Next, we show that

$$\lim_{n\to\infty}\mathrm{P}\left(\cap_{j=0}^{k_n}\{e_{n-j,n} = Q_\epsilon(U_{n-j,n}) + r(X_{i(j)}) - \hat{r}_n(X_{i(j)})\}\right) = 1, \tag{2.15}$$

that is the ordering of k largest residuals is determined by the ordering of U_i's. In view of (2.14), it is sufficient to show that with probability tending to one,

$$\min_{1\le i\le k_n}\left(Q_\epsilon(U_{n-i+1,n}) - Q_\epsilon(U_{n-i,n})\right) \ge 2\max_{1\le i\le k_n}|r(X_{i(j)}) - \hat{r}_n(X_{i(j)}|. \tag{2.16}$$

By the second order condition given in (2.7) and Theorem 2.3.9 in [17], for any small $\delta_1, \delta_2 > 0$, and n large enough,

$$\frac{Q_\epsilon(U_{n-i+1,n})}{Q_\epsilon(U_{n-i,n})} \ge W_i^\gamma + A_0\left(\frac{1}{1-U_{n-i,n}}\right)W_i^\gamma\frac{W_i^\rho - 1}{\rho} - \delta_1\left|A_0\left(\frac{1}{1-U_{n-i,n}}\right)\right|W_i^{\gamma+\rho+\delta_2}, \tag{2.17}$$

for $i = 1,\ldots,k_n$, where $W_i = \frac{1-U_{n-i,n}}{1-U_{n-i+1,n}}$ and $\lim_{t\to\infty}A_0(t)/A(t) = 1$. Observe that $\log W_i = \log\frac{1}{1-U_{n-i+1,n}} - \log\frac{1}{1-U_{n-i,n}} \stackrel{d}{=} E_{n-i+1,n} - E_{n-i,n}$ with E_i's i.i.d. standard exponential variables. Thus, by Rènyi's representation [49], we have

$$\{W_i, 1 \le i \le k_n\} \stackrel{d}{=} \left\{\exp\left(\frac{E_i}{i}\right), 1 \le i \le k_n\right\}.$$

From Proposition 2.4.9 in [17], we have $\frac{U_{n-k_n,n}}{1-\frac{k_n}{n}} \xrightarrow{P} 1$, which implies that $A_0\left(\frac{1}{1-U_{n-k_n,n}}\right) = O_p\left(A_0\left(\frac{n}{k_n}\right)\right)$. Using the fact that A_0 is regularly varying with index ϱ, hence $|A_0|$ is ultimately decreasing, we obtain for n sufficiently large and any $i = 1,\dots,k_n$,

$$\begin{aligned}\left|A_0\left(\frac{1}{1-U_{n-i,n}}\right)\right| &\le \left|A_0\left(\frac{1}{1-U_{n-k_n,n}}\right)\right| \\ &= \left|O_p\left(A_0\left(\frac{n}{k_n}\right)\right)\right| = \left|O_p\left(A\left(\frac{n}{k_n}\right)\right)\right| = \left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|, \end{aligned} \tag{2.18}$$

by the assumption $\sqrt{k_n}A\left(\frac{n}{k_n}\right) \to \lambda$.

For a sufficiently large u and any $k_n \ge 1$,

$$\begin{aligned}&\mathrm{P}\left(\max_{1\le i\le k_n}\frac{E_i}{i} \le u\right) = \prod_{i=1}^{k_n}\left(1-e^{-iu}\right) = \exp\left(\sum_{i=1}^{k_n}\log\left(1-e^{-iu}\right)\right) \\ &= \exp\left(-\sum_{i=1}^{k_n}\sum_{j=1}^{\infty} j^{-1}e^{-iuj}\right) \ge \exp\left(-\sum_{i=1}^{k_n} e^{-iu}\right) = \exp\left(\frac{1-e^{-ku}}{1-e^{u}}\right),\end{aligned}$$

which tends to one as $u \to \infty$. This implies that

$$\min_{1\le i\le k_n} W_i^{\varrho} \overset{d}{=} \exp\left(\varrho \max_{1\le i\le k_n}\frac{E_i}{i}\right) = O_p(1). \tag{2.19}$$

Thus, combining (2.17), (2.18) and (2.19), we have

$$\begin{aligned}&\min_{1\le i\le k_n}\frac{Q_\epsilon(U_{n-i+1,n})}{Q_\epsilon(U_{n-i,n})} - 1 \\ &\ge \min_{1\le i\le k_n} W_i^{\gamma}\left(1-\left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|\left(\frac{W_i^{\varrho}-1}{\varrho}+\delta_1 W_i^{\varrho+\delta_2}\right)\right) - 1 \\ &= \min_{1\le i\le k_n} W_i^{\gamma}\left(1-\left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|\right) - 1 \overset{d}{=} \exp\left(\gamma\frac{E_1}{k_n}\right)\left(1-\left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|\right) - 1 \\ &= \frac{\gamma E_1}{k_n}\left(1-\left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|\right),\end{aligned}$$

where the third equality follows from that $\min_{1\le i\le k_n}\frac{E_i}{i} \overset{d}{=} E_{1,k} \overset{d}{=} \frac{E_1}{k}$ by Rènyi's representation. Thus, we obtain that

$$\begin{aligned}\min_{1\le i\le k_n}\left(Q_\epsilon(U_{n-i+1,n}) - Q_\epsilon(U_{n-i,n})\right) &\ge \left(Q_\epsilon(U_{n-k_n,n})\frac{\gamma E_1}{k_n}\right)\left(1-\left|O_p\left(\frac{1}{\sqrt{k_n}}\right)\right|\right) \\ &= \left(\frac{n}{k_n}\right)^{\gamma} k_n^{-1}|O_p(1)|.\end{aligned}$$

Thus, (2.16) is proved by the assumption $k_n^{-1}\left(\frac{n}{k_n}\right)^{\gamma} >> n^{-\delta}$ and $\max_{1\le i\le k_n}|r(X_{i(j)}) - \hat{r}_n(X_{i(j)}| \le 2V_n = o_p\left(n^{-\delta}\right)$. Intuitively, (2.16) means that the difference between two successive upper order statistics of ϵ is larger than the error made in the estimation of $r(x)$.

As aforementioned, (2.14) and (2.16) together lead to (2.15), which further implies that with probability tending to one,

$$\max_{0\le j\le k_n}\left|\frac{e_{n-j,n}}{Q_\epsilon(U_{n-j,n})}-1\right|\le\frac{V_n}{Q_\epsilon(U_{n-k_n,n})}=o_p\left(n^{-\delta}\left(\frac{k_n}{n}\right)^{\gamma}\right). \tag{2.20}$$

By the definition of $\hat{\gamma}_n$ and (2.20), we can write the estimator as follows,

$$\begin{aligned}\hat{\gamma}_n&=\frac{1}{k_n}\sum_{i=0}^{k_n-1}\log\frac{e_{n-i,n}}{e_{n-k_n,n}}\\&=\frac{1}{k_n}\sum_{i=0}^{k_n-1}\log\frac{Q_\epsilon(U_{n-i,n})}{Q_\epsilon(U_{n-k_n,n})}+\left(\frac{1}{k_n}\sum_{i=0}^{k_n-1}\log\frac{e_{n-i,n}}{Q_\epsilon(U_{n-i,n})}-\log\frac{e_{n-k_n,n}}{Q_\epsilon(U_{n-k_n,n})}\right)\\&=:\hat{\gamma}_H+o_p\left(n^{-\delta}\left(\frac{k_n}{n}\right)^{\gamma}\right).\end{aligned}$$

The first part is the well known Hill estimator and we have by Theorem 3.2.5 in [17],

$$\sqrt{k_n}(\hat{\gamma}_H-\gamma))\xrightarrow{d}N\left(\frac{\lambda}{1-\varrho},\gamma^2\right).$$

Therefore we can conclude,

$$\sqrt{k_n}(\hat{\gamma}_n-\gamma)=\sqrt{k_n}(\hat{\gamma}_H-\gamma)+o_p\left(\sqrt{k_n}n^{-\delta}\left(\frac{k_n}{n}\right)^{\gamma}\right)\xrightarrow{d}N\left(\frac{\lambda}{1-\varrho},\gamma^2\right),$$

by the assumption that $k_n^{\gamma+1}n^{-\gamma-\delta}\to 0$.

We remark that the proof for Theorem 2.1 in [63] isn't completely rigorous, namely, the proof for (S.1) in the supplementary material of that paper is not right. We fix the problem while proving (2.20), which is an analogue to (S.1).

2.A.3. PROOF OF THEOREM 4

Before we proceed with the proof of Theorem 3, we state the asymptotic normality of $\hat{Q}_\epsilon(\tau_n)$ defined in (2.5) in the theorem below.

Theorem 4. *Let the conditions of Theorem 2 be satisfied. Assume* $np_n=o(k_n)$ *and* $\log(np_n)=o(\sqrt{k_n})$*, then, as* $n\to\infty$*,*

$$\frac{\sqrt{k_n}}{\log(k_n/(np_n))}\left(\frac{\hat{Q}_\epsilon(\tau_n)}{Q_\epsilon(\tau_n)}-1\right)\xrightarrow{d}N\left(\frac{\lambda}{1-\varrho},\gamma^2\right). \tag{2.21}$$

Theorem 4 can be proved in the same way as that for Theorem 2 in [63]. For the sake of completeness, we present the proof in this section.

Recall that $\hat{Q}_\epsilon(\tau_n)=\left(\frac{k_n}{np_n}\right)^{\hat{\gamma}_n}e_{n-k_n,n}=:d_n^{\hat{\gamma}_n}e_{n-k_n,n}$. First, note that from Theorem 2, we have $\sqrt{k_n}(\hat{\gamma}_n-\gamma)=\Gamma+o_p(1)$, where Γ is a random variable from $N\left(\frac{\lambda}{1-\varrho},\gamma^2\right)$. There-

fore,

$$d_n^{\hat{\gamma}_n-\gamma} = \exp\left((\hat{\gamma}_n-\gamma)\log d_n\right) = \exp\left(\frac{\log d_n}{\sqrt{k_n}}(\Gamma+o_p(1))\right)$$
$$=1+\frac{\log d_n}{\sqrt{k_n}}\Gamma+o_p(\frac{\log d_n}{\sqrt{k_n}}), \tag{2.22}$$

where the last step follows from the assumption that $\frac{\log d_n}{\sqrt{k_n}} \to 0$. Second, by Theorem 2.4.1,

$$\sqrt{k}\left(\frac{Q_\epsilon(U_{n-k_n,n})}{Q_\epsilon(1-k_n/n)}-1\right) \xrightarrow{d} N(0,\gamma^2).$$

In combination with (2.20), we have

$$\frac{e_{n-k_n,n}}{Q_\epsilon(1-k_n/n)} = \frac{e_{n-k_n,n}}{Q_\epsilon(U_{n-k_n,n})}\cdot\frac{Q_\epsilon(U_{n-k_n,n})}{Q_\epsilon(1-k_n/n)} = \left(1+o_p\left(n^{-\delta}\left(\frac{k_n}{n}\right)^\gamma\right)\right)\left(1+O_p\left(\frac{1}{\sqrt{k_n}}\right)\right)$$
$$=1+O_p\left(\frac{1}{\sqrt{k_n}}\right), \tag{2.23}$$

by the assumption that $k_n^{\gamma+1}n^{-\gamma-\delta} \to 0$. Last, by the second order condition given in (2.7) and Theorem 2.3.9 in [17],

$$\frac{Q_\epsilon(1-p_n)}{Q_\epsilon(1-k_n/n)d_n^\gamma} = 1+O(A(n/k_n)) = 1+O\left(\frac{1}{\sqrt{k_n}}\right). \tag{2.24}$$

Finally, combing (2.22), (2.23) and (2.24), we have

$$\frac{\hat{Q}_\epsilon(\tau_n)}{Q_\epsilon(\tau_n)} = \frac{d_n^{\hat{\gamma}}e_{n-k_n,n}}{Q_\epsilon(1-p_n)} = d_n^{\hat{\gamma}_n-\gamma}\frac{e_{n-k_n,n}}{Q_\epsilon(1-k_n/n)}\cdot\frac{Q_\epsilon(1-k_n/n)d_n^\gamma}{Q_\epsilon(1-p_n)}$$
$$=\left(1+\frac{\log d_n}{\sqrt{k_n}}\Gamma+o_p\left(\frac{\log d_n}{\sqrt{k_n}}\right)\right)\left(1+O_p\left(\frac{1}{\sqrt{k_n}}\right)\right)\left(1+O\left(\frac{1}{\sqrt{k_n}}\right)\right)$$
$$=1+\frac{\log d_n}{\sqrt{k_n}}\Gamma+o_p\left(\frac{\log d_n}{\sqrt{k_n}}\right),$$

by the assumption that $d_n \to \infty$. Thus, (2.21) follows immediately.

2.A.4. PROOF OF THEOREM 3

By definition of $Q_{Y|X}(\tau_n|x)$, $\hat{Q}_{Y|X}(\tau_n|x)$ and Theorem 1, we have,

$$\begin{aligned}
&\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)}\left(\frac{\hat{Q}_{Y|X}(\tau_n|x)}{Q_{Y|X}(\tau_n|x)}-1\right)\\
=&\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)Q_\epsilon(\tau_n)}\left(\hat{Q}_{Y|X}(\tau_n|x)-Q_{Y|X}(\tau_n|x)\right)\left(\frac{Q_\epsilon(\tau_n)}{Q_{Y|X}(\tau_n|x)}\right)\\
=&\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)Q_\epsilon(\tau_n)}\left(\hat{Q}_\epsilon(\tau_n)-Q_\epsilon(\tau_n)+\hat{r}_n(x)-r(x)\right)(1+o(1))\\
=&\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)Q_\epsilon(\tau_n)}\left(\hat{Q}_\epsilon(\tau_n)-Q_\epsilon(\tau_n)\right)(1+o(1))+O_p\left(\frac{\sqrt{k_n}n^{-\delta}}{\log\left(\frac{k_n}{np_n}p_n^{-\gamma}\right)}\right).
\end{aligned}$$

Thus it follows from Theorem 4 and the assumption $\frac{\sqrt{k}p_n^\gamma}{n^\delta\log\left(\frac{k_n}{np_n}\right)}\to 0$ that

$$\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{np_n}\right)}\left(\frac{\hat{Q}_{Y|X}(\tau_n|x)}{Q_{Y|X}(\tau_n|x)}-1\right)\xrightarrow{d}N\left(\frac{\lambda}{1-\varrho},\gamma^2\right).$$

3

GRADIENT BOOSTING EXTREMES

Extreme quantile regression provides estimates of conditional quantiles outside the range of the data. Classical methods such as quantile random forests perform poorly in such cases since data in the tail region are too scarce. Extreme value theory motivates to approximate the conditional distribution above a high threshold by a generalized Pareto distribution with covariate dependent parameters. This model allows for extrapolation beyond the range of observed values and estimation of conditional extreme quantiles. We propose a gradient boosting procedure to estimate a conditional generalized Pareto distribution by minimizing its deviance. Cross-validation is used for the choice of tuning parameters such as the number of trees and the tree depths. We discuss diagnostic plots such as variable importance and partial dependence plots, which help to interpret the fitted models. In simulation studies we show that our gradient boosting procedure outperforms classical methods from quantile regression and extreme value theory, especially for high-dimensional predictor spaces and complex parameter response surfaces. An application to statistical post-processing of weather forecasts with precipitation data in the Netherlands is proposed.

3.1. INTRODUCTION

In a regression setup the distribution of a quantitative response Y depends on a set of covariates (or predictors) $\mathbf{X} \in \mathbb{R}^d$. These predictors are typically easily available and can be used to predict conditional properties of the response variable Y. Machine learning offers a continuously growing set of tools to perform prediction tasks based on a sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of independent copies of a random vector $(\mathbf{X}, Y)$. The main objective is usually to predict the conditional mean $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$, which corresponds to minimizing the squared error prediction loss. While the mean summarizes the behavior of Y in the center of its distribution, applications in the field of risk assessment require knowledge of the distributional tail. For a probability level $\tau \in (0,1)$, an important quantity is the

Parts of this chapter have been submitted to Extremes.

conditional quantile

$$Q_{\mathbf{x}}(\tau) = F_Y^{-1}(\tau \mid \mathbf{X} = \mathbf{x}), \tag{3.1}$$

where $F_Y^{-1}(\cdot \mid \mathbf{X} = \mathbf{x})$ is the generalized inverse of the conditional distribution function of Y given $\mathbf{X} = \mathbf{x}$. There has been extensive research in statistics and machine learning to adapt mean prediction methods to other loss functions than squared error. For instance, quantile regression relies on minimizing the conditional quantile loss, which is based on the quantile check function [38]. This has been extended to more flexible regression functions such as the quantile regression forest [45] and the gradient forest [1], which both build on the original random forest [8]. Another popular tree-based method in machine learning is gradient boosting by [22]. This versatile method aims at optimizing an objective function with a recursive procedure akin to gradient descent.

Let n denote the sample size and $\tau = \tau_n$ the quantile level. The existing quantile regression methodology works well in the case of a fixed quantile level, or in the case of a quantile that is only moderately high, that is, $\tau_n \to 1$ and $n(1-\tau_n) \to \infty$ as $n \to \infty$, meaning that there are sufficient observations above the τ_n level. For more extreme quantiles with $n(1-\tau_n) \to 0$, the quantile loss function is no longer useful because observations become scarce at that level and extrapolation beyond the range of observed values is needed. Extreme value theory provides the statistical tools for a sensible extrapolation into the tail of the variable of interest Y. For a large threshold u close to the upper endpoint of the distribution of Y, the distribution of the threshold exceedance $Y - u \mid Y > u$ can be approximated by the generalized Pareto distribution (GPD)

$$H_{\gamma,\sigma}(y) = 1 - \left(1 + \gamma y/\sigma\right)_+^{-1/\gamma}, \quad y \geq 0, \tag{3.2}$$

where for $a \in \mathbb{R}$, $a_+ = \max(0, a)$, and $\gamma \in \mathbb{R}$ and $\sigma > 0$ are the shape and scale parameters, respectively.

There are two main streams in the literature focusing on the estimation of covariate-dependent extreme quantiles. First, [11] and [63] assume a linear form for the conditional quantile function in (3.1) and derive estimators for extreme quantiles and asymptotic properties of the estimators. The second stream first estimates the conditional quantiles at moderately high levels and then uses the GPD to model threshold exceedances with parameters σ and γ in (3.2) depending on the covariates in order to extrapolate to the extreme level. The covariate dependence is either modeled via parametric or semi-parametric structures such as linear models [15, 61] and generalized additive models [10, 66], or via local smoothing methods [e.g., 13, 26, 59]. While linear or additive models are restricted in their modelling flexibility, local smoothing methods on the other hand are known to be sensitive to the curse of dimensionality and work well only for a low-dimensional predictor space. To account for these issues for modern applications with complex data, tree-based methods are attractive due to their modelling flexibility and robustness in higher dimensions. A first attempt to use tree-based models in extreme value theory is the generalized Pareto regression tree by [20], but the model reduces to a single tree and suffers from limited performance.

Our goal is to estimate the extreme conditional quantile $Q_{\mathbf{x}}(\tau)$ in (3.1), where the dimension of covariates d is large and the response surface allows for complex non-linear

effects. To this end, we build a bridge between the predictive power of tree-based ensemble methods from machine learning and the theory of extrapolation from extreme value theory. Following the second stream of research mentioned above, we model the tail of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ using a GPD distribution in (3.2) with covariate-dependent parameters $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$. The main contribution of this chapter is that in order to estimate $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$, we propose a gradient boosting algorithm to optimize the deviance (negative log-likelihood) of the GPD model. In each boosting iteration, these parameters are updated based on an approximation of the deviance gradient by regression trees. The boosting algorithm has several tuning parameters, the most important ones being the number of trees and the tree depth. We show how they can be chosen effectively using cross-validation. The resulting model includes many trees and is flexible enough to account for a complex non-linear response surface. In two numerical experiments we illustrate that, for the task of extremal quantile estimation, our methodology outperforms quantile regression approaches that do not use tail extrapolation [1, 45] and methods from extreme value theory that assume simple forms for $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$ such as generalized additive models [66]. As a result, to the best of our knowledge, our gradient boosting is the first method that reliably estimates extreme quantiles in the case of complex predictor spaces and in the presence of possibly high-dimensional noise variables.

We apply the developed method to forecast the extreme quantiles of daily precipitation in the Netherlands using the output of numerical weather prediction models as covariates. Our diagnostic tools, namely variable importance score and partial dependence plots, are able to identify changes in the tail heaviness of precipitation as seasonality patterns in the shape parameter estimates $\gamma(\mathbf{x})$. We further investigate the contribution of weather prediction model outputs of neighbouring stations to forecasting the extreme precipitation of a specific location.

