Delft University of Technology
Faculty Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics
Department of Statistics

Tail characteristics of CRPS-based distributions

Report for the purpose of
Delft Institute of Applied Mathematics
as part of achieving

the degree of

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

JEROEN ROSEBOOM
4222679

Delft, Netherlands
January 2022

TU Delft

MSc report APPLIED MATHEMATICS

"Tail characteristics of CRPS-based distributions"

JEROEN ROSEBOOM

Delft University of Technology

Thesis defence at 14th of January 2022

Supervisor

Dr. P. Chen

Other committee members

Dr. D. Kurowicka	Dr. J. Cai	Dr. J. Velthoen
...	...	...

January, 2022 Delft

Preface

The following document, about tail characteristics of CRPS-based distributions, is the final stage of my studies to acquire the master degree in Applied Mathematics at the Delft University of Technology. My thesis is conducted within the TU Delft in the Department of Statistics of the faculty EEMCS. My supervisor is dr. P. Chen, and dr. J. Cai and dr. J. Velthoen have been my supervisors in the beginning of my master thesis.

In my thesis I researched the potential paths and pitfalls of the newly created "Taillardat index". This index uses the tail characteristics of several CRPS-based distributions to rank forecasters on how well they forecast, with a slight emphasis on extreme events. From my research I concluded that the "Taillardat Index" in its current form is unstable and should be avoided. Even with theoretical changes, such as moving away from using p-values as a ranking method, the ideas behind the "Taillardat index" have to be handled with precaution. I tried to construct a new index based on the ideas of the paper by Taillardat et al. (2019) and the general theory of forecaster validation, to no avail. The findings of my endeavours can be found after the reflection of the Taillardat index and in the discussion.

I would like to thank my supervisors, Piao, Juan and Jasper, for their continued support throughout my thesis. For proof reading my thesis, helping me find errors in my code and being part of my committee, you helped me very much and I am very thankful for that. I would also like to thank Geurt Jongbloed, Benjamin Slevin and Hicham Zahri, for proof reading my thesis, your comments were very appreciated. I would also like to thank Dorota Kurowicka, for taking time to review my thesis and being part of my committee alongside Piao, Juan and Jasper. And finally I would like to thank my family and friends, for their mental support throughout my thesis route, it was needed and I am thankful for that.

My personal thesis route did not come without hurdles. Mental illnesses plagued me ever since I started on this journey, from problems with concentration, to the feeling of not being good enough. The corona pandemic did not ease these mental illnesses and during the pandemic I also lost both my grandparents who were still alive. Ultimately I made it through this dark forest and I am very happy to present you my final report. I would like to commemorate my grandparents with this final report, they have been a special part of my life and thought me valuable life lessons. May you never be forgotten.

If anyone still busy on their master thesis is reading this, you will make it through this eventually, you can do it!

Jeroen Roseboom, 6th of January 2022

Contents

Page

List of Figures **10**

List of Tables **11**

List of Notations **13**

1 Introduction **15**

2 Key theoretical concepts **23**
2.1 The basis of forecasting . . . 23
2.2 What makes a forecaster a good forecaster? . . . 26
2.2.1 Consistency . . . 26
2.2.2 Quality . . . 27
2.2.3 Value . . . 29
2.3 Continuous Ranked Probability Score . . . 30
2.4 Calibration . . . 31
2.5 The Probability Integral Transform . . . 32
2.5.1 Probabilistic calibration and the PIT . . . 33

3 Simulation studies **35**
3.1 Introducing the model . . . 35
3.2 The PIT . . . 36
3.2.1 PIT-histogram results . . . 37
3.2.2 Creating a special overdispersed forecaster . . . 38
3.2.3 The impact of dataset sizes . . . 39
3.3 Results: CRPS score . . . 40
3.4 Results: p-values, CRPS score -connection . . . 42