The chapter is organized as follows. Section 3.2 introduces our methodology and algorithms for extreme quantile regression based on GPD modeling with gradient boosting. Practical questions such as parameter tuning and model interpretation are discussed in Section 3.3, while Section 3.4 is devoted to assessing the performance of our method in two simulation studies. The application to statistical post-processing of weather forecasts with precipitation data in the Netherlands is given in Section 3.5. We conclude the chapter with a summary and discussion section.

The gradient boosting method is implemented in an R package and can be downloaded from GitHub at `https://github.com/JVelthoen/gbex/`

3.2. EXTREME QUANTILE REGRESSION WITH GRADIENT BOOSTING

3.2.1. BACKGROUND ON EXTREME QUANTILE ESTIMATION

Extreme value theory provides the asymptotic results for extrapolating beyond the range of the data and statistical methodology has been developed to accurately estimate high quantiles. Most of these tools however do only apply in the case where $Y_1, \dots, Y_n$ are independent copies of Y and do not dependent on covariates.

In this case, the Pickands–de Haan–Balkema theorem [2, 47] states that under mild

regularity conditions on the tail of the distribution of Y, the rescaled distribution of exceedances over a high threshold converges to the generalized Pareto distribution. More precisely, if y^* denotes the upper endpoint of the distribution of Y then there exist a normalizing function $\sigma(u) > 0$ such that

$$\lim_{u \uparrow y^*} \mathbb{P}\left(\frac{Y-u}{\sigma(u)} > y \mid Y > u \right) = 1 - H_{\gamma,1}(y), \quad y \geq 0, \tag{3.3}$$

where H is defined in (3.2), with the convention $H_{0,\sigma}(y) = 1 - \exp(-y/\sigma)$, $y \geq 0$. The shape parameter $\gamma \in \mathbb{R}$ indicates the heaviness of the upper tail of Y, where $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$ correspond to distributions respectively with finite upper endpoint (e.g., uniform), light tails (e.g., Gaussian, exponential) and power tails (e.g., Student's t).

Moreover, the GPD is the only non-degenerate distribution that can arise as the limit of threshold exceedances as in (3.3), and therefore it is an asymptotically motivated model for tail extrapolation and high quantile estimation. By the limit relation in (3.3), for a large threshold u, the conditional distribution of $Y - u$ given $Y > u$ can be approximated by $H_{\gamma,\sigma}$ with $\sigma = \sigma(u)$. The threshold u can be chosen as the quantile $Q(\tau_0)$ of Y for some moderately high probability level $\tau_0 \in (0,1)$. Inverting the distribution function in (3.2) provides an approximation of the quantile for probability level $\tau > \tau_0$ by

$$Q(\tau) \approx Q(\tau_0) + \sigma \frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\gamma} - 1}{\gamma}. \tag{3.4}$$

3.2.2. SETUP FOR EXTREME QUANTILE REGRESSION

We consider here the setting where the response $Y_i \in \mathbb{R}$ depends on covariates $\mathbf{X}_i \in \mathbb{R}^d$ and our goal is to develop an estimator for the conditional quantile $Q_{\mathbf{x}}(\tau_n)$ defined by (3.1), where the probability level τ_n satisfies $\tau_n \to 1$ and $n(1-\tau_n) \to 0$ as $n \to \infty$. Such a quantile is extreme in the sense that the expected number of observations that exceed $Q_{\mathbf{x}}(\tau_n)$ converges to 0 as $n \to \infty$. Therefore empirical estimation is not feasible and extrapolation beyond observations is needed. Recall that $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote independent copies of the random vector $(\mathbf{X}, Y)$ with $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

In this setup, the intermediate threshold $Q(\tau_0)$, shape parameter γ and scale parameter σ in (3.4) may depend on covariates, and the extreme value approximation for the extreme conditional quantile becomes

$$Q_{\mathbf{x}}(\tau) \approx Q_{\mathbf{x}}(\tau_0) + \sigma(\mathbf{x}) \frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\gamma(\mathbf{x})} - 1}{\gamma(\mathbf{x})}, \quad \tau > \tau_0. \tag{3.5}$$

The triple $(Q_{\mathbf{x}}(\tau_0), \sigma(\mathbf{x}), \gamma(\mathbf{x}))$ provides a model for the tail (that is above the probability level τ_0) of the conditional law of Y given $\mathbf{X} = \mathbf{x}$. An estimator of conditional extreme quantiles $\hat{Q}_{\mathbf{x}}(\tau)$ is obtained by plugging in estimators $(\hat{Q}_{\mathbf{x}}(\tau_0), \hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x}))$ in Equation (3.5).

In the following we propose estimators for these three quantities. Our main contribution is a gradient boosting procedure for estimation of the GPD parameters $(\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ that allows flexible regression functions with possibly many covariates. For estimation

of the intermediate quantile $Q_{\mathbf{x}}(\tau_0)$, any method for (non-extreme) quantile regression can be used and we outline in Section 3.2.4 how the existing method of quantile random forests can be applied.

3.2.3. GPD MODELING WITH GRADIENT BOOSTING

Based on the asymptotic result in (3.3), the peaks-over-threshold approach assumes that, given $\mathbf{X} = \mathbf{x}$, the excess of Y above the threshold $Q_{\mathbf{x}}(\tau_0)$ follows approximately a GPD. In order to compute the sample of exceedances, we rely on a (non-extreme) quantile regression method providing an estimation $\hat{Q}_{\mathbf{x}}(\tau_0)$ of the intermediate quantile. We then define the exceedances of the data set $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ above threshold as

$$Z_i = \left(Y_i - \hat{Q}_{\mathbf{X}_i}(\tau_0)\right)_+, \quad i = 1, \ldots, n, \tag{3.6}$$

so that $Z_i = 0$ whenever the value Y_i is below threshold. We assume that the intermediate threshold is high enough so that the exceedances can be modeled by the generalized Pareto distribution and the approximation of conditional quantiles (3.5) is good. Our aim is to learn the conditional parameter $\theta(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ based on the sample of exceedances above the threshold. Following [22] and [23], we propose to use gradient boosting with a suitable objective function.

In absence of covariates, a standard way of estimating the GPD parameters $\theta = (\sigma, \gamma)$ is the maximum likelihood method, which provides asymptotically normal estimators in the unconditional case with $\gamma > -1/2$ [52]. Likewise, we use the negative log-likelihood, or equivalently the deviance of GPD distribution as the objective function of the gradient boosting procedure. Precisely, the deviance for an exceedance Z_i from a GPD distribution with parameters $\theta(\mathbf{X}_i) = (\sigma(\mathbf{X}_i), \gamma(\mathbf{X}_i))$ is given by

$$\ell_{Z_i}(\theta(\mathbf{X}_i)) = \left[(1 + 1/\gamma(\mathbf{X}_i)) \log\left(1 + \gamma(\mathbf{X}_i)\frac{Z_i}{\sigma(\mathbf{X}_i)}\right) + \log\sigma(\mathbf{X}_i)\right] \mathbb{I}_{Z_i > 0}. \tag{3.7}$$

Since the deviance depends on two parameters γ and σ, we need to build two sequences of trees, one for each parameter; [22] proposes a similar strategy in multiclass classification where several sequences of trees are trained to learn the different class probabilities.

The gradient boosting algorithm starts with an initial estimate, which is given by the unconditional maximum likelihood estimator, that is, by setting $\theta(\mathbf{X}_i) \equiv \theta$ in (3.7):

$$\theta_0(\mathbf{x}) \equiv \arg\min_{\theta} \sum_{i=1}^{n} \ell_{Z_i}(\theta). \tag{3.8}$$

The two sequences of gradient trees $(T_b^\sigma)_{1 \le b \le B}$ and $(T_b^\gamma)_{1 \le b \le B}$ are built recursively. We use the superscript $\delta \in \{\sigma, \gamma\}$ and the convenient notation $\theta = (\theta^\sigma, \theta^\gamma)$ to treat the two sequences simultaneously. Let $s \in (0, 1]$ denote a subsampling fraction. Sequentially for $b = 1, \ldots, B$, we draw a random subset $S_b \subset \{1, \ldots, n\}$ of size $\lfloor sn \rfloor$ and fit a pair of regression trees (T_b^σ, T_b^γ) to learn the gradient of the deviance on subsample S_b given by

$$r_{b,i}^\delta = \frac{\partial \ell_{Z_i}}{\partial \theta^\delta}(\theta_{b-1}(\mathbf{X}_i)), \quad i \in S_b, \delta \in \{\sigma, \gamma\}.$$

The regression tree T_b^δ is fitted on the sample $(\mathbf{X}_i, r_{b,i}^\delta)$, $i \in S_b$, by recursive binary splitting. Two further parameters are used to build the tree: the maximal depth D^δ gives the maximum number of splits between the root and a leaf in the tree; the minimal leaf size gives the minimum number of observations in each leaf. The leaves of the tree T_b^δ are denoted by $L_{b,j}^\delta$, $j = 1, \dots, J_b^\delta$, with J_b^δ the number of leaves.

Now that the tree T_b^δ is built, we need to update the value of parameter δ for each leaf such that the deviance is minimized. In theory this can be done by line search, that is, for leaf $L_{b,j}^\delta$, the updated value $\xi_{b,j}^\delta$ is obtained by minimizing the deviance, i.e.,

$$\xi_{b,j}^\delta = \arg\min_{\xi} \sum_{\mathbf{X}_i \in L_{b,j}^\delta} \ell_{Z_i}(\theta_{b-1}(\mathbf{X}_i) + \xi e_\delta), \quad j = 1, \dots, J_b^\delta, \tag{3.9}$$

where $e_\sigma = (1,0)$ and $e_\gamma = (0,1)$ give the directions of the line search corresponding to parameters σ and γ, respectively. In practice the line search (3.9) can be computationally expensive and $\xi_{b,j}^\delta$ is approximated by a Newton–Raphson step

$$\tilde{\xi}_{b,j}^\delta = -\frac{\sum_{\mathbf{X}_i \in L_{b,j}^\delta} \frac{\partial \ell_{Z_i}}{\partial \theta^\delta}(\theta_{b-1}(\mathbf{X}_i))}{\sum_{\mathbf{X}_i \in L_{b,j}^\delta} \frac{\partial^2 \ell_{Z_i}}{\partial (\theta^\delta)^2}(\theta_{b-1}(\mathbf{X}_i))}.$$

The derivatives of the deviance are provided in Appendix 3.A. Due to the instability of the derivatives of the GPD likelihood, we bound the absolute value of the Newton–Raphson step by 1 in order to mitigate the strong influence of extreme observations. We observe in practice that this results in better performance. This leads to the value

$$T_b^\delta(\mathbf{x}) = \sum_{j=1}^{J_b^\delta} \operatorname{sign}(\tilde{\xi}_{b,j}^\delta) \min(|\tilde{\xi}_{b,j}^\delta|, 1) \mathbb{I}_{\{\mathbf{x} \in L_{b,j}^\delta\}}. \tag{3.10}$$

for the gradient tree. The model $\theta_{b-1}(\mathbf{x})$ is then updated by

$$\theta_b^\delta(\mathbf{x}) = \theta_{b-1}^\delta(\mathbf{x}) + \lambda^\delta T_b^\delta(\mathbf{x}), \quad \delta \in \{\sigma, \gamma\}, \tag{3.11}$$

where the shrinkage parameters $\lambda^\sigma, \lambda^\gamma \in (0,1)$ are called learning rates. They are used to slow down the dynamic since a shrunken version of the trees is added to the current model.

The final output for the estimated parameters is the gradient boosting model

$$\hat{\theta}^\delta(\mathbf{x}) = \theta_0^\delta + \lambda^\delta \sum_{b=1}^{B} T_b^\delta(\mathbf{x}), \quad \delta \in \{\sigma, \gamma\}. \tag{3.12}$$

Algorithm 1 summarizes the procedure for GPD modeling of exceedances. In practice, the number of iterations B is an important parameter and its choice corresponds to a trade-off between bias and variance. The procedure is prone to overfitting as $B \to \infty$ and cross-validation is used to prevent this by early stopping; see Section 3.3.1 where

we discuss the interpretation of the different tuning parameters and their selection in practice.

Algorithm 1: gbex boosting algorithm for GPD modeling

Input:

- θ_0: the initial values of the parameters with default value as in (3.8);
- $(\mathbf{X}_i, Z_i)_{1\le i\le n}$: data sample of exceedances above threshold;
- B: number of gradient trees;
- D^σ, D^γ: maximum tree depth for the gradient trees;
- $\lambda^\sigma, \lambda^\gamma$: learning rates for the update of the GPD parameters σ and γ respectively;
- s: subsampling fraction;
- L^σ_{min}, L^γ_{min}: minimum leaf size of the nodes in the trees.

Algorithm: For $b = 1, \dots, B$:

1. Draw a random subsample $S_b \subset \{1, \dots, n\}$ of size $\lfloor sn \rfloor$.
2. Compute the deviance derivatives on the subsample S_b:
$$r^\sigma_{b,i} = \frac{\partial \ell_{Z_i}}{\partial \sigma}(\theta_{b-1}(\mathbf{X}_i)) \quad \text{and} \quad r^\gamma_{b,i} = \frac{\partial \ell_{Z_i}}{\partial \gamma}(\theta_{b-1}(\mathbf{X}_i)), \quad i \in S_b.$$
3. Fit regression trees T^σ_b, T^γ_b that predict the gradients $r^\sigma_{b,i}$ and $r^\gamma_{b,i}$ as functions of the covariates $\mathbf{X}_i$ on the sample $i \in S_b$; the trees are built with maximal depth (D^σ, D^γ) and minimal leaf size $(L^\sigma_{min}, L^\gamma_{min})$; for the tree values, use the truncated Newton–Raphson rule (3.10).
4. Update the GPD parameters $\theta_b(\mathbf{x}) = (\hat\sigma_b(\mathbf{x}), \hat\gamma_b(\mathbf{x}))$ with learning rates $(\lambda^\sigma, \lambda^\gamma)$, i.e.,
$$\hat\sigma_b(\mathbf{x}) = \hat\sigma_{b-1}(\mathbf{x}) + \lambda^\sigma T^\sigma_b(\mathbf{x}) \quad \text{and} \quad \hat\gamma_b(\mathbf{x}) = \hat\gamma_{b-1}(\mathbf{x}) + \lambda^\gamma T^\gamma_b(\mathbf{x}).$$

Output: Conditional GPD parameters $(\hat\sigma(\mathbf{x}), \hat\gamma(\mathbf{x})) = (\hat\sigma_B(\mathbf{x}), \hat\gamma_B(\mathbf{x}))$.

3.2.4. EXTREME QUANTILE REGRESSION

The input of Algorithm 1 are the exceedances Z_i defined (3.6). The conditional intermediate quantile $\hat Q_{\mathbf{X}_i}(\tau_0)$ used in this definition generally also depends on the covariate vector $\mathbf{X}_i$ and needs to be modeled first. For this task, any method for (non-extreme) quantile regression can be used, but we note that the quality of the approxmiation (3.5) of the extreme quantile will also depend on the accuracy of the intermediate quantile estimate. Together with the gradient boosting procedure for the GPD parameters in Section 3.2.3, we obtain an algorithm for extreme quantile prediction. We refer to this algorithm as the `gbex` method. It combines the flexibility of gradient boosting with the extrapolation technique from extreme value theory.

While in principle any quantile regression method can be used for estimation of the conditional intermediate quantiles $\hat Q_{\mathbf{X}_i}(\tau_0)$, we propose to use a quantile random forest. The reason for this is three-fold: first it requires no parametric assumptions on the quantile functions; secondly it exhibits good performance for high dimensional predic-

Algorithm 2: gbex algorithm for extreme quantile prediction

Input:

- $(\mathbf{X}_i, Y_i)_{1\leq i\leq n}$: data sample;
- τ_0: probability level for the threshold;
- τ: probability level for the prediction such that $\tau > \tau_0$;
- parameters of the gbex boosting algorithm for GPD modeling of exceedances (Algorithm 1).

Algorithm:

1. Fit a quantile regression to the sample $(\mathbf{X}_i, Y_i)_{1\leq i\leq n}$ that provides estimates $\hat{Q}_{\mathbf{X}}(\tau_0)$ of the conditional quantiles of order τ_0.
2. Compute the exceedances $Z_i = (Y_i - \hat{Q}_{\mathbf{X}_i}(\tau_0))_+$, $1 \leq i \leq n$.
3. Let $I = \{i : Z_i > 0\}$ be the index set of positive exceedances and run Algorithm 1 on the data set $(\mathbf{X}_i, Z_i)_{i\in I}$ to estimate the GPD parameters $(\hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x}))$.

Output: Estimation of the extreme conditional quantile

$$\hat{Q}_{\mathbf{X}}(\tau) = \hat{Q}_{\mathbf{X}}(\tau_0) + \hat{\sigma}(\mathbf{x})\frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\hat{\gamma}(\mathbf{x})} - 1}{\hat{\gamma}(\mathbf{x})}.$$

tor spaces; finally it requires minimal tuning for good results. Quantile regression forests were first proposed by [45] using the weights from a standard random forest [8]. The drawback of this method is that the criterion used in recursive binary splitting to build the trees of the random forest is not tailored to quantile regression. [60] therefore define a generalized random forest with splitting rule designed for that specific task, where the splitting criterion is related to the quantile loss function. In our case, we require the estimator of $Q_{\mathbf{x}}(\tau_0)$ at the sample points $\mathbf{x} \in \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and we recommend the use of out-of-bag estimation $\hat{Q}_{\mathbf{X}_i}(\tau_0) = \hat{Q}^{oob}_{\mathbf{X}_i}(\tau_0)$. This means that only the trees for which the ith observation is out-of-bag are kept for the quantile estimation at $\mathbf{x} = \mathbf{X}_i$, that is, trees based on sub-samples not containing the ith observation. This is necessary to avoid giving too much weight to the ith observation when predicting at $\mathbf{x} = \mathbf{X}_i$.

3.3. PARAMETER TUNING AND INTERPRETATION

3.3.1. PARAMETER TUNING

Our gradient boosting procedure for GPD modelling includes several parameters that need to be tuned properly for good results. We discuss in this section the interpretation of the different parameters and how to choose them. We introduce data driven choices based on cross validation for the most sensitive parameters and suggest sensible default values for the remaining parameters.

TREE NUMBER B

The number of trees is the most important regularization parameter. The boosting procedure starts from a constant model, that is usually an underfit, and adds recursively trees that adapt the model to the data, leading eventually to an overfit.

We recommend repeated K-fold cross-validation based on the deviance for a data driven choice of B. Given a maximal tree number B_{max} and a division of the data set into K folds $\mathscr{D}_1, \ldots, \mathscr{D}_K$, we repeatedly run the algorithm with B_{max} iterations on the data with one fold left-out and then compute the deviance on the left-out fold as a function of B. Adding up the deviances for the different folds, we obtain the cross-validation deviance. More formally, we define

$$\text{DEV}_{CV}(B) = \sum_{k=1}^{K} \sum_{i \in \mathscr{D}_k} \ell_{Z_i}(\hat{\theta}_B^{-\mathscr{D}_k}(\mathbf{X}_i)), \quad B = 0, \ldots, B_{max}, \tag{3.13}$$

where $\hat{\theta}_B^{-\mathscr{D}_k}$ denotes the model with B trees trained on the data sample with the kth fold $\mathscr{D}_k$ held out. Due to large values of the deviance on extreme observations, the cross-validation deviance is prone to fluctuations with respect to the partition into folds and we therefore recommend repeated cross-validation. A typical choice is $K = 5$ or 10 with 5 repetitions. The choice of B is then the minimizer of the cross-validation deviance.

TREE DEPTH (D^σ, D^γ)

The gradient boosting algorithm outputs a sum of tree functions. The complexity of the model is therefore determined by the depth parameters D^σ and D^γ, also called interaction depths [see 34, Section 10.11]. A zero depth tree corresponds to a constant tree with no split, so that $D^\sigma = 0$ or $D^\gamma = 0$ yield models with constant scale or shape parameters, respectively. Since the extreme value index γ is notoriously difficult to estimate, it is common in extreme value theory to assume a constant value $\gamma(\mathbf{x}) \equiv \gamma$ so that the case $D^\gamma = 0$ is particularly important. A tree with depth 1, also called a stump, makes only one single split on a single variable. As a result, $D^\sigma = 1$ (resp. $D^\gamma = 1$) corresponds to an additive model in the predictors for $\sigma(\mathbf{x})$ (resp. $\gamma(\mathbf{x})$). Trees with larger depth allow to introduce interaction effects between the predictors of order equal to the depth parameter. In practice, the depth parameter is quite hard to tune and we recommend to consider depth no larger than 3, also because interactions of higher order are difficult to interpret. Based on our experience, sensible default values are $D^\sigma = 2$ and $D^\gamma = 1$. But more interestingly, cross-validation can be used to select the depth parameters. The left panel of Figure 3.1 shows a typical cross-validation diagnostic in the context of the simulation study detailed in Section 3.4. Here $B_{max} = 500$ and depths parameter $(D^\sigma, D^\gamma) = (1,0)$, $(1,1)$, $(2,1)$ and $(2,2)$ are considered. The plot shows that sensible choices are $B \approx 200$ and $(D^\sigma, D^\gamma) = (1,0)$ or $(1,1)$ (more details given in Section 3.4). The histogram in the right panel shows that, depending on the randomly simulated sample, B typically lies in the range $[100, 250]$, where the deviance is relatively flat ($(D^\sigma, D^\gamma) = (1,0)$ is fixed here).

LEARNING RATES $(\lambda^\sigma, \lambda^\gamma)$

As usual in gradient boosting, there is a balance between the learning rate and the number of trees. As noted in [50], multiplying the learning rate by 0.1 roughly requires 10

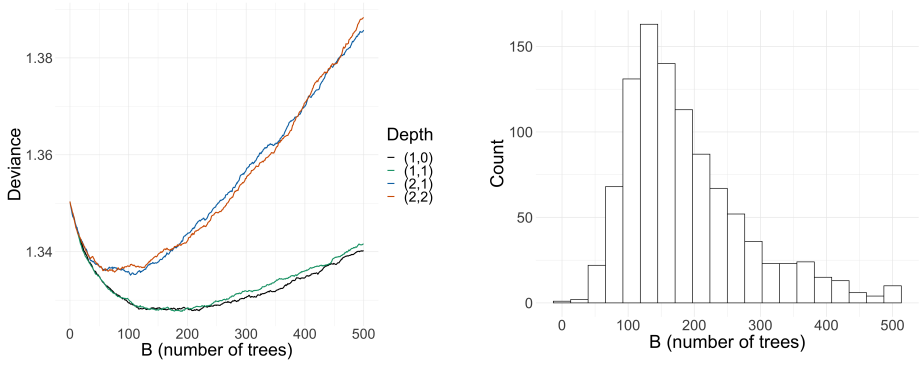


Figure 3.1: Left panel: cross-validation deviance given by (3.13) against B for one random sample and depth $(D^\sigma, D^\gamma) = (1,0), (1,1), (2,1)$ and $(2,2)$. Right panel: selected values of B for 1000 samples when $(D^\sigma, D^\gamma) = (1,0)$ is fixed. The design of the simulation study is Model 1 described in Section 3.4.

times more trees for a similar result. It is common to fix the learning rate to a small value, typically 0.01 or 0.001, and to consider the tree number as the main parameter. Since in our case we have two parallel gradient boosting procedures with different learning rates, we reparameterize them as $(\lambda_{scale}, \lambda_{ratio}) = (\lambda^\sigma, \lambda^\sigma/\lambda^\gamma)$. The balance described above is expressed between B and λ_{scale} and we propose the default $\lambda_{scale} = 0.01$, leaving the number of trees B as the primary parameter. The ratio of the learning rates is important as γ generally requires stronger regularization than σ and ranges on smaller scales. Therefore it is natural to choose $\lambda_{ratio} > 1$. Often a sensible default for λ_{ratio} falls in between 5 and 10.

REMAINING TUNING PARAMETERS

The minimum leaf sizes $L^\sigma_{min}, L^\gamma_{min}$ and subsample fraction s play the role of regularization parameters. The minimum leaf size makes sure that the splits do not try to isolate a single high observation of the gradient and that the leaves contain enough observations so that averaging provides a smoother gradient. Subsampling ensures that different trees are fitted on different sub-samples, mitigating the correlation between trees; see [23] and Section 10.12.2 of [34] for further discussion on the regularization effect of subsampling. In practice, we do not recommend to optimize these parameters but rather to use the sensible default parameters $L^\sigma_{min} = L^\gamma_{min} = \max(10, n/100)$ and $s = 75\%$.

The parameter τ_0 stands for the probability level of the intermediate quantile used as threshold. Threshold selection is a long standing problem in extreme value theory [e.g., 18, 19]. A higher threshold yields a better approximation by the GPD distribution but fewer exceedances, leading to reduced bias and higher variance. Some guidelines for threshold selection in practice are provided in Section 3.5, where we present an application to precipitation forecast statistical post-processing.

3.3.2. TOOLS FOR MODEL INTERPRETATION

Contrary to a single tree, boosting models that aggregate hundreds or thousands of trees are difficult to represent but diagnostic plots are available to ease the interpretation. We

briefly discuss variable importance and partial dependence plots, which are straightforward modifications to our framework of the tools detailed in Section 10.13 of [34].

VARIABLE IMPORTANCE

Boosting is quite robust to the curse of dimensionality and often provides good results even in the presence of high dimensional predictors and noise variables. Understanding which predictors are the most important is crucial for model interpretation. Variable importance is used for this purpose and we discuss here the permutation score and the relative importance.

The permutation score helps to evaluate the impact of a predictor on the model deviance and is not specific to boosting. The relation between a predictor and the response is disturbed by shuffling the values of this predictor and measuring the difference in the deviance before and after shuffling. More precisely, for predictor X_j, we define

$$I(X_j) = \sum_{i=1}^{n} \ell_{Z_i}\left(\hat{\theta}\left(\mathbf{X}_i^{(j)}\right)\right) - \sum_{i=1}^{n} \ell_{Z_i}\left(\hat{\theta}\left(\mathbf{X}_i\right)\right), \tag{3.14}$$

where $\hat{\theta}$ is the estimator given in (3.12) and $\mathbf{X}_1^{(j)}, \ldots, \mathbf{X}_n^{(j)}$ denote the same input vectors as $\mathbf{X}_1, \ldots, \mathbf{X}_n$ except that the jth components are randomly shuffled. A large permutation score $I(X_j)$ indicates a strong effect of X_j in the boosting model. Since the scores are relative, it is customary to assign to the largest the value of 100 and scale the others accordingly.

The relative importance is specific to tree based methods such as boosting or random forests and uses the structure of the trees in the model. It is discussed for instance in Section 10.13.1 of [34]. Recall that during the construction of the trees, the splits are performed so as to minimize the residual sum of squares (RSS) of the gradient and each split causes a decrease in the RSS. The more informative splits are those causing a large decrease in the RSS. The relative importance of a given variable X_j is obtained by considering all the splits due to this variable in the sequence of trees, and by summing up the decrease in RSS due to those splits. Because we have two sequences of trees, we compute relative importance of variable X_j in the estimation of σ and γ separately by considering the sequence of trees (T_b^{σ}) and (T_b^{γ}) respectively.

PARTIAL DEPENDENCE PLOT

Once the most relevant variables have been identified, the next attempt is to understand the dependence between the predictors and the response. Partial dependence plots offer a nice graphical diagnostic of the partial influence of a predictor X_j on the outputs $\hat{\sigma}(\mathbf{x})$, $\hat{\gamma}(\mathbf{x})$ or $\hat{Q}_{\mathbf{x}}(\tau)$; see Section 10.13.2 of [34]. The partial dependence plot for $\hat{\sigma}$ with respect to X_j is the graph of the function $x \mapsto \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}(\mathbf{X}_i^{-j,x})$, where the vector $\mathbf{X}_i^{-j,x}$ is equal to $\mathbf{X}_i$ except that the jth component has been replaced by x. Notice that dependence between the predictors is not taken into account so that this is not an estimate of $\mathbb{E}[\hat{\sigma}(\mathbf{X}) \mid X_j = x]$, except if X_j is independent of the other predictors. In the particular case when an additive model is built, i.e., $D^{\sigma} = 1$, the partial dependence plot with respect to X_j is equal to the effect of the variable X_j up to an additive constant. Partial dependence plots with respect to several covariates can be defined and plotted similarly, at least in dimension 2 or 3.

3.4. SIMULATION STUDIES

To demonstrate the performance of our method, we conduct two numerical experiments. We generate n independent samples with d covariates $\mathbf{X} = (X_1, \ldots, X_d)$ distributed from an independent uniform distribution on $[-1,1]^d$, with $(n,d) = (2000,40)$ or $(5000,10)$, depending on the complexity of the model. We aim to estimate the conditional quantile function $Q_{\mathbf{X}}(\tau)$ corresponding to extreme probability levels $\tau \in \{0.99, 0.995, 0.9995\}$. We choose the level $\tau_0 = 0.8$ for the intermediate quantile and it is worthwhile to note that the effective sample size $n(1-\tau_0)$ for the gradient boosting step is then only 400 for $n = 2000$.

The local smoothing based methods mentioned in the introduction [13, 26] become cumbersome in our simulation setting because of the sparsity of data in high dimension. We compare our `gbex` method to two quantile regression approaches, the quantile regression forest (`qrf`) from [45] and the generalized random forest (`grf`) from [1]. Moreover, we consider two existing methods from extreme value theory that use GPD modeling of the exceedances. One is the classical estimator of extreme quantile without using covariates, thus $\gamma(\mathbf{x}) \equiv \gamma$ and $\sigma(\mathbf{x}) \equiv \sigma$, which we call the `constant` method. The other one is the `evgam` method of [66] that assumes generalized additive models for $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$.

To evaluate the performance over the full predictor domain $[-1,1]^d$ we consider the integrated squared error (ISE) defined for a fixed quantile level τ and the ith replication of the data set by

$$\mathrm{ISE}_i = \int_{[-1,1]^d} \left(\hat{Q}_{\mathbf{X}}^{(i)}(\tau) - Q_{\mathbf{X}}(\tau) \right)^2 d\mathbf{x}, \tag{3.15}$$

where $\hat{Q}_{\mathbf{X}}^{(i)}(\tau)$ is the quantile estimated from the model. We use a Halton sequence, a low discrepancy quasi-random sequence [e.g., 46, p. 29], in order to efficiently evaluate the high dimensional integral in the ISE computation. Averaging over the $R = 1000$ replications, we obtain the mean integrated squared error (MISE).

Our first model is designed to check robustness of the methods against noise variables. This model is constructed in a similar way as the example studied in Section 5 of [1] and it has a predictor dimension of $d = 40$, of which one covariate is signal and the remaining are noise variables.

- **Model 1:** Given $\mathbf{X} = \mathbf{x} \in \mathbb{R}^{40}$, Y follows a Student's t-distribution with 4 degrees of freedom and scale

$$\mathrm{scale}(\mathbf{x}) = 1 + \mathbb{I}(x_1 > 0).$$

  This is a heavy tailed model where the GPD approximation has a constant shape parameter $\gamma(\mathbf{x}) \equiv 1/4$ and the scale parameter is a step function in X_1. More precisely, $\sigma(\mathbf{x}) = \sigma(\tau_0)(1+\mathbb{I}(x_1 > 0))$ where $\sigma(\tau_0)$ is a multiplicative constant depending on the threshold parameter τ_0.

In our second model, we consider a more complex response surface where both the scale and shape parameters depend on the covariates and interactions of order 2 are introduced.

- **Model 2:** Given $\mathbf{X} = \mathbf{x} \in \mathbb{R}^{10}$, Y follows a Student's t-distribution with degree of freedom $\mathrm{df}(\mathbf{x})$ depending on x_1 through

$$\mathrm{df}(\mathbf{x}) = 7\left(1 + \exp(4x_1 + 1.2)\right)^{-1} + 3,$$

and scale parameter $\mathrm{scale}(\mathbf{x})$ depending on (x_1, x_2) through

$$\mathrm{scale}(\mathbf{x}) = 1 + 6\varphi(x_1, x_2),$$

where φ denotes the density function of a bivariate normal distribution with standard normal margins and correlation 0.9. The numerical constants are chosen so that the GPD approximation of Y given $\mathbf{X} = \mathbf{x}$ has parameters $\gamma(\mathbf{x}) = 1/\mathrm{df}(\mathbf{x})$ in the range $[0.10, 0.33]$ for $\mathbf{x} \in [-1,1]^d$, $d = 10$.

3.4.1. TUNING PARAMETERS AND CROSS VALIDATION

We generate samples of size $n = 2000$ and 5000, respectively from Model 1 and Model 2. We set the following tuning parameters for `gbex`: the learning rate $\lambda_{scale} = 0.01$ and the sample fraction $s = 75\%$ for both models; $\lambda_{ratio} = 15$ for Model 1 and $\lambda_{ratio} = 7$ for Model 2.

As discussed in Section 3.3.1, the number of trees B is the most important regularization parameter and the depth parameters (D^σ, D^γ) determine the complexity of the fitted model. Therefore, we investigate how these tuning parameters influence the performance of our estimator in terms of MISE. Figure 3.2 shows the results for Model 1 (left panel) and for Model 2 (right panel). The curves represent the MISE of gbex as a function of B for various depth parameters (D^σ, D^γ). The right panel clearly shows that for Model 2 the choice $(D^\sigma, D^\gamma) = (1,1)$ does not account for the model complexity adequately, which leads to a high MISE. Indeed, boosting with depth one tries to fit an additive model but the scale parameter of Model 2 depends on (X_1, X_2) in a non-additive way. On the other hand, for Model 1, which is an additive model with the optimal depth $(D^\sigma, D^\gamma) = (1,0)$, the curves suggest that assuming unnecessary complexity of the model might lead to suboptimal behavior of the estimator: the choice $(2,1)$ yields higher MISE than the other two choices and the MISE stays low for a shorter range of B. In general, higher depths help the model to adapt the data faster but then overfitting is prone to occur more rapidly when B increases. The horizontal dashed lines in Figure 3.2 represent the resulting MISE of our estimator when B is chosen via cross-validation with deviance loss given in (3.13), with $K = 5$ folds and 10 replications. The plots confirm that the data driven choice of B results in near optimal MISE for fixed depth parameters (with dashed horizontal lines close to the minimum of the curve with the same color). We additionally apply cross-validation to select both B and (D^σ, D^γ) simultaneously. The resulting MISE is represented by the black dashed line, which is very close to the minimum of all the dashed lines. Overall, the results confirm the good performance of the proposed cross-validation procedure.

For the rest of the simulation study, we set $(D^\sigma, D^\gamma) = (1,1)$ for Model 1 and $(D^\sigma, D^\gamma) = (3,1)$ for Model 2 and choose B with cross-validation.

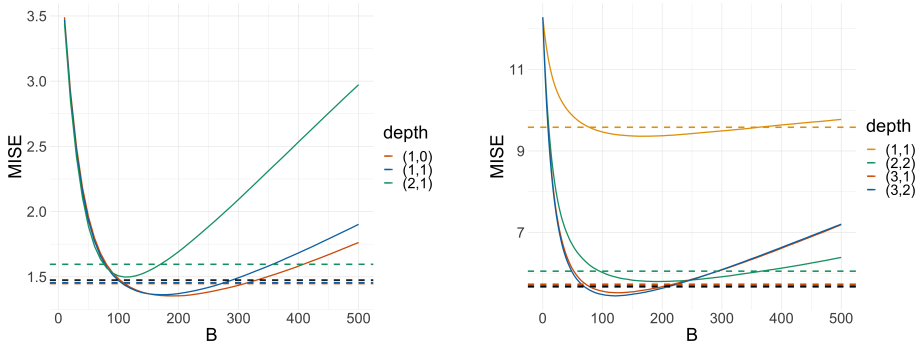


Figure 3.2: The MISE for Model 1 (left panel) and Model 2 (right panel) of the `gbex` extreme quantile estimator with probability level $\tau = 0.995$ as a function of B for various depth parameters (curves); the MISE of the `gbex` estimator with adaptive choice of B for various depth parameters (horizontal dashed lines); the MISE of the `gbex` estimator with both tree number and depth parameters selected by cross-validation (black dashed line).

3.4.2. COMPARISON WITH DIFFERENT METHODS

The comparison of our `gbex` method to the other three approaches `qrf`, `grf` and `constant`, is presented in Figure 3.3. The results for Model 1 and Model 2 are given in the first and second row, respectively. For the probability level $\tau = 0.99$, 0.99 and 0.9995 in the left, middle and right column, the figure shows the boxplots of ISE defined in (3.15) and the MISE represented by the vertical black line. The MISE grows as the probability level increases for all methods, however `gbex` clearly outperforms the other three approaches with a much smaller MISE and a much lower variation of ISE. When the probability level τ is close to or larger than $(1 - 1/n)$ (right column), both `grf` and `qrf` lead to extremely large ISE outliers so that the ISE mean is larger than the third quartile (black line outside the box). Some extreme outliers of ISE are left out of the boxplots to have a clear comparison.

Because Model 2 does not satisfy the additive model assumption of `evgam`, the comparison between `evgam` and `gbex` is based on data generated from Model 1 only. Figure 3.4 presents the ISE and MISE of `evgam` and `gbex` for probability level $\tau = 0.995$. To have a fair comparison, we use the same forest based estimation of quantile $\hat{Q}_{\mathbf{x}}(\tau_0)$ for the intermediate threshold. Because of the computational burden and numerical instability of `evgam` in high dimensions, we have restricted the dimension to $d \leq 10$ for this method, while `gbex` is still considered with $d = 40$. For `evgam`, the boxplots show a steady increase of the MISE with respect to the dimension and we can see that the MISE of `gbex` with $d = 40$ is similar to the MISE of `evgam` with $d = 4$. This clearly demonstrates the robustness of `gbex` against the curse of dimensionality and noise variables, which is a prominent advantage of tree based methods.

3.4.3. DIAGNOSTIC PLOTS

We finally look at the model interpretation diagnostics. Figure 3.5 shows the permutation importance scores defined in (3.14) for both models, based on 1000 replications. The boxplots clearly show that this score is able to identify the signal variable(s). Note

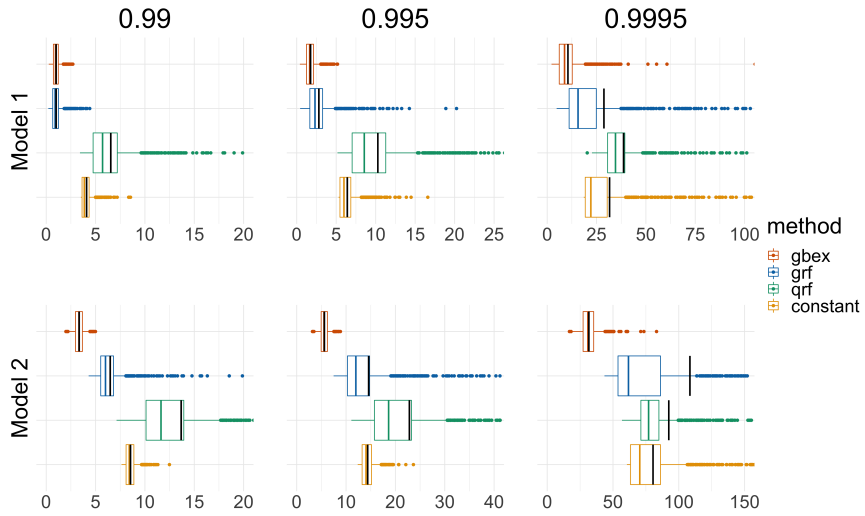


Figure 3.3: Boxplot of ISE based on 1000 replications for the four quantile estimators (`gbex`, `grf`, `qrf` and `constant`) at different probability levels $\tau = 0.99$ (left), 0.995 (middle) and 0.9995 (right) for Model 1 (top) and Model 2 (bottom). Some outliers of `grf` and `qrf` are left out for a clearer comparison. The black vertical lines indicate the MISE.

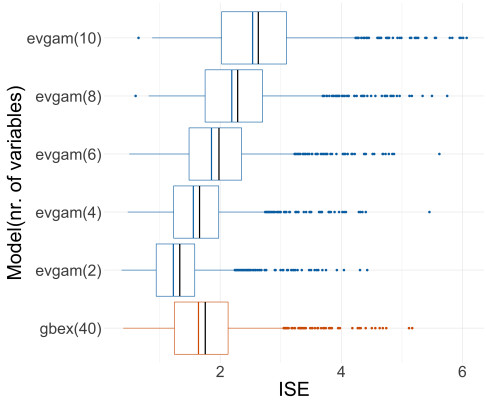


Figure 3.4: Boxplots of ISE based on 1000 replications of Model 1 for `evgam` estimator in dimension $d = 2, 4, 6, 8, 10$ (blue) and `gbex` estimator in dimension $d = 40$ (red) at probability levels $\tau = 0.995$. The vertical black lines indicate the MISE. Note that many more noise variables are used for the experiment with `gbex`, showing its robustness against the curse of dimensionality.

that there are 39 noise variables for Model 1 and 8 for Model 2. The scores of the noise variables behave all similarly and only a limited number are displayed. For Model 2, the permutation score is higher for X_1 than for X_2, due to the fact that X_1 contributes to both

shape and scale functions while X_2 only contributes to the scale function.