4 Extreme events and the CRPS **45**
4.1 Theoretical basis of the Taillardat index . . . 46
4.1.1 A CRPS-based tool . . . 46
4.1.2 The extreme value behaviour of the CRPS . . . 48
4.1.3 The index . . . 50
4.2 Simulation of the Taillardat index . . . 52
4.2.1 Our results . . . 53
4.3 Our reflection on the Taillardat index . . . 58
4.3.1 Reflection on the results . . . 58
4.3.2 Reflection on the fundamentals of the Taillardat index . . . 59
4.3.3 The shift towards extreme events . . . 60

4.4 Future improvement paths of the Taillardat index 61
4.4.1 The theory 61
4.4.2 The research and results 62

5 Conclusions and future research **67**
5.1 Conclusions 67
5.2 Secondary research paths results 67
5.2.1 Designing new models 67
5.2.2 A new weight function for the CRPS 70
5.3 Closure and recommendation for future research 72

Appendices **73**

A PIT-histograms simulation chapter **75**

B Histograms of uniformity p-values simulation chapter **77**

C γ- and σ-estimates **83**
C.1 Difference between γ-estimates 85
C.2 Correlation figure full size 86
C.3 Index percentages 87
C.4 Index percentages 88
C.5 Index differences 89

D Uniform-Pareto model, climatological tail index **91**

E Beta-Pareto model, climatological tail index **93**

F Results new weight functions on NN-model **95**

List of Figures

1.1 Climatological probability density of the average daily temperature of an arbitrary day in June in the Netherlands in $°C$ (fictional density for sake of an easy illustration). . . . 17
1.2 Left: forecaster X (low resolution), right: forecaster Y (higher resolution). 17
1.3 Flowchart of the research. . . . 21

2.1 Climatological probability density of the average daily temperature in January in the Netherlands in $°C$ (fictional density for sake of easy illustration). . . . 24
2.2 Three distinct forecasts of the average daily temperature of tomorrow in January in the Netherlands in $°C$ (fictional forecasts for sake of illustration). . . . 25
2.3 On the left we see the issued forecast and the observed observation, while on the right we see the calculated CRPS score, $(F_t(x) - \mathbb{1}\{x \geq y_t\})$ in black and $(F_t(x) - \mathbb{1}\{x \geq y_t\})^2$ in red, of the forecast-observation pair. . . . 30

3.1 Biased PIT-histogram. . . . 37
3.2 Overdispersed PIT-histogram. . . . 37
3.3 Sign-Biased PIT-histogram. . . . 38
3.4 Hamill's Unfocused PIT-histogram. . . . 38
3.5 PIT-histogram of the overdispersed sign-biased forecaster, on a dataset of 100 000. 39
3.6 Distribution of the CRPS scores of the perfect forecaster, based on the p-values of the Anderson-Darling uniformity test. . . . 42
3.7 Distribution of the CRPS scores of the climatological forecaster, based on the p-values of the Cramér-von Mises uniformity test. . . . 43
3.8 Distribution of the CRPS scores of the perfect forecaster, based on the p-values of the Kolmogorov-Smirnov uniformity test. . . . 44
3.9 Distribution of the CRPS scores of Hamill's forecaster, based on the p-values of the Anderson-Darling uniformity test. . . . 44

4.1 γ and σ estimations at the 0.75-quantile. . . . 54
4.2 Correlation γ-estimates using obs. or CRPS values at the different values of u. . 55
4.3 Index values of the perfect and climatological forecaster using u is the 0.75-quantile and 10 000 forecast-observation pairs. . . . 56
4.4 Course of the standard deviation of the mean index value, depending on the number of shuffles used (perfect forecaster). . . . 63

5.1 Probability density functions of the tail index of the perfect forecaster when using the (red) beta(1,5), (blue) beta(5,1) or the (yellow) uniform distribution model, rescaled to the [0,0.25]-range. . . . 69