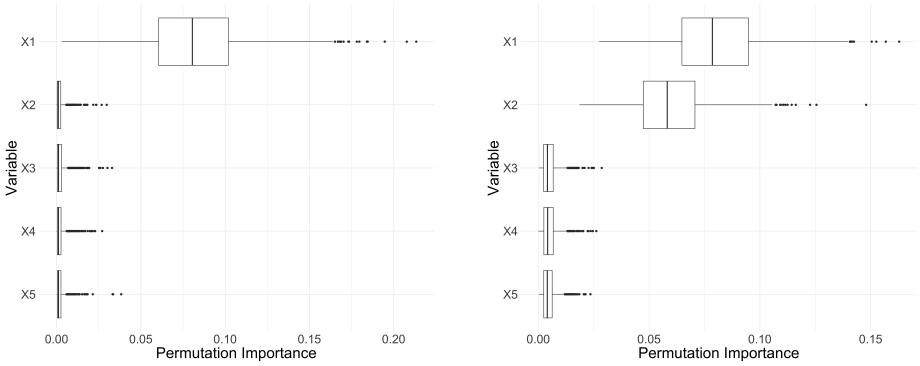


Figure 3.5: Boxplots of permutation scores defined in (3.14) for X_j, $j = 1,\ldots,5$, based on 1000 samples. Left panel: Model 1, where only X_1 contains signal. Right panel: Model 2, where only X_1 and X_2 contain signal.

The left panel of Figure 3.6 presents a typical partial dependence plot (Section 3.3.2) for $\hat{\sigma}$ based on one random sample from Model 1. This plot clearly suggests that $\hat{\sigma}$ is a step function of X_1 and does not depend on the noise variables. The partial dependence plot for $\hat{\gamma}$ indicates that the shape does not change with respect to any of the covariates. For this model, the partial dependence plots are in perfect agreement with the simulation design. For Model 2, the left panel of Figure 3.7 shows the partial dependence plot of the scale parameter with respect to X_1 and X_2. We see that the model detects the right pattern of larger values on the diagonal and in the center. The right panel shows that the model identifies the impact of X_1 on the shape parameter while the partial dependence plot of the other variables is fairly constant, again in agreement with the simulation design.

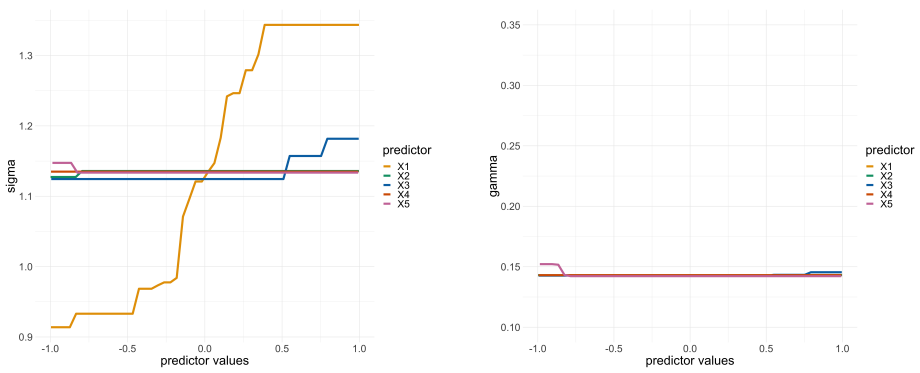


Figure 3.6: Partial dependence plots of $\hat{\sigma}$ (left panel) and of $\hat{\gamma}$ (right panel) with respect to X_j, $j = 1,\ldots,5$, based on one random sample of Model 1.

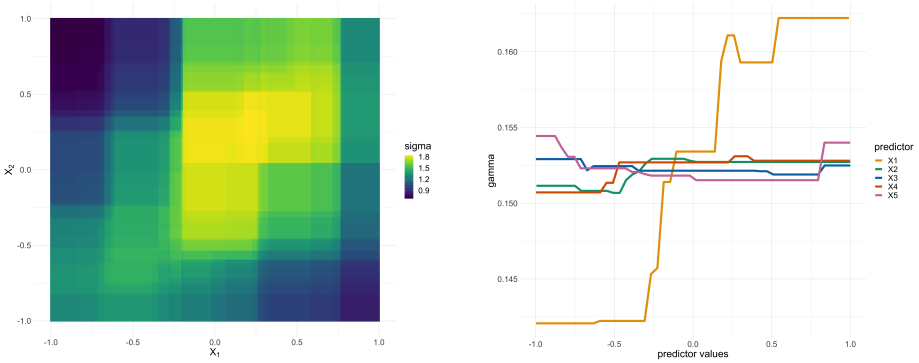


Figure 3.7: Left panel: partial dependence plots of $\hat{\sigma}$ with respect to (X_1, X_2). Right panel: partial dependence plot of $\hat{\gamma}$ with respect to X_j, $j = 1,\ldots,5$. Both experiments corresponds to one random sample of Model 2.

3.5. APPLICATION TO PRECIPITATION FORECAST

Extreme precipitation events can have disruptive consequences on our society. Accurate predictions are vital for taking preventive measures such as pumping water out of the system to prevent flooding. We apply our `gbex` method to predict extreme quantiles of daily precipitation using the output of numerical weather prediction (NWP) models.

Weather forecasts rely on NWP models that are based on non-linear differential equations from physics describing the atmospheric flow. The solutions to these equations with respect to initial conditions and parametrizations of unresolved processes form a forecast that is deterministic in nature. Introducing uncertainty in these initializations yields an ensemble forecast that consists of multiple members. In this application, we use the ensemble forecast from the European Centre for Medium-Range Weather Forecasts (ECMWF) as covariates in `gbex` to predict the daily precipitation. Using NWP output for further statistical inference to improve forecasts is known as statistical post-processing.

3.5.1. PRECIPITATION DATA

Our data set consists of ECMWF ensemble forecasts of daily accumulated precipitation and the corresponding observations at seven meteorological stations spread across the Netherlands (De Bilt, De Kooy, Eelde, Schiphol, Maastricht, Twente and Vlissingen)[1]. We use about 9 years of data, from January 1st, 2011, until November 30th, 2019, with sample size $n = 3256$. We fit separate models for each station with response variable Y equal to the observed precipitation at the station between 00 UTC and 24 UTC.

As for the covariates, we use ECMWF ensemble forecasts of daily accumulated precipitation that is computed the day before at 12UTC. The ensemble forecast contains 51 members. For efficiency, we use two summary statistics, namely the standard deviation of the ensemble members and the upper order statistics (the maximum of the ensemble members). Because most part of the Netherlands is flat and the distance be-

[1] Observed daily precipitation can be obtained from
`http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi`

tween stations is not large, we include the ensemble summary statistics of all stations as covariates for the model of each station. To account for seasonality, we additionally consider the sine and cosine with a period of 365 for the day of the year. The total covariate dimension is $d = 7 \times 2 + 2 = 16$, for each model. We denote our data as $(Y_i^{(l)}, \mathbf{X}_i)$, where $\mathbf{X}_i \in \mathbb{R}^{16}$, $i = 1, \ldots, n = 3256$ and $l = 1, \ldots, 7$. For station l, we apply the `gbex` Algorithm 2 to $\{(Y_i^{(l)}, \mathbf{X}_i), i = 1, \ldots, n\}$ to obtain estimates of $Q_{\mathbf{X}}^{(l)}(\tau)$.

3.5.2. MODEL FITTING

For model fitting, we have observed in a preliminary analysis that the output is sensitive to the initial value of (γ, σ) and we propose a specific strategy that provides better results than the default initialization. We consider a common initial value for the shape γ for all the stations and different initial values of σ for the different stations, which leads to $\theta_0 = (\gamma, \sigma_1, \ldots, \sigma_7)$. More precisely, we obtain the initial values by optimizing the log-likelihood function

$$L(\theta_0) = \sum_{l=1}^{7} \sum_{i=1}^{n} \left[(1 + 1/\gamma) \log\left(1 + \gamma \frac{Y_i^{(l)} - c}{\sigma_l}\right) + \log \sigma_l \right] \mathbb{I}_{\{Y_i^{(l)} - c > 0\}},$$

where c is a large threshold chosen such that the estimate of γ becomes stable.

We apply `gbex` as detailed in Algorithm 2 with $\tau_0 = 0.8$ for each model. We choose all tuning parameters except for B to be the same for the seven models, in such a way to achieve the overall best combined deviance score for all stations. This prevents overfitting for a specific station and it results in the following choices:
$(D^\sigma, D^\gamma) = (2,1), (\lambda_{scale}, \lambda_{ratio}) = (0.01, 12), s = 50\%$, and $(L^\sigma_{min}, L^\gamma_{min}) = (15, 45)$. Figure 3.8 shows the cross-validated deviance as a function of the number of trees B for different depth levels at two stations. The deviance behaves quite similar for the two stations and we choose $(D^\sigma, D^\gamma) = (2,1)$ for all stations. The optimal B for each station is then chosen as the minimizer of the cross-validated deviance.

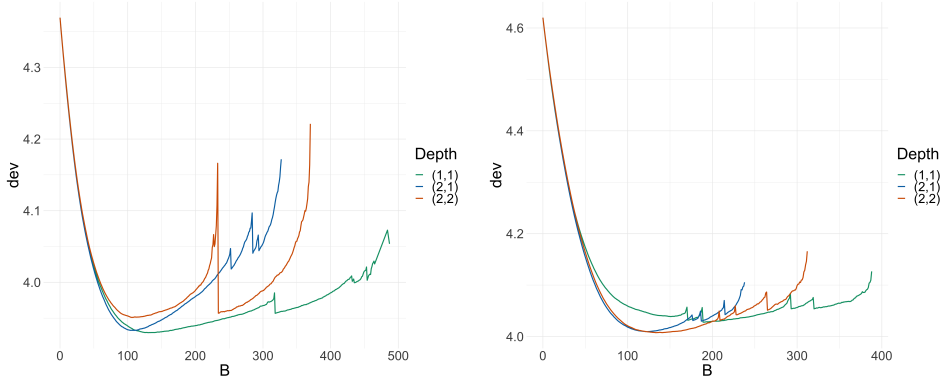


Figure 3.8: Cross-validation deviance given by (3.13) against B for the data at stations Eelde (left) and Schiphol (right) in the application in Section 3.5.

3.5.3. RESULTS

We first look into the variable importance scores for the fitted models and focus on the relative importance to understand which variables affect the scale and shape parameters, respectively. Figure 3.9 shows the relative importance for γ and σ. It is interesting to note that for the shape γ, the `day of year` is the most important variable in six out of seven models. This motivates to investigate the seasonality pattern in the extreme precipitation. The partial dependence plots of $\hat{\gamma}^{(l)}$ (left panel) and $\hat{Q}_{\mathbf{X}}^{(l)}(0.995)$ (right panel) with respect to the `day of year` are presented in Figure 3.10 for all stations. They indicate that the tail of the precipitation is heavier in summer and autumn than in winter and spring. The curves in the left panel resemble step functions and higher values of $\hat{\gamma}$ correspond to June, July and August for five stations. For the other two stations Twente and Vlissingen, it is shifted towards autumn.

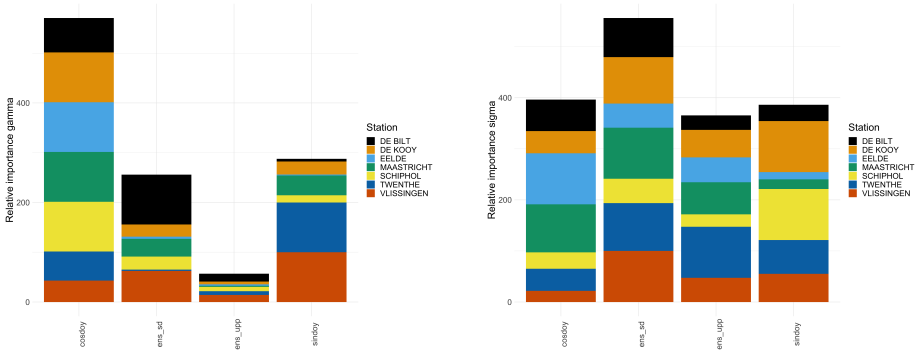


Figure 3.9: Relative variable importance score for γ (left) and σ (right). For each model, the scores are normalized such that the maximum score is 100.

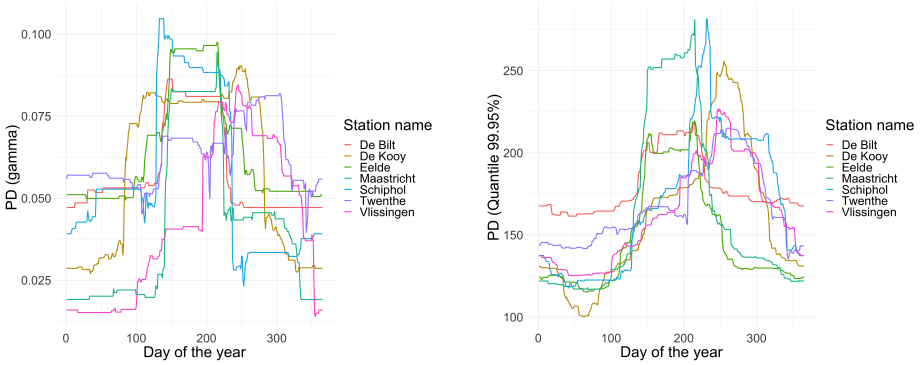


Figure 3.10: Partial dependence plots of $\hat{\gamma}^{(l)}$ (left panel) and $\hat{Q}_{\mathbf{X}}^{(l)}(0.995)$ (right panel, in mm) with respect to `day of year`.

Another relevant question concerns the contribution of ensemble statistics of other stations in forecasting the extreme precipitation of a specific location. To this end, we

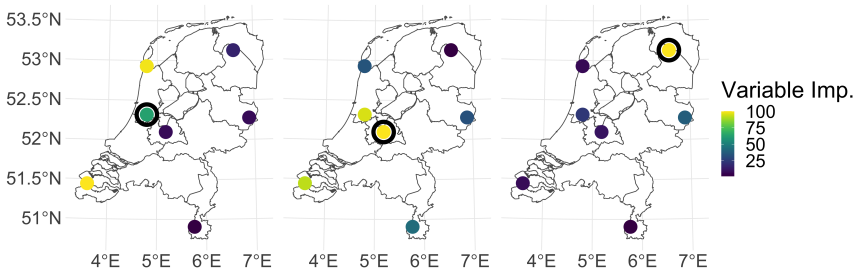


Figure 3.11: Normalized permutation scores of ensemble statistics per location for three models: Schiphol (left), De Bilt (middle), Eelde (right). The black circle indicates the station for which the model is fitted. From North to South, the stations are: Eelde, De Kooy, Twente, Schiphol, De Bilt, Vlissingen, Maastricht.

add the permutation scores of ensemble standard deviation and ensemble upper order statistics per station, resulting in seven scores for each model. We then normalize these scores such that the maximum score is 100. The results for three stations are visualized in Figure 3.11. First, quite surprisingly, when forecasting the extreme precipitation at Schiphol (left plot), the ensemble forecast relies on the information from Vlissingen and De Kooy even more than the information at Schiphol, which might be explained by a coastal effect. Similarly, the model at De Bilt (middle plot) uses the information from Schiphol and Vlissingen. For other stations like Eelde (right plot), the own information of the station is the most important. The maps of the other four stations (De Kooy, Maastricht, Vlissingen and Twente) are very similar to that of Eelde.

We finally assess the goodness of fit of our GPD model and produce QQ-plots comparing the empirical and theoretical quantiles of exceedances above threshold. We use a transformation to the exponential distribution to compare observations stemming from different stations with different covariate values. More precisely, denoting by $Z_i^{(l)}$ the ith exceedance above threshold at station l, then if our model is well-specified $Z_i^{(l)} \sim \text{GPD}\left(\hat{\sigma}^{(l)}(\mathbf{X}_i), \hat{\gamma}^{(l)}(\mathbf{X}_i)\right)$, and therefore

$$\frac{1}{\hat{\gamma}^{(l)}(\mathbf{X}_i)} \log\left(1 + \frac{\hat{\gamma}^{(l)}(\mathbf{X}_i) Z_i^{(l)}}{\hat{\sigma}^{(l)}(\mathbf{X}_i)}\right) \sim \text{Exp}(1). \tag{3.16}$$

The corresponding QQ-plots graphically assess the goodness of fit and we can see in Figure 3.12 that the `gbex` model (left panel) fits the data well at all stations, outperforming the `constant` model (right panel).

3.6. SUMMARY AND DISCUSSION

In this chapter, we have developed a gradient boosting procedure for extreme quantile regression that can handle non-linear complex problems and relatively high dimensional feature space. The tail distribution of the response Y is modelled with a Generalized Pareto Distibution (GPD), the parameters of which depend on the covariate $\mathbf{X}$. Based on exceedances over high threshold, gradient boosting produces a tree ensemble

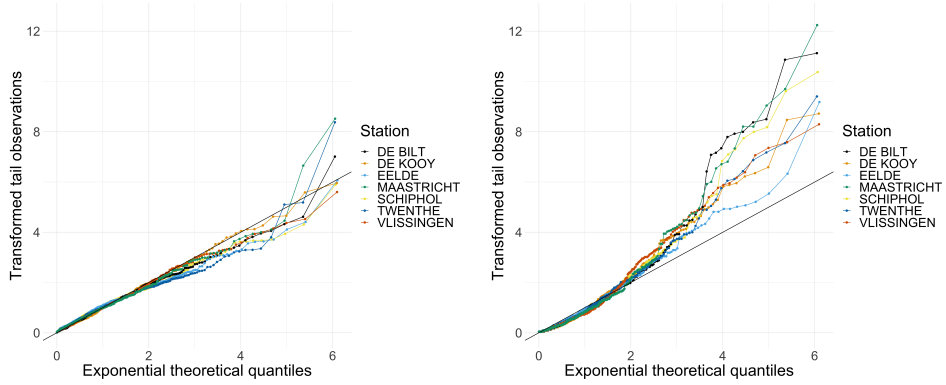


Figure 3.12: QQ-plots based on (3.16) for the estimated models at seven stations via `gbex` (left panel) and via the `constant` method (right panel).

estimating the GPD parameters using the deviance (negative log-likelihood) as the objective function to minimize. The whole procedure requires several tuning parameters and we have proposed cross-validation for tuning the key parameters and suggested default for others. Our method borrows strengths from two fields: machine learning and extreme value theory. The performance and advantages of our method is demonstrated on several numerical experiments. First, the robustness against the curse of dimension and noise variables is evidenced and the diagnostic tools are able to identify the signal variables. Second, the tree ensemble is able estimate non-linear and non-monotonic response surfaces. Third, it outperforms two machine learning methods and two classic extreme value theory methods, showing the advantage of combining both approaches. The method can be applied to complex real-worl data sets and we shown its interest for the post-processing of extreme precipitation forecast in the Netherlands.

The literature about extreme quantile regression with high dimensional covariates is rather limited and our methodological contribution fills in a gap in this area. A very natural yet challenging direction for future research is the theoretical property of our gradient boosting procedure. A consistency result for large samples is desirable but we note that all results in this direction in the existing literature on gradient boosting asume the convexity of objective function, see e.g. [7]. In our approach, the GPD deviance used as objective function is not convex in the shape parameter γ, making the theorical properties of gradient boosting very hard to address. Substantial research work would be needed to address this issue which remains outside the scope of the present chapter.

3.A. LIKELIHOOD DERIVATIVES

The gradient boosting algorithm for GPD modeling makes use of the first and second order derivatives of the negative log likelihood $\ell_z(\theta)$, $\theta = (\sigma, \gamma)$ and $z > 0$. They are re-

spectively given by

$$\frac{\partial \ell_z}{\partial \sigma}(\theta) = \frac{1}{\sigma}\Big(1 - \frac{(1+\gamma)z}{\sigma+\gamma z}\Big),$$
$$\frac{\partial \ell_z}{\partial \gamma}(\theta) = -\frac{1}{\gamma^2}\log\Big(1+\gamma\frac{z}{\sigma}\Big) + \frac{(1+1/\gamma)z}{\sigma+\gamma z},$$

and

$$\frac{\partial^2 \ell_z}{\partial \sigma^2}(\theta) = \frac{1}{\sigma(\sigma+\gamma z)}\Big(\frac{z}{\sigma} + \frac{z-\sigma}{\sigma+\gamma z}\Big),$$
$$\frac{\partial^2 \ell_z}{\partial \gamma^2}(\theta) = \frac{2}{\gamma^3}\log\big(\gamma\frac{z}{\sigma}+1\big) - \frac{2z}{\gamma^2(\sigma+\gamma z)} - \frac{(1+1/\gamma)z^2}{(\sigma+\gamma z)^2}.$$

4

INTERPRETABLE RANDOM FOREST MODELS THROUGH FORWARD VARIABLE SELECTION

Random forest is a popular prediction approach for handling high dimensional covariates. However, it often becomes infeasible to interpret the obtained high dimensional and non-parametric model. Aiming for obtaining an interpretable predictive model, we develop a forward variable selection method using the continuous ranked probability score (CRPS) as the loss function. Our stepwise procedure leads to a smallest set of variables that optimizes the CRPS risk by performing at each step a hypothesis test on a significant decrease in CRPS risk. We provide mathematical motivation for our method by proving that in population sense the method attains the optimal set. Additionally, we show that the test is consistent provided that the random forest estimator of a quantile function is consistent.

In a simulation study, we compare the performance of our method with an existing variable selection method, for different sample sizes and different correlation strength of covariates. Our method is observed to have a much lower false positive rate. We also demonstrate an application of our method to statistical post-processing of daily maximum temperature forecasts in the Netherlands. Our method selects about 10% covariates while retaining the same predictive power.

4.1. INTRODUCTION

In the past decades, random forests [8] have gained traction in many areas of application simply because random forests provide good predictive power. A random forest combines several trees, each obtained by recursively making axis-aligned splits in the covariate space until a stopping criterion is reached. The initial algorithm for random

Parts of this chapter have been submitted to the Journal of Applied statistics.

forests in [8] provides a good approach for conditional mean regression and classification. Later on, the approach was extended to estimate quantiles by [45] and further improvements were made in [1], which introduced a quantile based splitting criterion. Due to the results in [45] and [1], random forests are also used for estimating the conditional quantile function.

These quantile forests have been used in statistical post-processing to obtain probabilistic forecasts, e.g. [56], [55] and [64]. Post-processing is used as a second step in weather forecasting following a first step of physical modelling, see [35]. This first step entails a numerical weather prediction (NWP) model that uses non-linear partial differential equations of atmospheric flow on a spatial and temporal grid. Together with parametrizations of unresolved physical processes within the grid cells and an estimated initial condition, which is obtained from observational data and a so called first guess (i.e. a forecast for that time based on a previous NWP model run), the NWP model approximates the solution to the partial differential equations. An ensemble prediction system (EPS) adds uncertainty quantification to the NWP model by computing an ensemble of forecasts for perturbed initial conditions and/or the parametrization schemes [35].

Generally there is still a need for bias correction and calibration of numerical weather forecasts, which motivates the second step: statistical post-processing. Historical forecasts together with the corresponding observations are used in post-processing to estimate their statistical relationship. This relationship can then be used in order to calibrate future forecasts.

When post-processing forecasts of a weather phenomenon, a better performance is often attained by adding more information from the NWP models as predictors. For example, [64] showed that the post-processed precipitation forecasts perform substantially better when indices of atmospheric instability from the NWP models are used in modelling the statistical relation. The improvement is due to the fact that the indices of atmospheric instability help to distinguish between different types of precipitation. A full day of drizzle might accumulate to the same amount as a quick shower. However, the distributions of precipitation under these two different weather conditions are very different. Incorporating NWP forecasts of other weather phenomena enables the model to capture such differences.

A natural question is now: “ Which additional forecasts contain useful information on the phenomenon that one is post-processing?" The set of potential forecasts to include in the statistical model is generally very large and furthermore they exhibit large correlations. In practice, including too many variables often leads to a decrease in statistical efficiency, and more importantly the model becomes hard to interpret. For a practitioner, it is important to understand which variables play key roles in the statistical model and how they calibrate the EPS forecast. This motivates variable selection procedures in statistical post-processing.

A random forest is generally seen as a method that deals rather well with high dimensional covariates. This property comes from the fact that in the tree fitting algorithm, a random forest chooses, the split variables and split points, in a greedy way based on a certain criterion, e.g. the variance. This is often rather effective in the beginning of the tree fitting as many observations are split, but deep down in the tree there are fewer

observations which makes the splitting criterion subject to higher variances. Therefore global variable selection methods are considered in the literature to improve statistical efficiency and interpretation of the random forest model.

Variable selection in random forests is mainly done in terms of two types of importance measures. The first type calculates the decrease in impurity of a split made in a tree. In [43] consistency of these measures is shown on fully randomized trees. But in practice in a random forest setting these impurity measures are shown to exhibit biases ([54]). The second type is the permutation measure introduced in [8]. This measure computes how much the predictive performance decreases by randomly permuting one single predictor, which breaks the relation between response and the predictor. A popular approach is to perform a backward selection based on the permutation measures, where the model with the best predictive performance is chosen, see e.g. [28], [21] and [33].

Correlation between predictors has a large effect on the permutation importance scores. An initial approach of dealing with this is to consider conditional importance scores, [53]. This has the downside that in some way the conditioning variable has to be chosen. A more precise analysis of the effect of correlation on permutation measures is done in [33], where they conclude that a backward selection is better able to handle correlation between predictors than other strategies incorporating variable importance measures. We show in our simulation study that although the correct variables are often selected by the backward selection, there is no control on the rate of selected noise variables, i.e. the false positives.

In this chapter, we propose a new method of selecting variables with random forests. By using the so-called continuous ranked probability score as the loss function (cf. (4.5)), we are able to select variables that are informative for the entire conditional distribution instead of just for the conditional mean. The procedure estimates the predictive risk based on the so-called out-of-bag samples (cf. Section 4.3.2), which is similar to leave-one-out cross validation. Finally, we introduce a hypothesis test for each selection step to test whether a variable significantly decreases the predictive risk. We show by a detailed simulation study that our method controls the false positive rate much better than the backward selection method introduced in [33], even in the presence of high correlations.

The outline of the chapter is as follows. In Section 4.2, we give a detailed description of the mathematical set-up of the variable selection procedure. Then in Section 4.3, we give a small introduction to random forests and show how the variable selection can be applied to the random forest set-up. A comparison with backward selection based on permutation measures is made in Section 4.4. In Section 4.5, we apply the method to a practical example of post-processing maximum temperature forecasts and compare it to a standard method in post-processing. Finally, we end with a discussion in Section 4.6.

4.2. FORWARD SELECTION

In this section, we describe the mathematical set-up of our forward variable selection method. We provide the intuition of the procedure together with some theoretical motivation. For now, we consider a pair of random observations $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. Let $J \subset \{1, \dots, d\}$ denote a set of indices corresponding to the entries of the covari-

ate vector $\mathbf{X}$ and $\mathbf{X}^J$ denote the vector with the entries from $\mathbf{X}$ corresponding to J.

Let $F_{Y|\mathbf{X}^J}$ denote the conditional distribution function of Y given $\mathbf{X}^J$. And, let $L(Y,\mathbf{X}^J,F_{Y|\mathbf{X}^J})$ denote a loss functional measuring the loss between the observation Y, the quantity that we want to predict, and $F_{Y|\mathbf{X}^J}$, e.g. the squared error loss $(Y-\int z\mathrm{d}F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J))^2$. In this section, we work from the population perspective and use exact distribution functions. The next section will be concerned with the estimation of the conditional distributions using random forests.

Corresponding to the loss functional, we can now define a risk functional for the subset of variables corresponding to J,

$$R(J) = \mathbb{E}_{Y,\mathbf{X}}[L(Y,\mathbf{X}^J,F_{Y|\mathbf{X}^J})]. \tag{4.1}$$

4

In our approach an ideal variable selection procedure selects the set of variables corresponding to J that minimize this risk functional. Define $m_R = \min_{J\subset\{1,\ldots,d\}} R(J)$. Then, the optimal set of variables denoted by $\mathbf{X}^{J^*}$ is such that

$$R(J^*) = m_R \quad \text{and} \quad |J^*| = \min\{|J| : R(J) = m_R\} \tag{4.2}$$

where $|J|$ denotes the cardinality of J. The goal is to identify the smallest model that reaches an optimal risk. This is desirable when it comes to estimating the conditional distribution of Y. It is important to note that J^* is not necessarily unique. For example two collinear covariates X_1 and X_2 both contain the same information of Y, then including any of the two covariates would result in the same expected loss.

In order to obtain J^*, one could evaluate $R(J)$ for all 2^d possible sets, which is often computationally infeasible. Instead we propose a forward variable selection approach as follows. We construct a sequence of length $d+1$ of nested sets J_j for $j = 0,\ldots,d$ where $J_0 = \emptyset$ and

$$J_j = J_{j-1} \cup \left\{ \underset{q\notin J_{j-1}}{\arg\min}\, R\left(J_{j-1}\cup\{q\}\right) \right\}. \tag{4.3}$$

Our proposed forward selection procedure selects an optimal set J^o such that it is the smallest set attaining the minimum risk among J_j, $j = 0,\ldots d$. More precisely,

$$J^o = J_{\min\{j:R(J_j)=\min_{0\le i\le d} R(J_i)\}}. \tag{4.4}$$

From this point on in the chapter, we will choose the loss function equal to the Continuous Rank Probability Score (CRPS), see [30], defined by,

$$L(Y,\mathbf{X}^J,F_{Y|\mathbf{X}^J}) = \int_{-\infty}^{\infty} \left(I(Y\le z) - F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J)\right)^2 \mathrm{d}z. \tag{4.5}$$

The CRPS compares the distribution $F_{Y|\mathbf{X}^J}$ with the ideal deterministic forecast, of which the distribution function equals the step function at the observation Y. The CRPS is a proper scoring rule for a large class of distribution functions; see Section 4.2 in [30].

In the theorem below we show that under the assumption of independent covariates, the set J^o and J^* coincide.

Theorem 5. *Let $X_1,\dots,X_d$ and ϵ be independent random variables. Let $h:\mathbb{R}^{|J^*|+1}\to\mathbb{R}$ be a real valued measurable function and define $Y=h(\mathbf{X}^{J^*},\epsilon)$, where $J^*\subseteq\{1,\dots,d\}$. Assume that $\mathbb{E}[Y^2]<\infty$, and for any $I\subsetneq J\subseteq J^*$, there exists a set $S\subseteq\mathbb{R}$ with positive Lebesgue measure such that $\mathbb{E}[I(Y\le z)|\mathbf{X}^J]$ is not $\sigma(\mathbf{X}^I)$ measurable for all $z\in S$. Then J^* is the unique subset of $\{1,\dots,d\}$ satisfying (4.2), and $J^0=J^*$.*

Proof. Let $(\Omega,\mathscr{A},\mu)$ denote the probability space supporting $X_1,\dots,X_p$ and ϵ. Define the standard inner product on $L^2(\Omega,\mathscr{A},\mu)$ by $(Z_1,Z_2)=\mathbb{E}(Z_1Z_2)$, for any random variables Z_1 and Z_2 on $(\Omega,\mathscr{A},\mu)$. Then $L^2(\Omega,\sigma(X_1,\dots,X_p,\epsilon),\mu)$ becomes a Hilbert space, where the conditional expectation $\mathbb{E}(Z|\mathbf{X}^J)$ is the orthogonal projection of Z onto the closed linear subspace $L^2(\Omega,\sigma(\mathbf{X}^J),\mu)$. Now we have

$$\begin{aligned}R(J)&=\int_{-\infty}^{\infty}\mathbb{E}\left[\left(I(Y\le z)-F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J)\right)^2\right]\mathrm{d}z\\&=\int_{-\infty}^{\infty}\mathbb{E}\left[\left(I(Y\le z)-\mathbb{E}[I(Y\le z)|\mathbf{X}^J]\right)^2\right]\mathrm{d}z\\&=:\int_{-\infty}^{\infty}g_J(z)\mathrm{d}z.\end{aligned}$$

As the conditional expectation equals the orthogonal projection, for any $z\in\mathbb{R}$,

$$g_J(z)=\min_{G\in\sigma(X^J)}\mathbb{E}[(I(Y\le z)-G)^2]. \tag{4.6}$$

Therefore, for any $z\in\mathbb{R}$, if $J_1\subset J_2$, we have

$$g_{J_1}(z)\ge g_{J_2}(z). \tag{4.7}$$

This implies that $R(J_1)\ge R(J_2)$.

Next, note that if $J_2=J_1\cup\{j\}$ and $j\notin J^*$, then for any $z\in\mathbb{R}$,

$$g_{J_1}(z)=g_{J_2}(z). \tag{4.8}$$

This is because $\mathbb{E}[I(Y\le z)|\mathbf{X}^{J_2}]=\mathbb{E}\left[\mathbb{E}[I(Y\le z)|\mathbf{X}^{J_1}]\middle|\mathbf{X}^{J_2}\right]=\mathbb{E}[I(Y\le z)|\mathbf{X}^{J_1}]$ by the independence of Y and X_j and the independence between X_j and the other covariates. In this case $R(J_1)=R(J_2)$.

Finally, we show that if $J_2=J_1\cup\{j\}$, where $j\in J^*$ then $R(J_1)>R(J_2)$. We prove by contradiction. If not, then $R(J_1)=R(J_2)$, which means in view of (4.7) that $g_{J_1}(z)=g_{J_2}(z)$, for all $z\in\mathbb{R}\setminus C$, where C has zero Lebesgue measure.

From here we denote $I(Y\le z)$ by I_z to simplify notation. Expanding the squares and using the tower property of conditional expectation we see that

$$\begin{aligned}g_{J_1}(z)-g_{J_2}(z)&=\mathbb{E}\left[\left(I_z-\mathbb{E}[I_z|\mathbf{X}^{J_1}]\right)^2-\left(I_z-\mathbb{E}[I_z|\mathbf{X}^{J_2}]\right)^2\right]\\&=\mathbb{E}\left[-2I_z\mathbb{E}[I_z|\mathbf{X}^{J_1}]+\mathbb{E}[I_z|\mathbf{X}^{J_1}]^2+2I_z\mathbb{E}[I_z|\mathbf{X}^{J_2}]-\mathbb{E}[I_z|\mathbf{X}^{J_2}]^2\right]\\&=\mathbb{E}\left[\mathbb{E}\left[-2I_z\mathbb{E}[I_z|\mathbf{X}^{J_1}]+\mathbb{E}[I_z|\mathbf{X}^{J_1}]^2+2I_z\mathbb{E}[I_z|\mathbf{X}^{J_2}]-\mathbb{E}[I_z|\mathbf{X}^{J_2}]^2\middle|\mathbf{X}^{J_2}\right]\right]\\&=\mathbb{E}\left[(\mathbb{E}[I_z|\mathbf{X}^{J_1}]-\mathbb{E}[I_z|\mathbf{X}^{J_2}])^2\right]=0.\end{aligned}$$

From this we conclude that $\mathbb{E}[I_z|\mathbf{X}^{J_1}] = \mathbb{E}[I_z|\mathbf{X}^{J_2}]$ for all $z \in \mathbb{R} \setminus C$. This implies that $\mathbb{E}[I_z|\mathbf{X}^{J_2}]$ is $\sigma(X_{J_1})$ measurable which contradicts our assumption, hence $R(J_1) > R(J_2)$.

We can now observe that the forward sets are built by adding variables from J^* until all variables of J^* have been added, therefore $J^0 = J^*$. □

Remark 5. *The assumption : $\mathbb{E}[I(Y \leq z)|\mathbf{X}^J]$ is not $\sigma(\mathbf{X}^I)$-measurable for any $I \subsetneq J \subseteq J^*$, is used to prove the uniqueness of J^*. As we know that $R(J^*) = R(J^* \cup \{j\})$ for $j \notin J^*$, there are many sets, which have minimal risk in population sense. The assumption essentially ensures that J^* does contain only indices j such that the function h is not constant for x_j almost everywhere with respect of the distribution of X.*

Remark 6. *The choice of the CRPS loss function is motivated by our application. Though for different loss functions L that focus on a specific part of the conditional distribution, the procedure explained in this section could still be applied.*

4

4.3. FORWARD SELECTION USING RANDOM FORESTS

We use a random forest to estimate the conditional distribution function $F_{Y|\mathbf{X}^J}$ and the risk. Now, we make a little excursion to explain the random forest algorithm. We follow the tree construction algorithm proposed in [60] and the extension for quantile estimation from [1]. We choose this approach because it is the only approach that makes splits based on a quantile criterion, additionally in [1] asymptotic normality for the quantile estimates is established.

4.3.1. INTERMEZZO: RANDOM FORESTS

Denote the data set by $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. A random forest is defined as a collection of trees. Each tree T is obtained by recursively splitting a set of observations by making axis-aligned splits in the covariate space, meaning a split is made on a single covariate value at a time. As a result, every tree induces a partitioning of the covariate space in possibly semi-infinite hyper rectangles. Denote the conditional quantile function by $Q_{Y|\mathbf{X}}(\tau|\cdot)$, where $\tau \in (0,1)$ denotes a probability level. In this section we focus on fitting a forest in order to estimate the function $Q_{Y|\mathbf{X}}(\tau|\cdot)$. The estimation procedure for $Q_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J)$ works exactly the same by fitting a forest based on $\{(\mathbf{X}_1^J, Y_1), \ldots, (\mathbf{X}_n^J, Y_n)\}$.

Recurrent splits are made starting with parent node P, a node in the current partition, creating two child nodes C_1 and C_2, such that $P = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$. This split should be informative with respect to $Q_{Y|\mathbf{X}}(\tau|\cdot)$ and is chosen to maximize,

$$e(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P} (Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in C_1) - Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in C_2))^2, \tag{4.9}$$

where n_P, n_{C_1}, n_{C_2} are the number of observations $\mathbf{X}_i$ in each node. In practice this makes the the algorithm very slow as it requires the computation of two quantiles for each possible split. Instead in [1] a relabelling step is proposed and defined as $\mathbb{I}(Y_i > Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in P))$ for the τ quantile. Now a standard regression split, as used in a standard random forest [8], is made on the labels. This means to maximize the squared difference between the average label in both child nodes.

The trees fitted in [60] and [1] are called honest trees and are slightly different from the standard structure of tree fitting. A tree is fit by first sub-sampling a set of indices

from $\{1,\dots,n\}$ of size $s << n$ and then randomly splitting this sub-sample in two sets $\mathscr{K}$ and $\mathscr{J}$ both of size $s/2$ each. Recursive splits of $\mathbb{R}^d$ are then made based on criterion (4.9), with data points $(Y_i, \mathbf{X}_i) : i \in \mathscr{K}$. The tree becomes honest by removing all the data points indexed by set $\mathscr{K}$ and using only the data points indexed by set $\mathscr{J}$ for estimation of $Q_{Y|\mathbf{X}}(\tau|\mathbf{x})$ for a new observations $\mathbf{X}$.

A random forest is then obtained by fitting B trees. Denote by $l_b(\mathbf{X})$ the leaf node of tree b in which $\mathbf{X}$ falls. Then for $1 \le i \le n$, the weight for $(\mathbf{X}_i, Y_i)$ induced by the bth tree is given by,

$$w_{i,b}(\mathbf{X}) = \frac{\mathbb{I}(i \in \mathscr{J} \ \& \ \mathbf{X}_i \in l_b(\mathbf{X}))}{\sum_{j \in \mathscr{J}} \mathbb{I}(\mathbf{X}_j \in l_b(\mathbf{X}))}, \tag{4.10}$$

where $\frac{0}{0} = 0$. The forest weights are obtained by averaging the tree weights over the B trees, $w_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} w_{i,b}(\mathbf{X})$. An estimate of $\hat{Q}_{Y|\mathbf{X}}$ is then given by the locally weighted estimated quantile,

$$\hat{Q}_{Y|\mathbf{X}}(\tau|\mathbf{x}) = \arg\min_{\theta} \sum_{i=1}^{n} w_i(\mathbf{x}) \rho_\tau(Y_i - \theta), \tag{4.11}$$

with $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ the quantile check function. Note that the structure is similar to kernel regression, but instead of a deterministic kernel with bandwidth h the weights are determined by the data via the forest. Random forests are sometimes called adaptive nearest neighbour estimators for this reason.