A.1 Perfect PIT-histogram. . . . 75

A.2 Climatological PIT-histogram. 75
A.3 Unfocused PIT-histogram. 76
A.4 Hamill PIT-histogram. 76

B.1 Histograms of p-values of the perfect forecaster. 78
B.2 Histograms of p-values of the climatological forecaster. 79
B.3 Histograms of p-values of Hamill's unfocused forecaster. 80
B.4 Histograms of p-values of Hamill's forecaster. 81

C.1 γ-estimates using the 0.75-quantile for u. 83
C.2 γ-estimates using the 0.90-quantile for u. 83
C.3 γ-estimates using the 0.95-quantile for u. 84
C.4 Differences in γ-estimates at the 0.75-quantile. 85
C.5 Differences in γ-estimates at the 0.90-quantile. 85
C.6 Differences in γ-estimates at the 0.95-quantile. 85
C.7 Correlation γ-estimates using obs. or CRPS values at the different values of u. . 86
C.8 Course of the standard deviation of the mean index value, depending on the number of shuffles used (unfocused forecaster). 88
C.9 Course of the standard deviation of the mean index value, depending on the number of shuffles used (extremist forecaster). 88
C.10 The index values of the three forecasters. 89
C.11 Difference in index value between the perfect forecaster and the unfocused forecaster. 89
C.12 Difference between the index value of the extremist forecaster and the element wise maximum of the index values of the perfect forecaster and the unfocused forecaster. 90

List of Tables

1 List of Notations. 13

3.1 The Normal-Normal model. 36
3.2 Percentage of cases where the test statistic rejects the uniformity of the PIT-histogram (percentages are based on 100 000 simulations). 40
3.3 Mean and standard deviation of the CRPS score of all forecasters of the Normal-Normal model (taken over 100 simulations). 41

4.1 The Gamma-Exponential model. 48
4.2 The simplified Gamma-Exponential model. 52
4.3 Percentages of the index values of the climatological forecaster equal to 0. 57

5.1 The Uniform-Pareto model. 68
5.2 The $B(5,1)$-Pareto model. 68
5.3 The $B(1,5)$-Pareto model. 68

C.1 Percentages of the index values of the climatological forecaster equal to 0. 87

F.1 The Normal-Normal model. 95
F.2 Mean scores of forecasters using different weight functions (the order of the forecasters in this table is according to their ranking using no weight function). . . . 96
F.3 Ranking order of forecasters using different weight functions. 96
F.4 Mean percentile change between the original CRPS value and the CRPS value using a different weight functions. 97

List of Notations

To ensure our notation is clear we compiled a list of notations here, which features the notations used in our research.

t	timestep or specific point in time
y_t	observation at timestep t
Y_t	random variable at timestep t, from which y_t is generated
Δ_t	information at timestep t
H_t	the true cdf of $Y_t \vert \Delta_t$
G	the climatological cdf of Y, G is also the true cdf of Y_t if there is no information
f_t	the forecast issued at timestep t, which is a pdf
j_t	the forecaster's own best judgement at timestep t
F_t	the cdf of the forecaster at timestep t, so $F_t(x) = \int_{-\infty}^{x} f_t(u)\ du$
$s(\dots)$	a scoring function s
$\mathcal{F}$	a forecasting system, which issues forecasts
F_t^P	the forecast of the perfect forecaster at timestep t. More generally, the superscript indicates which forecaster is meant, P for perfect forecaster here

Table 1: List of Notations.

Note 1: G and G_t are the same since G is invariable over time given a stable climate.

Note 2: we use two different subscripts, $_i$ and $_t$. The difference between subscripts $_i$ and $_t$ is similar to the difference between permutations and combinations. When using $_t$, the moment in time is of importance, while using $_i$, the moment in time is irrelevant.

Example 1: $CRPS(F_t, y_t)$ is the CRPS value of the forecast-observation pair at timestep t.

Example 2: $Y \vert (\mathcal{F} = F_i)$ is the set of observations given that the (cdf of the) forecast issued beforehand is equal to F_i.

If $_t$ was used in example 2, the incorrect assumption can be made that only the observation at timestep t is part of this set. However, if the same forecast is issued at two different timesteps, $F_{t_1} = F_{t_2}$, $t_1 \neq t_2$, then $Y \vert (\mathcal{F} = F_{t_1})$ contains both y_{t_1} and y_{t_2} in our research. To avoid misconceptions, the subscript $_i$ is used in these cases.

Note 3: we use the point as the decimal separator (e.g. 3.5) and we use a brief space between thousands to ease the reading of those numbers (e.g. 2 492 384).

Note 4: in literature, both "scoring rule(s)" and "scoring function(s)" are used for the exact same concept. In our research we will call these scoring functions, even when quoting from a source which uses the term "scoring rules".

Chapter 1

Introduction

Forecasting has a rich history with the first recorded weather predictions starting in 1861 (Moore (2015)), although we can suspect that undocumented or spoken predictions predate these written predictions by many years. The early forecasts were deterministic point forecasters like, 'tomorrow it will be $18°C$', or deterministic categorical forecasters like, 'tomorrow it will be cloudy'. Over the years, the models generating the forecasts became more and more sophisticated and the number of different forecasters grew as well. The increase in the number of forecasters and the increasement of their sophistication caused an increasing interest in the question "which forecaster is the best forecaster?". This question sparked the field of forecaster validation, whose goal is to answer the aforementioned question.
Lately, the scientific community gained an increased interest in the correct forecasting of extreme events, due to their societal impact and the increased interest by the public and the media in these events. In the recent paper by Taillardat et al. (2019), a new and interesting index has been constructed which tries to validate forecasters on their ability to accurately forecast extreme events while not disregarding their overall quality. Our research tries to understand this index and investigates its potential paths and pitfalls. In the following part of this introduction we will introduce the aspects of the field of forecaster validation that we will use in our research, followed by a reading guide with a detailed overview of our research by chapter.

In 1987, Murphy & Winkler (1987) set up a framework for forecaster validation based on the joint distribution of forecasts and observations. In their paper, Murphy and Winkler state that the joint distribution of forecasts and observations contains all relevant non-temporal information to evaluate the quality of a forecaster.
Murphy (1993) built upon the aforementioned framework and introduced his three "goodnesses-of-fit". His proposed goodnesses-of-fit are called: *consistency*, *quality* and *value*.
Consistency can be described as the match between a forecaster's own best judgement and the forecast which was actually issued. In case of a perfect match, total consistency is achieved.
Murphy defines a forecaster's own best judgement and a forecaster's forecast as follows:
He assumes that a forecaster derives its *forecast* about a future event from a knowledge base which can consist of information from different sources. He furthermore assumes that the forecasting process ends up in the formulation of a *judgement* the future event. He defines these *judgements* as internal, "recorded only in the forecaster's mind", and that they are not yet an external statement, or *forecast*.
The forecaster's own best judgement can vary from the forecaster's forecast when the entity in control of issuing the forecasts alters their forecasts to, for example, achieve a better evaluation of their forecaster or when the forecast's "spatial or temporal specificity" differs from that of the judgement of the forecaster. The altering of forecasts to achieve a better evaluation is called

hedging (Jolliffe (2008), Lerch et al. (2017)), which can be countered by *proper* scoring functions (Murphy (1993)) which we will investigate in further detail in chapter 2.
Quality is the match between the forecasts and their corresponding observations. Due to the underlying framework of joint distributions by Murphy & Winkler (1987), quality becomes naturally multifaceted and therefore Murphy subdivided quality into a set of aspects of quality.
The third and final goodness-of-fit is *value*, which is the amount of benefit a forecaster contributes to its user.
The majority of scientific papers focus on Murphy's first and second goodness-of-fit, since they are more theoretical in nature. Value, on the other hand, rarely receives extensive research and is mostly only coined for completeness of the theory.
In our research, we focus on *consistency* and *quality*. However, we will come back to *value* in our recommendations.