In the variable selection procedure we aim to select variables that are predictive for the conditional distribution. Therefore, instead of building random forests with respect to a single τ quantile, consider a sequence of quantiles $0 < \tau_1, \dots, \tau_K < 1$. This needs a different type of relabelling than for a single quantile as explained above. They define the relabelling then by,

$$Z_i = \sum_{k=1}^{K} I(Y_i \le \hat{Q}_{Y|\mathbf{X}}(\tau_k|\mathbf{X} \in P)).$$

The best split is then chosen to maximize the following multi class classification rule:

$$\hat{e}(C_1, C_2) = \frac{\sum_{k=1}^{K} \left[\sum_{X_i \in C_1} I(Z_i = k)\right]^2}{n_{C_1}} + \frac{\sum_{k=1}^{K} \left[\sum_{X_i \in C_2} I(Z_i = k)\right]^2}{n_{C_2}}.$$

4.3.2. ESTIMATION OF PREDICTIVE LOSS

The main quantity in the theoretical framework from Section 4.2 is the CRPS risk.To make use of the random forest quantile estimator, we use an equivalent expression of the CRPS loss (4.5), given that the second moment of $F_{Y|\mathbf{X}^J}$ exists, $\int_{-\infty}^{\infty} \left(I(Y \le z) - F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J)\right)^2 \mathrm{d}z = 2\int_0^1 \rho_\tau(Y - Q_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J))\mathrm{d}\tau$. The equivalence of these two definitions is shown in the appendix. Plugging in the estimated quantile function, we obtain the following targeted loss in the estimation context:

$$L(Y, \mathbf{X}^J, \hat{Q}_{Y|\mathbf{X}^J}) = 2\int_0^1 \rho_\tau(Y - \hat{Q}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J))\mathrm{d}\tau. \tag{4.12}$$

Here we denote $\hat{Q}_{Y|\mathbf{X}^J}$ as the random forest estimator of the conditional quantile function with respect to the dataset $\{(\mathbf{X}_1^J, Y_1), \ldots, (\mathbf{X}_n^J, Y_n)\}$ and with two arguments, a probability level τ and the covariate vector $\mathbf{X}^J$.

A naive way to estimate the expected loss (that is the expectation of (4.12)), would be considering

$$\frac{2}{n}\sum_{i=1}^{n}\int_0^1 \rho_\tau(Y_i - \hat{Q}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}_i^J))\mathrm{d}\tau.$$

However, this would lead to over-fitting because the training set (data for estimating $Q_{Y|\mathbf{X}^J}$) are the same as the testing set (data for estimating the expectation). This problem can be circumvented by using so called out-of-bag samples as test set.

The out-of-bag samples for the bth tree are defined as the samples that are not used for generating the tree. For each observation $(\mathbf{X}_i^J, Y_i)$, a sub forest $\mathscr{F}_i$ is defined by $\mathscr{F}_i = \{T_b : i \notin (\mathscr{K}_b \cup \mathscr{J}_b)\}$. Namely, this sub forest consists of trees for which $(\mathbf{X}_i^J, Y_i)$ is out-of-bag. Observe that the number of trees in $\mathscr{F}_i$ is random and hence not necessarily the same for all i. The expected number of trees for each sub forest is $B\left(1 - \frac{s}{n}\right)$.

We use the sub forest $\mathscr{F}_i$ to estimate the conditional quantile function and denote it with $\hat{Q}^{\mathscr{F}_i}_{Y|\mathbf{X}}(\tau|\mathbf{X}^J)$. Since the trees in sub forest $\mathscr{F}_i$ do not use observation $(\mathbf{X}_i^J, Y_i)$, we use this quantile estimator to evaluate the CRPS loss for $(\mathbf{X}_i^J, Y_i)$. Doing this for all observations, we obtain the estimated CRPS risk given by,

$$\hat{R}(J) := \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \rho_\tau(Y_i - \hat{Q}^{\mathscr{F}_i}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}_i^J))\mathrm{d}\tau. \tag{4.13}$$

In the sequel, we write $\hat{Q}^{\mathscr{F}_i}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}_i^J) = \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^J)$ for simplicity.

This out-of-bag procedure for estimating risk has similarities to leave-one-out cross validation. For validating the ith observations we use all trees which do not use the ith observation. The difference is that sub forests have in expectation the same size, but not exactly. Computationally the out-of-bag sample approach is also much faster compared to leave-one-out cross validation. Note that a tree has $n-s$ out-of-bag samples and hence the tree is used is used in $n-s$ sub-forests. On the other hand leave one out cross validation does not reuse trees and estimates a new forest for each element in the summation of (4.13).

4.3.3. ONE STEP FORWARD

The forward variable selection sequentially adds variables such that the predictive loss is minimized. We here explain how each step is performed. Recall that for a index set J, the estimated risk $\hat{R}(J)$ is given by (4.13). Suppose that we have selected the first $j-1$ variables with indices in $\hat{J}_{j-1}$. Then the jth variable $X_{\hat{i}_j}$ is selected based on

$$\hat{i}_j = \underset{q \notin \hat{J}_{j-1}}{\operatorname{argmin}}\, \hat{R}(\hat{J}_{j-1} \cup \{q\}). \tag{4.14}$$

and $\hat{J}_j = \hat{J}_{j-1} \cup \{\hat{i}_j\}$. The procedure of a single step forward is detailed in Algorithm 3.

Algorithm 3: A forward step

Result: $\hat{\imath}$, $\hat{R}(J\cup\{q\}) : q \notin J$
Define data $(Y_1, \mathbf{X}_1), \dots (Y_n, \mathbf{X}_n)$;
Define set $J \subset \{1, \dots p\}$;
for $q \notin J$ **do**
  construct a forest with $(Y, \mathbf{X}^{J\cup\{q\}})$;
  Calculate $\hat{R}(J\cup\{q\})$;
end
$\hat{\imath} = \text{argmin}_q \hat{R}(J\cup\{q\})$

4.3.4. STOPPING ON TIME

Motivated by the result in Theorem 5, we stop selecting variables when there is no further decrease in CRPS risk. From the proof of Theorem 5, adding variables that are not in J^* does not have an effect on the CRPS risk. In practice, where we are working with finite samples, additional covariates decrease in fact the statistical efficiency of the random forest which leads to higher CRPS values. Because of the random component in the forest procedure, different forests will have different risk. In general this can be avoided by fitting an enormous number of trees to reduce the random component, but in practice this is infeasible. Instead we use the randomness to test the following hypothesis at each step,

$$H_0 : R(\hat{J}_{j-1}) - R(\hat{J}_j) = 0$$
$$H_A : R(\hat{J}_{j-1}) - R(\hat{J}_j) > 0.$$

The fitted forests at j-th and $(j+1)$-th steps are used to obtain several estimates of $R(\hat{J}_{j-1}) - R(\hat{J}_j)$. More precisely, we estimate this difference by $\hat{R}(\hat{J}_{j-1}\cup\{q\}) - \hat{R}(\hat{J}_j\cup\{q\})$, where $q \notin \hat{J}_j$. Note that $\hat{R}(\hat{J}_{j-1}\cup\{q\})$ is computed at the j-th step for identifying $\hat{\imath}_j$ and $\hat{R}(\hat{J}_j\cup\{q\})$ at the $(j+1)$-th step for identifying $\hat{\imath}_{j+1}$. So, the testing procedure does not require any extra forest fitting. We propose the following test statistics:

$$W_q = \sum_{q\notin\hat{J}_j} \mathbb{I}(\hat{R}(\hat{J}_{j-1}\cup\{q\}) - \hat{R}(\hat{J}_j\cup\{q\}) > 0). \tag{4.15}$$

Under the null hypothesis, the variable added on the jth step does not contribute to the predictive performance of the model. As a result both risks are asymptotically equal (see the proof for Theorem 6), meaning that the test-statistic approximately has a binomial distribution, $\text{Bin}(M_j, 0.5)$, where $M_j = d - |\hat{J}_j|$. We reject H_0 if $W > C^j_{1-\alpha}$, where $C^j_{1-\alpha}$ is the $1-\alpha$ quantile of $\text{Bin}(M_j, 0.5)$. The consistency of this test is established in the theorem below.

Theorem 6. *Assume that for any $\tau \in (0,1)$, as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} \left| \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}^J_i) - Q(\tau|\mathbf{X}^J_i) \right| \xrightarrow{p} 0, \tag{4.16}$$

where $J = \hat{J}_j \cup \{q\}$ *or* $J = \hat{J}_{j-1} \cup \{q\}$, $q \notin \hat{J}_j$. *Then, under the assumptions of Theorem 5,*

$$P(W > C^j_{1-\alpha}) \to 1 \quad \textit{Under hypothesis } H_A, \tag{4.17}$$

as $n \to \infty$.

Proof. It suffices to prove that under H_A, as $n \to \infty$

$$\mathbb{E}[W] \to \omega_0,$$

where $\omega_0 > C^j_{1-\alpha}$.
Denote $I_q := \hat{J}_{j-1} \cup \{q\}$ and $K_q := \hat{J}_j \cup \{q\}$. Then, we have

$$\begin{aligned}
&\hat{R}(I_q) - \hat{R}(K_q) \\
=&\frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1 \rho_\tau(Y_i - \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{I_q}))\mathrm{d}\tau - \int_0^1 \rho_\tau(Y_i - \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{K_q}))\mathrm{d}\tau\right) \\
=&\frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1 \rho_\tau(Y_i - \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{I_q}))\mathrm{d}\tau - \int_0^1 \rho_\tau(Y_i - Q(\tau|\mathbf{X}_i^{I_q}))\mathrm{d}\tau\right) \\
&+\frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1 \rho_\tau(Y_i - Q(\tau|\mathbf{X}_i^{K_q}))\mathrm{d}\tau - \int_0^1 \rho_\tau(Y_i - \hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{K_q}))\mathrm{d}\tau\right) \\
&+\frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1 \rho_\tau(Y_i - Q(\tau|\mathbf{X}_i^{I_q}))\mathrm{d}\tau - \int_0^1 \rho_\tau(Y_i - Q(\tau|\mathbf{X}_i^{K_q}))\mathrm{d}\tau\right) \\
=:&S_1 + S_2 + S_3.
\end{aligned}$$

Applying the Knight's identity, $\rho_\tau(u-v) - \rho_\tau(u) = -v(\tau - I(u<0)) + \int_0^v (I(u \le s) - I(u \le 0))\mathrm{d}s$, which implies that $|\rho_\tau(u-v) - \rho_\tau(u)| \le 2|v|$, we have

$$|S_1| \le \frac{4}{n}\sum_{i=1}^{n}\int_0^1 \left|\hat{Q}^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{I_q}) - Q^{\mathscr{F}_i}(\tau|\mathbf{X}_i^{I_q})\right|\mathrm{d}\tau \xrightarrow{p} 0,$$

by (4.16). The same result holds for S_2.
Observe that S_3 is the sample mean of I.I.D. random variables with expectation $R(I_q) - R(K_q)$. Applying law of large number, $S_3 \xrightarrow{p} R(I_q) - R(K_q)$. Combing with the results for S_1 and S_2, we have

$$\hat{R}(I_q) - \hat{R}(K_q) \xrightarrow{p} R(I_q) - R(K_q).$$

Under H_a, $\hat{i}_j \in J^*$, thus, by the proof for Theorem 5, for all $q \notin \hat{J}_j$,

$$R(I_q) - R(K_q) > 0.^{1}$$

This implies that

$$\mathbb{E}[W] = \sum_{q \notin \hat{J}_j} \mathrm{P}\left(\hat{R}(I_q) - \hat{R}(K_q) > 0\right) \xrightarrow{p} M_j > B_{1-\alpha}.$$

□

[1] Obviously under H_0, $R(I_q) - R(K_q) = 0$.

Remark 7. *The expression in condition (4.16) ensures that the average out-of-bag estimation errors of the quantile random forest converge to zero. This is comparable to a leave-one-out cross validation setting, but instead of n different random forests, n sub-forests are used for estimation.*

In practice, the integration in (4.13) is numerically approximated. Let $\tau_t = \frac{t}{k+1}, t = 1,\dots,k$, where k is a pre-specified integer. The estimated risk $\hat{R}(J)$ in (4.13) is approximated by

$$\hat{R}(J) = \frac{2}{k}\sum_{t=1}^{k} \rho_{\tau_t}(Y_i - \hat{Q}_{Y|\mathbf{X}}(\tau_t|\mathbf{X})). \tag{4.18}$$

The complete procedure is given in Algorithm 4.

Algorithm 4: Forward variable selection

Result: J^o
Set data $(Y_1,\mathbf{X}_1),\dots(Y_n,\mathbf{X}_n)$;
Set $j = 1$, $J_0 = \emptyset$, α;
Calculate J_1 with Algorithm 3 using $J = J_0$;
repeat
  $j = j + 1$;
  Calculate J_j with Algorithm 3 using $J = J_{j-1}$;
  Calculate W_j from equation 4.15;
until $W_j \leq C^j_{1-\alpha} \mid j == p$;
$J^o = J_{j-1}$;

4.4. COMPARISON BASED ON SIMULATION

In this section we compare the performance of our variable selection procedure with the backward selection based on a permutation measure with a mean squared error criterion proposed in [33]; details of the method are stated later in this section. We compare with this method as it is currently the only method that deals with correlated predictors for random forests and we will refer to it as the backMSE method. For the comparison we simulate data from the following model,

$$Y = \mu(\mathbf{X}) + \sigma(\mathbf{X})\epsilon, \tag{4.19}$$

where ϵ follows a standard normal distribution and independent of this, $\mathbf{X} \in \mathbb{R}^{25}$ follows a multivariate normal distribution. For the covariance structure of $\mathbf{X}$ we split up the covariates into blocks $I_l = \{(l-1)*5+1,\dots,(l-1)*5+5\}$ for $l = 1,\dots,5$. The covariance function of $\mathbf{X}$ is then given by,

$$\mathrm{Cov}(X_j, X_i) = \begin{cases} 1, & \text{if } i = j; \\ \rho, & \text{if } i,j \in I_l \text{ for the same } l; \\ 0, & \text{otherwise.} \end{cases} \tag{4.20}$$

The two selection methods are compared for $\rho \in \{0, 0.4, 0.8\}$. For the functions μ and σ three different models are considered:

$$\mu_1(\mathbf{X}) = X_1 + \frac{X_6}{2} + \frac{X_{11}}{4}, \qquad \sigma_1(\mathbf{X}) = 1;$$

$$\mu_2(\mathbf{X}) = X_1, \qquad \sigma_2(\mathbf{X}) = \exp\left(\frac{X_6}{2} + \frac{X_{11}}{3}\right);$$

$$\text{and } \mu_3(\mathbf{X}) = \begin{cases} X_6^2, \text{ if } X_1 \geq 0, \\ -X_6, \text{ if } X_1 < 0, \end{cases} \qquad \sigma_3(\mathbf{X}) = \begin{cases} 2, \text{ if } X_{11} \geq 0, \\ 1, \text{ if } X_{11} < 0. \end{cases}$$

The first model is a model where the covariates only influence the mean, in the second model the influence is mainly on the variance. The third model considers discontinuous covariate dependence in both mean and variance. Finally, we choose sample sizes $n \in \{500, 1000, 2500\}$.

The backMSE method evaluates the relevance of a covariate by its permutation importance measure, which is defined as

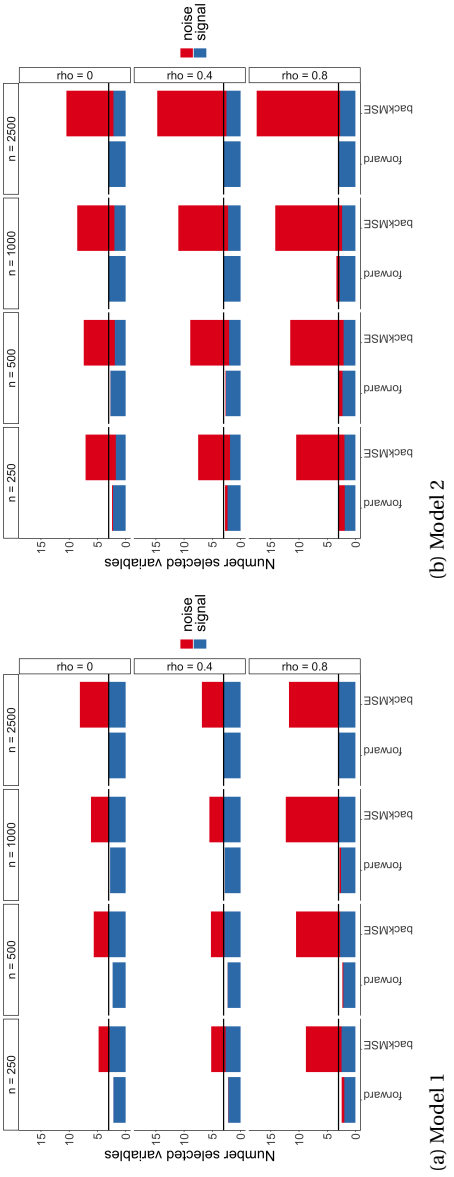
$$I(X_j) = \mathbb{E}\left[(Y - \mathbb{E}(Y|\mathbf{X}_{(j)}))^2\right] - \mathbb{E}\left[(Y - \mathbb{E}(Y|\mathbf{X}))^2\right],$$

where $\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_d)$ such that $X'_j =^d X_j$ and X'_j is independent of Y and of the other covariates. A large score of $I(X_j)$ indicates that covariate X_j is important. The method randomly permutes the values of X_j to mimic a random sample of X'_j. An estimator of $I(X_j)$ using out of bag samples is given in (2.1) in [33].

In [33] it is shown that the order of the permutation importance measures can not be naturally interpreted in the presence of correlation between the covariates, as variables that are correlated share their importance. As a result, the importance of the important variables is lower than it should be. The backMSE deals with this problem by iteratively removing the least important variable and refitting the model and calculating the importance scores. This process is repeated until no variables are left. The optimal model is then chosen as the model that minimizes the out-of-bag mean squared error. Why this works is easily seen with two highly correlated informative variables. Initially they do not seem important because they share their importance, but by removing one the importance is not shared any more. The left over variable shows the true importance and will therefore be in the selected set.

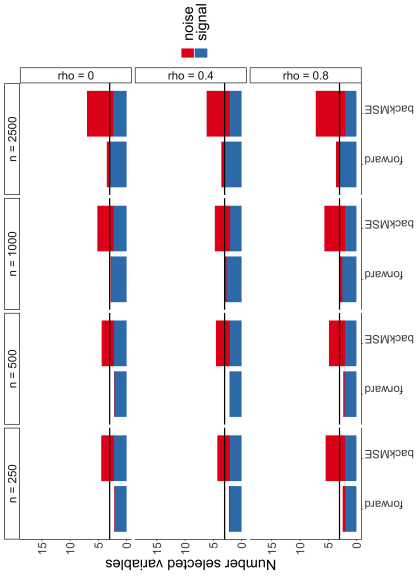
It is recommended in [33] to compute several forests and take averages to stabilize the variable importance scores and the error estimates. We compute for each step 20 forests where each forest contains 2000 trees. For this method, we follow the standard forest algorithm from [8], fitting trees based on bootstrap samples of size n, $mtry$ is set to the default value for regression $p/3$ and taking a minimum leaf size of 5.

For our method we also take fixed parameters with sub sampling fraction $s = 0.5$, a minimum node size of 1, $mtry = p$ and 1000 trees. We have tested the influence of these tuning parameters on several simulation models and the results are rather robust to different choices. Our selection model adds variables one at a time and stops when additional variables do not increase performance. As the model is therefore often small it makes sense to not over randomize by setting $mtry$ to smaller than p. We advise to choose a small s for large datasets in order to reduce computation time.



(a) Model 1

(b) Model 2



(c) Model 3

Figure 4.1: Average number of selected variables over 100 simulations. Total number of variables in height of the bar blue for signals variables and red for noise variables. Bars from left to right correspond to the sample sizes and top to bottom corresponds to different levels of correlation. The horizontal line coincides the total number of signal variables in the model.

For each model we simulate 100 data sets. The results are summarized in Figure 4.1. For the first model we see that the backMSE method retrieves more signal variables than the forward selection for low sample sizes and that as the sample size grows the forward variable selection also recovers all signal variables. A large difference is seen in the number of noise variables that are selected. The forward selection performs much better in this than the backMSE, which systematically selects noise variables and tends to even select more as the sample size increases. This phenomenon is also visible for Models 2 and 3 as seen in Figures 4.1b and 4.1c. For these two models where the variance is dependent on covariates, the CRPS criterion clearly has an edge over the backMSE that selects variables based on the mean squared error and therefore has a hard time selecting these variables.