Over the years, increasingly better forecasters and measurements steadily lowered the uncertainty[1] surrounding forecasts. However, they could never fully eliminate them.
Murphy & Winkler (1984) say that a drawback of the deterministic point forecasters and deterministic interval forecasters is that these forecasts come without margins of said uncertainty. As a result of the drawback, the user is unable to make a well-informed decision. If a forecast includes a margin of uncertainty, the user has an insight in the possible outcomes and the probabilities of the outcomes and therefore the user can make a more calculated decision. In order to quantify the uncertainty within the forecasts, forecasters shifted from the usage of deterministic (point) forecasters, such as "rain/no rain" or "$18°C$", progressively towards the usage of probabilistic forecasters, such as "20% chance of rain" (binary case) or a probability density function of the expected temperature (see for example Figure 1.2) [2]. Even though this shift has taken place, deterministic point forecasters have not completely vanished. In some practical situations, deterministic point forecasters still persist due to decision making, reporting requirements, or tradition, among others (Gneiting (2011), Ehm et al. (2016)). Dawid (1984) and Gneiting (2008) urge that forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities and events, in order to provide suitable measures of the uncertainty associated with the predictions.
In our research, we adhere to these urges and therefore most forecasts in our research take the form of probability density functions.

These new probabilistic models required new definitions of Murphy's aspects of *quality*, which originally were recorded for deterministic forecasting. In chapter 7 of the book of Stevenson et al. (2006), they investigate Murphy's quality aspects with respect to probabilistic forecasters on binary cases. They state that *reliability* and *resolution* are the two key aspects of quality for probabilistic forecasters. Here, *reliability* is the match between the probabilities given in a forecast and the observations occurring after these forecasts were made. *Resolution* is the ability of a forecaster to issue forecasts distinct from the climatological probabilities. In other words, *resolution* is the ability of a forecaster to "discriminate in advance between situations that lead to different observed events" Stevenson et al. (2006).
To illustrate what resolution is, let us artificially create the following example:
We want to forecast the average daily temperature of an arbitrary day in June in the Netherlands. Let us assume that the climatological probability density for the average daily temperature of an arbitrary day in June in the Netherlands are:

[1] Note that uncertainty here means the uncertainty in the forecast, not the uncertainty that Murphy describes as one of the 6 aspects of his *quality* goodness-of-fit.

[2] A more in-depth history about this change can be found in Murphy & Winkler (1984).

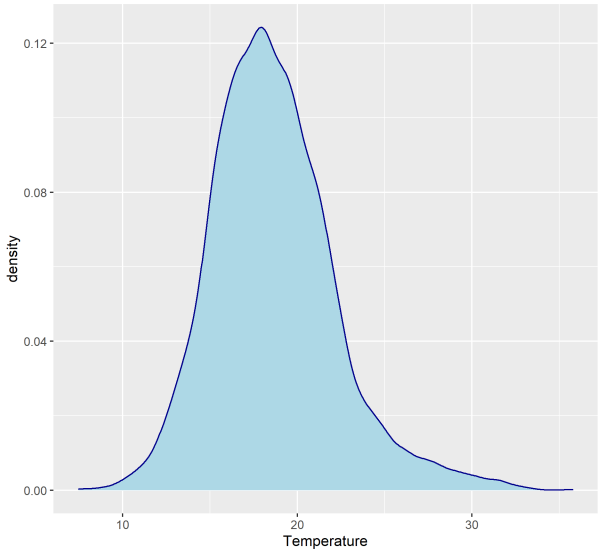

Figure 1.1: Climatological probability density of the average daily temperature of an arbitrary day in June in the Netherlands in $°C$ (fictional density for sake of an easy illustration).

We assume that we have two different forecasters, forecaster X and forecaster Y, which each try to forecast the average daily temperature of an arbitrary day in June in the Netherlands. For this example, we assume that we have 30 different sets of information and both forecasters, X and Y, each issue a forecast for each of these sets of information. This results for both forecasters, X and Y, in 3 distinct forecasts, forecast A, forecast B and forecast C. Each of these forecasts have been issued an equal number of times by the forecasters over the 30 different sets of information.[3] Forecaster X and Y are probabilistic forecasters, where the cumulative distribution of an interval can be calculated by taking the area under the density function of that interval. The forecasts of forecaster X and Y are the following:

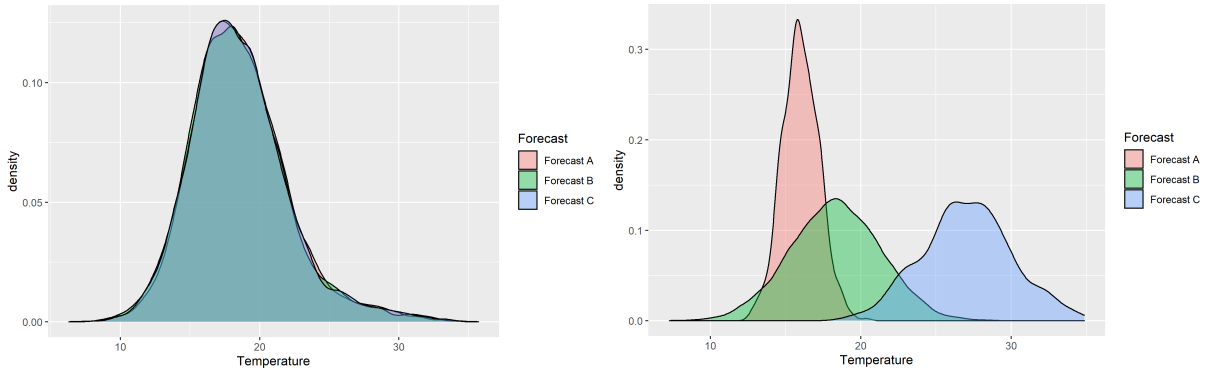

Figure 1.2: Left: forecaster X (low resolution), right: forecaster Y (higher resolution).

We see that the forecasts of forecaster X are all very close to the climatological density. Therefore, forecaster X has a low resolution. The forecasts of forecaster Y are more distant from the climatological density than the forecasts of forecaster X. Therefore, forecaster Y has a higher

[3]Meaning that forecaster X forecasts its forecast A, forecast B and forecast C each 10 times and the same holds for forecaster Y.

resolution than forecaster X. A high resolution is beneficial because it signifies that a forecaster can distinguish different cases into different forecasts. A forecaster can obtain a resolution which is too high, however this can be compensated with the reliability aspect, more on this can be found in the theory in chapter 2.

Stevenson et al. (2006) also state that the reliability of a forecaster can be improved by *calibration* and that if a forecaster is totally reliable, its resolution is identical to its *sharpness*.
Calibration in most papers is used synonymously with reliability, while *sharpness* can be defined as the average variance of the forecasts of a forecaster. Given the aforementioned example, forecast A of forecaster Y is much sharper than any of the other forecasts of forecaster X and/or forecaster Y. More broadly, the forecasts of forecaster Y are all sharper than the forecasts of forecaster X due to their densities being more concentrated, causing these forecasts to have less variance. Therefore, forecaster Y has more sharpness than forecaster X.
As a result of the aforementioned statements about reliability, resolution, calibration and sharpness, Gneiting et al. (2007) suggests to maximise for sharpness given some level of calibration.
We will study *calibration* and *sharpness* in further detail in our research and quantify what level of calibration is required for "Maximise sharpness given some level of calibration". Tsyplakov (2013) states that proper scoring functions provide the right balance between sharpness and calibration, meaning that these validation methods score forecasters accurately with regards to the aforementioned "Maximise sharpness, given calibration"-rule.