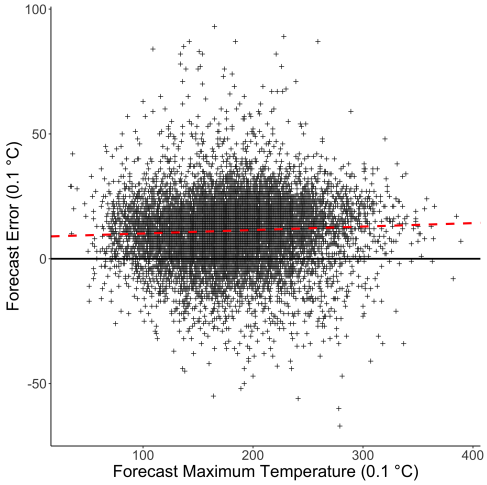
The reason why the backMSE selects many noise variables is two-fold. First the backMSE method selects the optimal set based on a predictive mean squared error criterion. This approach does not account for the inherent variable selection within the random forest, where at each node the split that reduces the variance the most is chosen. As a result the random forest is able to ignore noise variables partially. In practice this means that in an out-of-bag performance measure the addition of a single noise variable cannot be detected. Therefore the variables that are selected will not be the smallest set, but instead a set with maximum number of noise variables maintain the lowest performance. Secondly, the backMSE does not adequately deal with the correlation. For example in Model 1 with $\rho = 0.8$ all variables $X_1, \ldots, X_5$ have higher variable importance compared to X_6, which means that if X_6 is in the model, so are $X_2, \ldots, X_5$.

Thanks to our testing approach, a small number of noise variables is selected with the forward selection. Using the randomness induced by the random forest, our testing procedure selects a variable that leads to a significant reduction of the predictive loss. The significance level naturally controls the number of selected noise variables by the nature of the testing procedure. We have set the significance level to 5% for all simulations in the chapter.

4.5. POST-PROCESSING MAXIMUM TEMPERATURE FORECASTS

There are substantial risks related to extremely high temperatures. Consecutive days of high temperatures, i.e. heat waves, lead to higher mortality, especially older people. Besides high temperatures can cause train rails to expand and thereby potentially disrupt the train system. Additionally, in the absence of rain they likely cause severe droughts as seen in 2018 in The Netherlands, which has had large consequences for nature areas and agriculture. The Royal Netherlands Meteorological Institute (KNMI) issues alarms for persistent warm weather. To design a good alarm system it is essential to have good quality weather forecasts. One of the most used ensemble models, the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble model, has a negative bias in the maximum temperature forecast. As an illustration, Figure 4.2 shows the forecast bias for data observed at weather station de Bilt where KNMI is located. For accurate forecasts, this bias needs to be corrected for. This can be easily done by estimating the linear relation between the forecasts and the observations. Although this quickly improves the maximum temperature forecast, this leaves unused a vast amount of forecast data for other weather types. We will show that using a wide range of potential covariates, the

Figure 4.2: Scatter plot of error of the ECMWF high resolution deterministic run maximum temperature forecast against that deterministic forecast in the warm half year for the years 2011-2019 at De Bilt. The black line indicates a zero error and the red dashed line is the linear regression of the data points.



maximum temperature forecasts are improved further than by a simple bias correction. By performing the variable selection we then also investigate in more detail what effect different covariates have on the forecast distribution estimated using the random forest model.

We use maximum temperature observed at seven stations spread across The Netherlands, namely Den Helder, Schiphol, De Bilt, Eelde, Twente, Vlissingen and Maastricht (http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi). The focus is on high temperatures, hence we consider only observations from mid-April until mid-October. In total, we look at 9 years of data ranging from 2011 to 2019.

As covariates we use the output of the ECMWF model, which contains a 51 member ensemble and a higher resolution deterministic run. These forecasts are initiated two times a day, at 00 UTC and at 12 UTC, but here we use only forecasts of the latter run. We define the lead time of the forecast as the time difference between the start of the day for which the forecast is valid and the initiation time of the forecast. For this analysis we will consider forecasts with lead times equal to $36+24k$ hours for $k = 0,1,2,3,4,5$. The ensemble contains 51 exchangeable members and in order to use them we compute a set of summary statistics from the ensemble. These summary statistics are the mean, standard deviation, quantiles and number of ensemble members exceeding a pre-specified threshold. For the quantiles in our application we choose the 25, 50 and 75 percent quantiles. Thresholds are chosen as to extract different types of information from the ensemble relative to the weather phenomenon itself. For cloud cover we use three thresholds, 20 percent, 50 percent and 80 percent of cloud cover to create variables measuring probabilities of a few to no clouds, partly clouded weather and clouded weather.

Apart from the forecasts for maximum temperature and cloud cover, we consider other covariates including forecasts for daily average temperature at 2m, dew point tem-

perature, minimum temperature, daily average wind speed and daily accumulated precipitation. For long lead times, predictability of these typical weather phenomena decreases, but the range of predictability of for example flow pattern at 500 hPa extends much further. Therefore the first three principal components flow pattern at 500 hPa over Europe are also used as predictors [40]. Note that these covariates are the same for each station.

For the response variable we consider the forecast error, which we obtain by subtracting the deterministic forecast run from the observed maximum temperature. By doing so, the seasonality in the temperature is largely reduced. In Figure 4.2, the forecast error is clearly visible as the distance between the red linear regression line and the x-axis is rather large. Additionally it is clear that the spread of error changes as a function of the deterministic forecast. A possible explanation is that there is still remain seasonality effects that are not taken care of by a simple linear effect. Therefore, also the sine and cosine of the day of the year with a period of one year and half a year are included as two predictors. In total this gives us 71 covariates. For a given lead time an observation on a given day is denoted by $(Y, \mathbf{X})$, with Y the error of the deterministic run and $\mathbf{X}$ the 71 dimensional covariate vector.

In this section, we will explore 3 methods, quantile random forests as in [1] with all variables, quantile random forests with variables selected by our forward variable selection and Non-homogeneous Gaussian Regression (NGR) [31]. This third method is known in the meteorology literature as an EMOS (Ensemble Model Output Statistics) method and is used as a standard approach in post-processing. The NGR method assumes the data follow a Gaussian model,

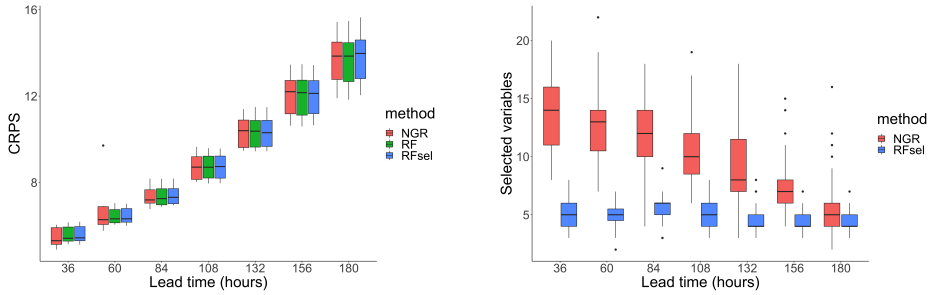
$$Y|\mathbf{X} = \mathbf{x} \sim N\left(\mathbf{x}^T\beta, \exp(\mathbf{x}^T\gamma)\right).$$

The parameters β and γ are then estimated by maximum likelihood. For this model, we select variables based on the Bayesian Information Criterion (BIC) by a forward and backward stepwise approach.

For each station and lead time, we fit a separate model. The models are estimated with a 9-fold cross validation, each time leaving out a single year. In Figure 4.3a the CRPS risk is shown as a function of lead time, where the box-plots contain the CRPS risk for all stations. Then in Figure 4.3b the number of selected variables is shown for our method and NGR, where we leave out the random forest with all 71 variables.

Based on the CRPS, all methods perform comparably. This is also confirmed by other verification measures such as reliability diagrams, quantile reliability diagrams and probability integral transform histograms, which are not shown in this chapter. A selection of these diagrams is shown in the appendix. The interesting part comes from the number of selected variables. Our method selects a small portion (less than 10%) of covariates, substantially less than NGR. We investigate this further by considering which variables are selected. The result is visualized in Figure 4.4. For each lead time, the color indicates the frequency of a covariate being selected by 63 estimated models (7 locations and 9 cross-validation sets per location). An extremely important variable would be selected all 63 times.

Yellow boxes correspond to a few variables that are always selected. But the number of light blue boxes is much smaller for our method compared to NGR. From this



(a) CRPS

(b) Selected variables

Figure 4.3: (Left panel) the CRPS risk against leadtime where the boxplots contain the CRPS risk for each station. (Right panel) the number of selected variables against leadtime.

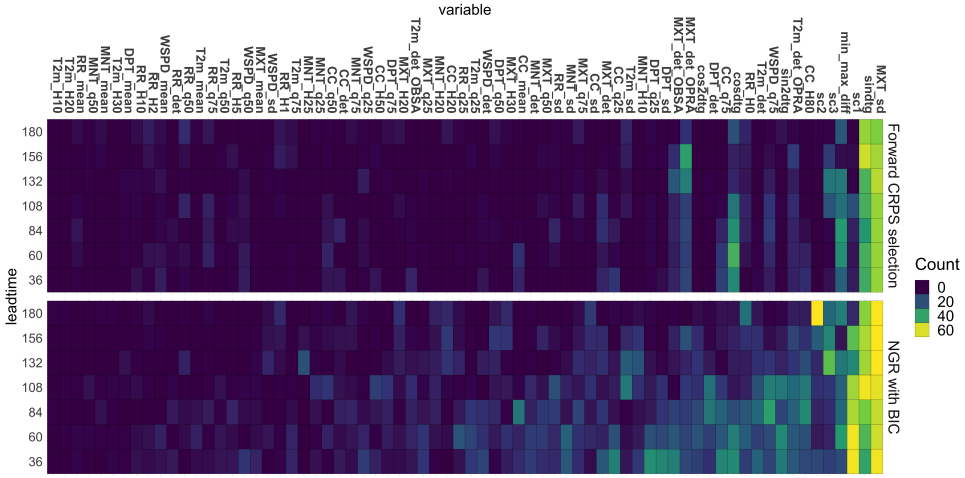


Figure 4.4: Selected variables for each lead time. All cross validations and all stations have been aggregated where the maximum number of times a variable can be selected is 63.

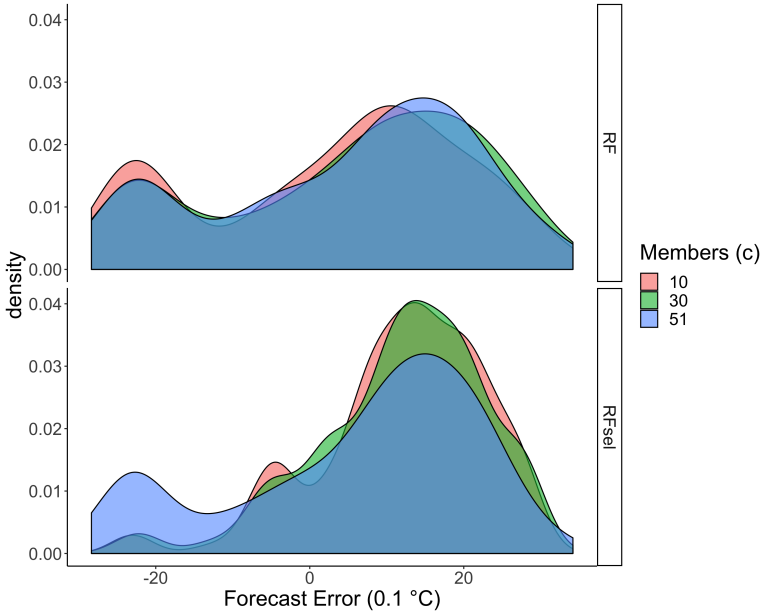


Figure 4.5: The conditonal distribution of the forecast error, based on estimated models for De Bilt with lead time 36 hours. Different colors indicate different values of cloud cover while the values of other covariates are fixed to be the same as that for 31-05-2018 at de Bilt. Top figure for the random forest model with all variables and the bottom figure for the random forest model with the selection of variables.

we conclude that our method selects fewer variables and it also selects similar variables for different stations. This suggests that our variable selection method is more robust compared to the NGR method for short lead times, where a diverse set of variables is selected.

The main variables that our method selects are the sine of the day of the year, the standard deviation of the ensemble forecast and variables related to cloud cover. Since our procedure typically selects a small set of variables, it is then feasible to interpret the estimated model. For instance, to investigate how a selected covariate, say X_j influences the forecast distribution of Y, one can compare the conditional distribution of Y given different values of X_j while the other covariates denoted by $\mathbf{X}_{(-j)}$ are kept the same. We consider Y, the forecast error at de Bilt with lead time 36 hours and X_j the cloud cover, which is the number of ensemble members with cloud cover exceeding 50%. The values of other covariates are fixed the same as the data of 31-05-2018 at De Bilt, denoted by $\mathbf{X}_{(-j)} = \mathbf{x}^*_{(-j)}$. Figure 4.5 shows the conditional density of Y given $(X_j = c, \mathbf{X}_{(-j)} = \mathbf{x}^*_{(-j)})$, where different colors indicate three different values of c. Note that all 51 ensemble members exceed 50% cloud cover. As shown in the lower panel of Figure 4.5, cloud cover clearly has an effect based on the estimation of our method: $c = 51$ yields a bimodal distribution while $c = 10$ leads to a unimodal distribution. This suggests that in this configuration, higher cloud cover implies a higher chance for a negative forecast error (left mode in the plot). However, the distributions obtained by random forest (without vari-

able selection) are very similar; see the upper panel of the figure. This is because that there are other covariates correlated to cloud cover, and these covariates still indicate that there is a high cloud cover even when the number of ensemble members exceeding 50% could cover is set to 10. In other words, changing the value of a single variable in a random forest with many correlated covariates is *not* interpretable. Such a random forest model fails to capture the effect of a signal variable.

4.6. SUMMARY AND DISCUSSION

In this chapter, we have proposed a general framework for a forward variable selection with respect to a loss function. We show in population sense that under an independence assumption between covariates and by choosing the continuous ranked probability score as loss function that the forward selected variables form the correct set with respect to the CRPS risk functional. Applying the method in a random forest set-up, we show that the out-of-bag samples can be efficiently used to asses predictive performance. The main difficulty in the procedure is determining the stopping time, that is when selecting more variables does not add in predictive performance. Due to randomness and the inherent greedy variable selection procedure in the random forest algorithm this can not be determined by the calculated predictive performance. Instead in a single forward selection step we use the predictive performance of each possible set to construct a test to detect increasing predictive performance. The procedure then stops a null hypothesis of non increasing predictive performance can not be rejected. We show that this test is consistent.

With a simple simulation study we show that our variable selection method, compared to a backward selection based on a permutation importance measure, is more capable of discriminating between signal variables and noise variables. This improvement is shown for various sample sizes and correlations between the covariates.

In an application on post-processing maximum temperature, our method shows consistency in the number of selected variables and in the variables being selected over several stations. Moreover, our method selects less than 10 percent of the covariates and still attains the same predictive power as the quantile random forest with all covariates. Further, it is easier to interpret our resulting model, due to the largely reduced number of covariates. Without variable selection, it is hardly possible to analyse the effect of a single covariate in a random forest model when it is heavily correlated to other covariates. In our data example, in the presence of thick cloud cover, our random forest model indicates that there is a higher risk of over forecasting (lower panel of Figure 4.5) instead of under-forecasting which was indicated by Figure 4.2.

There are two interesting directions for future research. First, the theoretical results in Sections 4.2 and 4.3 are derived under the assumption that the covariates are independent. However, the ability of our method to select signal variables from a correlated setting is evidenced by our simulation study and data application. It is interesting to investigate such a setting. Second, we focus in this chapter on how this forward method behaves for the CRPS, but the mathematical set-up in Section 4.2 is much more general and allows to select variables with respect to other loss functions. It would be interesting to extend the current results to a more general set of loss functions.

4.A. CRPS CALCULATIONS

Here we show for an observation y and a distribution function F that the CRPS calculated from the quantile perspective as well as from the distribution function perspective are equivalent as shown in [41], i.e we show that

$$2\int_0^1 \rho_\tau(y - F^{-1}(\tau))\mathrm{d}\tau = \int_{-\infty}^{\infty} (I(y \le z) - F(z))^2 \mathrm{d}z. \tag{4.21}$$

We first assume that the distribution with distribution function F has a finite second moment, then we have,

$$\begin{aligned}
2\int_0^1 \rho_\tau(y - F^{-1}(\tau))\mathrm{d}\tau &= 2\int_0^1 (I(y \le F^{-1}(\tau)) - \tau)(F^{-1}(\tau) - y)\mathrm{d}\tau \\
&= 2\int_{-\infty}^{\infty} (I(y \le z) - F(z))(z - y) f(z)\mathrm{d}z \\
&= -(I(y \le z) - F(z))^2 (z - y)\Big|_{-\infty}^{\infty} + \\
&\int_{-\infty}^{\infty} (I(y \le z) - F(z))^2 \mathrm{d}z \\
&= \int_{-\infty}^{\infty} (I(y \le z) - F(z))^2 \mathrm{d}z
\end{aligned}$$

Here we use a substitution in the second line of $\tau = F(z)$ and in the third line we apply integration by parts. The finite second moment is used in the fourth line such that the first term converges to 0.

4.B. CALIBRATION OF FORECASTS FOR LEAD TIME 60 AND STATION DE BILT

Figure 4.6 shows a histogram of the $\hat{F}(Y)$ where $\hat{F}$ is the forecast distribution for observation Y. If F is calibrated the histogram should look like the histogram based on standard uniform random variable.

Figure 4.7 shows reliability diagrams. Let t be a threshold and define $p = \hat{F}(t)$ and $I = I(Y \le t)$ for each forecast. A reliability diagram bins the probabilities p in equally sized bins. The average indicator I should be the same as the average p. Hence plotting these averages they should be approximately on the identity line; for detailed explanation we refer to [65].

Figure 4.8 shows quantile reliability diagrams. Let τ be a probability level and $\hat{Q}$ the forecast quantile function. Define $q = \hat{Q}(\tau)$ for each forecast. A quantile reliability diagram bins the quantiles q in equally sized bins. The τ quantile of observation Y should be the same as the average q. Hence plotting these against each other should be approximately on the identity line; for detailed explanation we refer to [6].

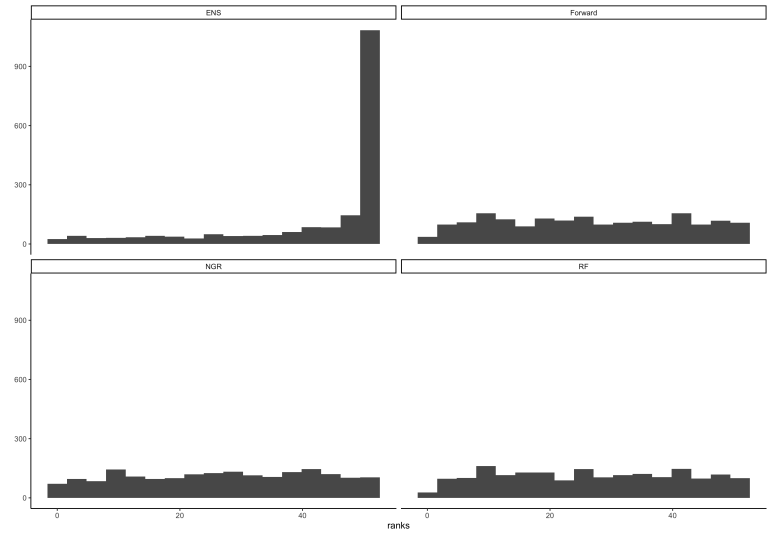


Figure 4.6: Rank histograms for lead time 60 h and station De Bilt. For forward selection, NGR, random forest and the raw ensemble forecast

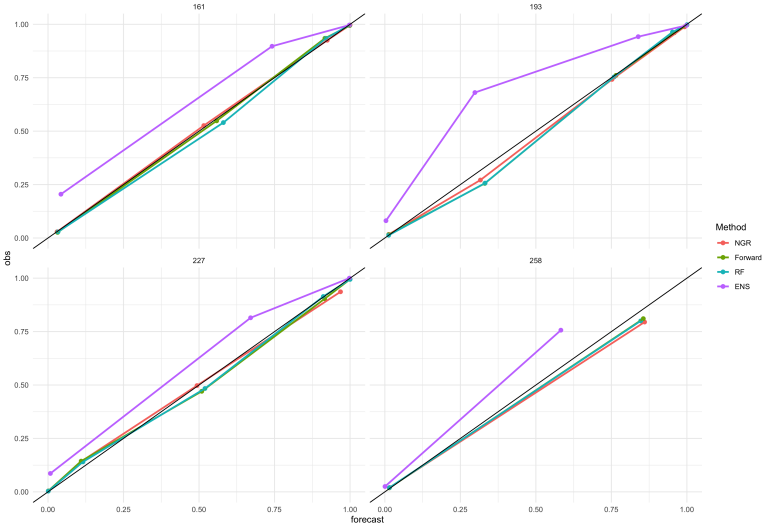


Figure 4.7: Reliability diagrams with thresholds equal to the 0.25, 0.5, 0.75 and 0.9 observational quantiles for lead time 60 and station De Bilt. Methods compared are: forward selection, NGR, random forest and the raw ensemble forecast

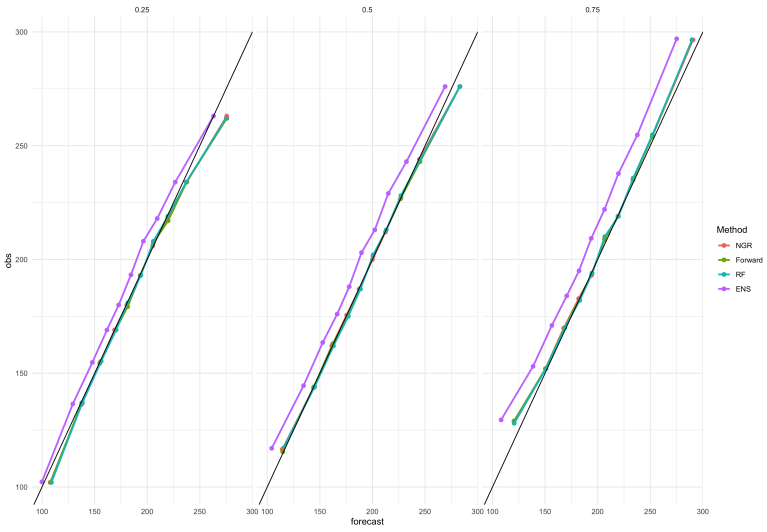


Figure 4.8: Quantile reliability diagrams for quantile levels equal to 0.25, 0.5 and 0.75 for lead time 60 and station De Bilt. Methods compared are: forward selection, NGR, random forest and the raw ensemble forecast

5

CONCLUSION

In this thesis, we discuss two important aspects in the topic of statistical post-processing for weather forecast. First, we address the problem of forecasting weather events by estimating the tail of the forecast distribution. Second, we investigate the problem of variable selection within post-processing when a large set of potential predictors are available. In this conclusion section, we summarize the results and discuss possible directions for future research for both topics.