In our research we make a clear distinction between verification methods and validation methods.
Verification methods are used to verify if a forecaster does or does not possess a certain aspect or trait. An example of a verification method is checking if someone has a passport before they can enter a certain country.
Validation methods are used to validate how well a forecaster performs with regard to a certain attribute. An example of a validation method is a school test which grades the test takers with a score between 1 and 10.
Both verification methods and validation methods are part of forecaster validation, which is the umbrella term of validating forecasters. Each type of method, verification and validation, comes with its own benefits and shortcomings, which we will explore in our research. The methods we will use in our research are the Probability Integral Transform (PIT) (Dawid (1984), Rosenblatt (1952), Pearson (1933)) and the Continuous Ranked Probability Score (CRPS) (Matheson & Winkler (1976)). The PIT is a verification method, meaning that the PIT can verify if a forecaster possesses a certain specific aspect or trait, which is vital to verify the *given calibration*-part of the rule to "Maximise sharpness, given calibration". The CRPS is a validation method, meaning that the CRPS can be used to validate how good a forecaster is compared to other forecasters, which is vital for the *Maximise sharpness*-part of the rule to "Maximise sharpness, given calibration". We want to emphasize this difference, because verification methods cannot validate forecasters and validation methods cannot verify if a forecaster possesses a certain aspect or trait in most of the cases.

Recently, the important field of extreme event forecasting has become more popular in scientific papers (as mentioned in Gneiting & Katzfuss (2014), Friederichs & Thorarinsdottir (2012), Casati et al. (2008), Lerch et al. (2017) and Brehmer et al. (2019)). Two reasons for its importance are the impact that extreme events have and the attention given to extreme events by the public and the media. As a result, it has become imperative to adequately predict these extreme events. In the latter parts of our research we discovered that this shift towards forecasting ex-

treme events has a lot of ties to Murphy's goodness-of-fit *value*. Although we did not have time to include our own extensive research on *value*, we will provide some recommendations towards researching this goodness-of-fit, as it might help with the field of forecasting extreme events.

The main focus of our research is concentrated around an inspiring new paper by Taillardat et al. (2019). Their paper uses the field of Extreme Value Theory to look for tools that could benefit the research on extreme event forecasting. Moreover, their paper argues that the way we use the scores coming from validation methods can be enriched. In general, a validation method produces a value for every forecast-observation pair from a set of forecast-observation pairs and the final score of a forecaster is calculated by taking the mean over all these individual values. Taillardat et al. (2019) argues that the set of scores possesses much more information than its mean and uses the set of individual values to construct a new index which can be used to rank forecasters. This index has no name as of yet and so we will refer to this index as the "Taillardat index".
In their discussion, Taillardat et al. call upon the scientific community to "study the specific properties of this, their index (ed.), CRPS-based tool and its potential paths and pitfalls". Our original aim was to construct different models to test the Taillardat index on. However, we discovered that the Taillardat index is not as robust as their paper seems to suggest. Therefore our research focuses on exploring their questions, scrutinizing their steps, illuminating their index and studying their results. Our research reveals some structural weaknesses in their index, which makes their index unreliable[4] for general use.