In order to forecast extreme weather events we estimate the quantiles for the forecast distribution $Q_{Y|X}(\tau)$ for probabilities τ close to 1. For Y, the weather phenomenon of interest, and a covariate set X which can contain the uncalibrated forecast and additional predictor variables.

In this thesis, we proposed two ways of estimating the extreme conditional quantile curve $x \mapsto Q_{Y|X}(\tau|x)$. Both methods are fitted in a two-step procedure. In the first step, an intermediate quantile is estimated using all the data. This serves as the 'threshold' for classifying the extreme observations. In the second step, the extreme quantile is extrapolated from the intermediate quantile using the extreme observations.

Chapter 2 proposes the common shaped tail (CST) estimator. The underlying statistical model assumes quantile curves for τ within close to 1 are equidistant, i.e. the distance between two curves at a point x is the same for all values x. Within the estimator the intermediate conditional quantile is estimated using local linear quantile regression. This allows a flexible nonparametric estimation of the threshold instead of the conventionally fixed threshold. Subsequently, the extreme quantile is estimated by applying a Weissman type estimator on the exceedances of the intermediate quantile.

Through proving uniform consistency for the local linear quantile curve we show that the estimators for the extreme value index and the extreme quantile are asymptotically consistent and normally distributed. An extensive simulation study shows that, given the model assumptions, the estimator outperforms an extreme quantile method assuming linear quantile curves. In the case of deviations from the model assumptions, the performance is slightly worse.

In Chapter 3 we proposes a different method for extreme quantile regression which

we call gradient boosting for extremes (GBEX). Here the intermediate conditional quantile is estimated using quantile random forests, while the extreme conditional quantile is estimated using a gradient boosting approach. This methodology is much more flexible than the CST estimator due to its data-driven nature. Through the tree-like structure of the boosting procedure, the GBEX estimator is also able to fit non-monotone quantile curves, without assuming equidistant quantile curves, while at the same time retaining good predictive power in relatively high dimensional predictor space.

The flexibility and efficiency of the GBEX methodology is confirmed in an extensive simulation study where, in the presence of many noise variables, the estimator out performs both conventional machine learning approaches as well as standard extreme quantile estimation methods.

Both the CST and the GBEX methods are applied to a dataset of ECMWF precipitation forecasts and observations at several stations throughout the Netherlands. The CST is applied on only the summer data as they could be assumed to be heavy tailed. It outperforms the linear extreme quantile estimator for the upper ensemble member $\tau = 51/52$ based on a quantile verification skill score.

The GBEX method is applied to data of the entire year and is able to capture the seasonality in the data in terms of the tail parameters. This seasonality shows that summer precipitation has a heavier tail compared to winter precipitation. This is due to the local convective events that are observed in summer that are very uncertain, but could lead to large shower events.

The estimation of tail quantiles are subject to high uncertainties, because the number of extreme observations is by definition small. Additionally, most information concerning tail behaviour is captured in the very largest observations. This means that, especially in situations where heavy-tailed distributions play a role, removing just the single largest observation can change the tail estimation substantially.

As standard machine learning models are fully data driven they will quickly start over fitting by focussing on the largest observation. In order to adapt these machine learning methods to the problem of tail quantile estimation they need to be regularized heavily.

The GBEX method is one of the first approaches that adapts a machine learning model, gradient boosting, to obtain robust and consistent estimations with less risk of over fitting. By setting an upper bound on iterative updates of the algorithm, heavy regularization is applied such that no single iteration can have too high influence. At the same time, rigorous sub sampling and additional regularization of the weak learners lead to consistent results for the tail quantiles.

Gradient boosting is very suitable for this estimation problem by its ability to regularize the estimation in many different ways. This strong regularization allows the model to learn the structure in the data slowly without allowing single extreme events to dominate all the estimates.

Future research in this method is required to get a better understanding in choosing the tuning parameters. The current model is sensitive to the large number of tuning parameters. A reduction of these parameters or a clear strategy to choose them is needed to make the method easily applicable in other domains of research. A specific tuning parameter that would be interesting for further investigation is the ratio between learning rates for γ and σ. Although the tail quantiles do not strongly depend on this ratio, the

specific fitted values of γ and σ do change. The main reason for this is that both parameters influence each other in the estimation procedure. The optimal choice of this ratio is problem dependent and is important for good consistent results. Finally, more theoretical theory is needed for general gradient boosting methods in order to better understand how the method can be applied to the estimation of tail quantiles.

In Chapter 4 we discuss the problem of variable selection within a post-processing framework. This problem is interesting for three reasons. First, there is an extremely large number of potential features coming from the NWP models that can be used in the post-processing models. The strong correlations between them make the interpretability of the model difficult. Secondly, within post-processing the focus is on probabilistic forecasts, which makes it important to select features that are informative for the entire probabilistic forecast and not only the conditional mean. And thirdly, selected features need to have predictive skill such that the resulting models have strong predictive skill.

We propose a methodology for random forest to select variables that provide predictive skill for the entire forecast distribution. The methodology is a stepwise procedure that selects in each step the feature that reduces the predictive performance of the model most. The predictive performance is calculated using the continuous ranked probability score, which allows us to include the entire forecast distribution in selecting the best variable at each step.

In order to be able to estimate the predictive performance we use the randomly left out samples that are not used in fitting trees in the random forest bootstrap procedure to compute the performance. These out-of-bag samples allow us to approximate predictive performance without a cross-validation set-up or by splitting data in train and test sets beforehand.

As the number of features can become very large we developed an early stopping method that checks iterative improvement of the predictive performance of adding an additional variable to the model. At each step we test the null hypothesis of non-decreasing predictive performance. When the predictive performance stagnates and we can not significantly reject the null hypothesis of non decreasing predictive performance we stop selecting variables. The resulting model is a model of minimal size but maximal predictive performance.

We show in a population sense that under an independence assumption between covariates that the selected variables are form an optimal set with respect to the continuous ranked probability score. Also we show under regularity conditions that the testing procedure is consistent.

In an extensive simulation study we show that our variable selection method, compares favourably to a backward selection method based on a permutation importance measure. Our proposed method is more capable of discriminating between signal variables and noise variables. We investigate in a simulation study many different models, where we vary the signal strength and the correlation structure of the predictors. Additionally, we explore how sample size influences these results. In all cases our methodology outperforms the competing methods and provides even in the presence of high correlations satisfactory results.

In an application on post-processing maximum temperature, our method shows consistency in the number of selected variables and in the variables being selected over sev-

eral stations. Moreover, our method selects less than 10 percent of the covariates and still attains similar predictive power as the quantile random forest with all covariates. Further, it is easier to interpret how a single variable contributes to the estimated model. This becomes much harder in the presence of many correlated predictors. Due to correlation, the contribution of a single variable is split over several correlated variables. In our data example, in the presence of thick cloud cover, the fitted random forest model indicates that there is a higher risk of over forecasting (lower panel of Figure 4.5) instead of under-forecasting which was indicated by Figure 4.2.

At this stage, two interesting directions for future research can be identified. First, the theoretical results in Sections 4.2 and 4.3 are derived under the assumption that the covariates are independent. However, the ability of our method to select signal variables from a correlated setting is evidenced by our simulation study and data application. It is interesting to investigate the theoretical properties of the method in the setting with dependent covariates. Second, we focus in this paper on how this forward method behaves for the continuous ranked probability score. Hereby, we are trying to select variables that are predictive for the entire forecast distribution. The mathematical set-up in Chapter 4.2 is much more general and allows to select variables with respect to other loss functions or even a general set of loss functions. For example the quantile loss function could be used to select variables that are predictive for a specific forecast quantile. Weighted loss functions that focus on extreme values could potentially help selecting variables that are predictive for extreme events.

ACKNOWLEDGEMENTS

This thesis would not be possible if not for the help of many people. Only with their encouragement and company was I able to overcome the struggles, complete this manuscript and create fond memories during my PhD journey. This section is dedicated to them.

This thesis is part of the research project “Probabilistic forecasts of extreme weather utilizing advanced methods from extreme value theory" with project number 14612 which is financed by the Netherlands Organisation for Scientific Research (NWO).

First, I want to express my sincere gratitude to my TU-Delft supervisors, Geurt Jongbloed and Juan-Juan Cai. My passions for statistics and data are because of you. After the first lecture of Introduction to Statistics, taught by Geurt, I remember being thrilled how simple numbers can become insightful using statistics. A few years later, during my bachelor thesis, Juan showed me the amazing strength of extreme value statistics – that we can better understand the events which are rarely, if ever, observed. That excitement has never left me and has resulted in this thesis.

Juan, thank you for giving me the freedom to pursue my own ideas, but always having time to help me when I encountered an obstacle. Geurt, thank you for teaching me how to break down problems and showing me the importance of describing ideas in precise and simple language.

Next, I would like to thank my KNMI supervisors Maurice Schmeits and Kiri Whan. When I joined this project, I did not have any background in meteorology and statistical post-processing. Maurice, thank you for integrating me in the statistical post-processing group and the many sit-downs to provide me with the meteorological perspectives on our work. Kiri, thank you for the many friendly discussions and brainstorming sessions of new ideas.

Upon its completion, my thesis was examined by committee members Annoesjka Cabo, Valerie Chavez-Demoulin, Philippe Naveau, Chen Zhou and Arnold Heemink. I would like to thank them for carefully reading the manuscripts and providing useful comments.

Thank you to Clément Dombry and Sebastian Engelke for a fruitful collaboration on one chapter of this thesis. Special thanks to Clement for hosting me in Besançon during my research visit, which I thoroughly enjoyed.

This research project consists of a user committee, whose members I would like to acknowledge. Many thanks to Etienne, Pier, Kees, Brian and Dirk for their insights into the real-world problems they deal with on a day-to-day basis.

I would like to thank Kate Saunders for the many, sometimes heated, discussions and the many pointers on my coding. I am a better programmer because of it.

I have spent 11 years at Delft Institute of Applied Mathematics as a bachelor, master and PhD student where I was taught everything I know by all the professors here. Thank you all for being great teachers and for your guidance on my mathematical journey. In particular from Statistics and Probability, Jacob, Joris, Rik, Frank, Nestor, Robbert,

Alessandra, André, Cor, Frank R., Dorota, Ludolf and Tina, thank you for all the interesting conversations at the coffee machine and during lunch, it was great to be at such a lively department.

The greatest part of my time I have spent amongst the PhD colleagues in the department from Statistics and Probability. To all my PhD (and post-doc) colleagues, Andrea B. Andrea F., Bart, Birbal, Dan, Eliza, Eni, Francesca, Inoni, Larissa, Lixue, Lörinc, Marc, Mario, Martina, Rik, Sebastiano and Simone. Thank you for all the discussions, coffees, lunches, and great conversations. Special thanks to Martina for being the heart of our group and for tolerating me before the third coffee in the early morning.

Thanks to everyone at KNMI for the many interesting conversations and the opportunities to read some great statistics books together.

Thanks to the secretaries and support staffs at DIAM. Thank you for taking care of all my non-research related problems.

On a personal note, thank you, Sven, Laura, Emma, Zaza, Marieke, Irene, Richard, Jelle, Niels and Paul, for being there when frustrations were high and providing with the needed social distractions.

To my family. Thank you, Mom and Dad, for always believing in me, always listening to me and for providing me with a shoulder when needed. Thank you, Marjolein and Yvo, for always understanding what I am going through and giving great advice.

Finally, I am most grateful to my wife Phyllis, for your endless support through many thesis struggles, for your patience and for always making time to proofread bits of this thesis.

REFERENCES

[1] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[2] A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.

[3] Sándor Baran and Sebastian Lerch. Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141(691):2289–2299, 2015.

[4] Jan Beirlant, Tertius De Wet, and Yuri Goegebeur. Nonparametric estimation of extreme conditional quantiles. *Journal of Statistical Computation and Simulation*, 74(8):567–580, 2004.

[5] Sabrina Bentzien and Petra Friederichs. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution nwp model cosmo-de. *Weather and Forecasting*, 27(4):988–1002, 2012.

[6] Sabrina Bentzien and Petra Friederichs. Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1924–1934, 2014.

[7] G. Biau and B. Cadre. Optimization by gradient boosting. In A. Daouia and A. Ruiz-Gazen, editors, *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*, pages 23–44. Springer International Publishing, 2021.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] T.A. Buishand, L. de Haan, and C. Zhou. On spatial extremes: with application to a rainfall problem. *Ann. Appl. Stat.*, 2:624–642, 2008.

[10] V. Chavez-Demoulin and A. Davison. Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society, series C*, 54, 2005.

[11] V. Chernozhukov. Extremal quantile regression. *Ann. Statist.*, 33(2):806–839, 2005.

[12] S.G. Coles and J.A. Tawn. Modelling extremes of the areal rainfall process. *J. R. Stat. Soc. Ser. B.*, 58:329–347, 1996.

[13] Abdelaati Daouia, Laurent Gardes, and Stéphane Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589, 2013.

[14] Abdelaati Daouia, Laurent Gardes, Stéphane Girard, and Alexandre Lekina. Kernel estimators of extreme level curves. *Test*, 20(2):311–333, 2011.

[15] A. Davison and R. Smith. Models for exceedances over high threshold. *Journal of the Royal Statistical Society, series B*, 52, 1990.

[16] Anthony C Davison and Richard L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442, 1990.

[17] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[18] Holger Drees, Anja Janßen, Sidney I Resnick, and Tiandong Wang. On a minimum distance procedure for threshold selection in tail analysis. *SIAM Journal on Mathematics of Data Science*, 2(1):75–102, 2020.

[19] Debbie J. Dupuis. Exceedances over high thresholds: A guide to threshold selection. *Extremes*, 1(3):251–261, 1999.

[20] Sébastien Farkas, Olivier Lopez, and Maud Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021.

[21] Eric W. Fox, Ryan A. Hill, Scott G. Leibowitz, Anthony R. Olsen, Darren J. Thornbrugh, and Marc H. Weber. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, 189(7):316, 2017.

[22] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[23] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[24] L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions: an application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204, 2010.

[25] Laurent Gardes, Stéphane Girard, and Alexandre Lekina. Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101(2):419–433, 2010.

[26] Laurent Gardes and Gilles Stupfler. An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT*, 17(1):109–144, 2019.

[27] Laurent Gardes and Gilles Stupfler. An integrated functional Weissman estimator for conditional extreme quantiles. *Revstat Statistical Journal*, forthcoming.

[28] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[29] Harry R Glahn and Dale A Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, 11(8):1203–1211, 1972.

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

[31] Tilmann Gneiting, Adrian E Raftery, Anton H Westveld III, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

[32] Yuri Goegebeur, Armelle Guillou, and Michael Osmann. A local moment type estimator for the extreme value index in regression with random covariates. *Canadian Journal of Statistics*, 42(3):487–507, 2014.

[33] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[35] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

[36] William H Klein, Billy M Lewis, and Isadore Enger. Objective prediction of five-day mean temperatures during winter. *Journal of Atmospheric Sciences*, 16(6):672–682, 1959.

[37] Roger Koenker. *Quantile regression*. Cambridge university press, 2005.

[38] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[39] Efang Kong, Oliver Linton, and Yingcun Xia. Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory*, 26(5):1529–1564, 2010.

[40] S Kruizinga. Objective classification of daily 500 mbar patterns. In *Preprints sixth conference on probability and statistics in atmospheric sciences*, volume 9, page 12. American Meterological Society Boston, MA, 1979.

[41] Francesco Laio and Stefania Tamea. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277, 2007.

[42] Sebastian Lerch and Thordis L Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):21206, 2013.

[43] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 431–439. Curran Associates, Inc., 2013.

[44] Carlos Martins-Filho, Feng Yao, and Maximo Torero. Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory*, 34(1):23–67, 2018.

[45] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[46] Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*, volume 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[47] James Pickands III. Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1):119–131, 01 1975.

[48] Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018.

[49] Alfréd Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4(3-4):191–231, 1953.

[50] G. Ridgeway. Generalized boosting models: a guide to the gbm package, 2007. URL https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

[51] Michael Scheuerer. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1086–1096, 2014.

[52] Richard L. Smith. Estimating tails of probability distributions. *Ann. Stat.*, 15:1174–1207, 1987.

[53] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

[54] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[55] Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34(3):617–634, 2019.

[56] Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

[57] Thordis L Thorarinsdottir and Tilmann Gneiting. Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):371–388, 2010.

[58] Simon Veldkamp, Kirien Whan, Sjoerd Dirksen, and Maurice Schmeits. Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149(4):1141–1152, 2021.

[59] Jasper Velthoen, Juan-Juan Cai, Geurt Jongbloed, and Maurice Schmeits. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019.

[60] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[61] Hansheng Wang and Chih-Ling Tsai. Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240, 2009.

[62] Huixia Judy Wang and Deyuan Li. Estimation of extreme conditional quantiles through power transformation. *Journal of the American Statistical Association*, 108(503):1062–1074, 2013.

[63] Huixia Judy Wang, Deyuan Li, and Xuming He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464, 2012.

[64] Kirien Whan and Maurice Schmeits. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Monthly Weather Review*, 146(11):3651–3673, 2018.

[65] Daniel S Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, 2011.

[66] Benjamin D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019.

[67] Keming Yu and MC Jones. Local linear quantile regression. *Journal of the American statistical Association*, 93(441):228–237, 1998.

CURRICULUM VITÆ

Jasper Jonathan VELTHOEN

30-07-1993 Born in Driebergen-Rijsenburg, The Netherlands.

EDUCATION

2005–2011	High School Christelijk Lyceum Zeist
2011–2015	Bachelor in Applied Mathematics Delft University of Technology
2015–2016	Master in Applied Mathematics Delft University of Technology
2021	PhD. Statistics Delft University of Technology
Thesis:	Statistical Post-processing for extreme weather forecasts
Promotor:	Prof. dr. G. Jongbloed

LIST OF PUBLICATIONS

3. **J. Velthoen, C. Dombry, J.J. Cai, S. Engelke**, *Gradient boosting for extreme quantile regression*, arXiv preprint arXiv:2005.05113.

2. **J. Velthoen, J.J. Cai, G. Jongbloed**, *Forward variable selection for random forest models*, To be published in Journal of Applied Statistics.

1. **J. Velthoen, J.J. Cai, G. Jongbloed, M. Schmeits**, *Improving precipitation forecasts using extreme quantile regression*, Extremes **22**, 599 (2019).