Our research is documented in the following way.
In chapter 2, we will lay the theoretical basis for forecasting and forecaster validation. In this chapter, we will discuss Murphy's three goodnesses-of-fit.
For *consistency*, we will examine a phenomenon which can harm consistency and a counter measure against this phenomenon which gives rise to a property that most state-of-the-art score functions should possess.
For *quality*, we will mention the aspects of quality and extensively review the majority of them. Our review amends the definitions of reliability and resolution for deterministic forecasts with definitions of reliability and resolution for probabilistic forecasts, in order to increase our intuition when working on forecaster validation when using probabilistic forecasters in a practical setting. Furthermore, our review explains why only a few of the aspects of quality are needed to accurately gauge a forecaster's quality.
In our theoretical basis, we will only briefly mention *value*, because its potential was only discovered at the end of our research.
With the important aspects of quality in mind, we will introduce two methods: the Continuous Ranked Probability Score (CRPS) and the Probability Integral Transform (PIT), which we will use in our research to rank the available forecasters on their quality.
In chapter 3 we will conduct a simulation to show how the PIT and the CRPS work in practice. In this simulation we will show why the PIT can only be used to verify if forecasters possess a certain desired aspect and why the PIT is unable to validate forecasters, which is needed to accurately rank them among each other. For the PIT we will also expand a statement made in Hamill (2001), this statement says that "A U-shaped rank histogram, PIT-histogram (ed.), typically thought of as indicating undervariability in the ensemble, could also indicate that the ensemble is sampling a population with some combination of conditional biases.". In our simulation we will show that even overdispersed forecasters can generate a U-shaped PIT-histogram. Furthermore, we will shed more light on why Hamill's forecaster and Hamill's unfocused fore-

[4]Unreliable in the broad sense of the word, not affiliated with Murphy's aspect of quality called reliability.

caster from Hamill (2001) generate a uniform PIT-histogram, even though they theoretically should not be uniform. The reasons mentioned in Hamill's paper are somewhat limited, and we hope to clarify why the uniform PIT-histograms are generated and which fundamental part of the PIT is responsible for this occurrence.
For the CRPS we will show that caution should be exercised when using this validation method on its own. Most papers agree with Gneiting et al. (2007) to "Maximise sharpness, given calibration", however the CRPS on its own cannot rank the forecasters according to this rule. As shown in our simulation, some forecasters with a lower calibration do score better than others with a higher calibration. Therefore, verification methods have to be used in order to verify that all forecasters meet the *given calibration* requirement.
After understanding these methods in a general environment, we will use chapter 4 to shift towards ranking forecasters with an emphasis on extreme events. We will explain the initial problem which gave rise to the shift in the scientific community towards the emphasis on forecasting extreme events. Our main focus of this first part is to set the scene for a new and inspiring paper (Taillardat et al. (2019)).
We will discuss their theory which culminates in the "Taillardat index". The main part of this chapter consists of explaining their model and investigating their simulation and their model. After extensive research, we have found that their index has some structural flaws which we will explain in the reflection section of chapter 4. This section contains three reflection angles; a reflection on the results of the Taillardat index, a reflection on the fundamental building blocks of the Taillardat index, and a reflection of the shift in the scientific community towards the emphasis on forecasting extreme events. After the reflections, we will investigate another part of the paper by Taillardat et al., which could be used to construct a different, new, index. We constructed this new index and tested whether this index could be used for the original purpose of the Taillardat index. This is not the case according to our research.
Our final chapter lists our conclusions, after which we highlight some secondary research paths that we have conducted but which would not fit within our main research. In these secondary research paths we constructed new models which have some key fundamental differences compared to the model used in Taillardat et al. (2019). This research was conducted when we had our original aim, which was to study how the Taillardat index would fair under different fundamental circumstances. Due to the shift in the aim of our research, this research became secondary, however we expect that our research will be valuable for others who will continue working on this subject. The other secondary research path consists of different weight functions for the CRPS and how well these weight functions work on the practical model used in our simulation in chapter 3. We researched this topic because Tsyplakov (2013) stated that proper scoring functions score forecasters accurately with regards to the aforementioned "Maximise sharpness, given calibration"-rule. However, chapter 3 shows us that the strictly proper scoring function *CRPS* on its own does score forecasters without any calibration over some forecasters which possess some type of calibration, which is a counter argument to what Tsyplakov stated in its paper. All of these new weight functions have some deficit to them which causes them to rank at least one alternative forecaster over the perfect forecaster. We will close this chapter with an exhaustive list of recommendations.

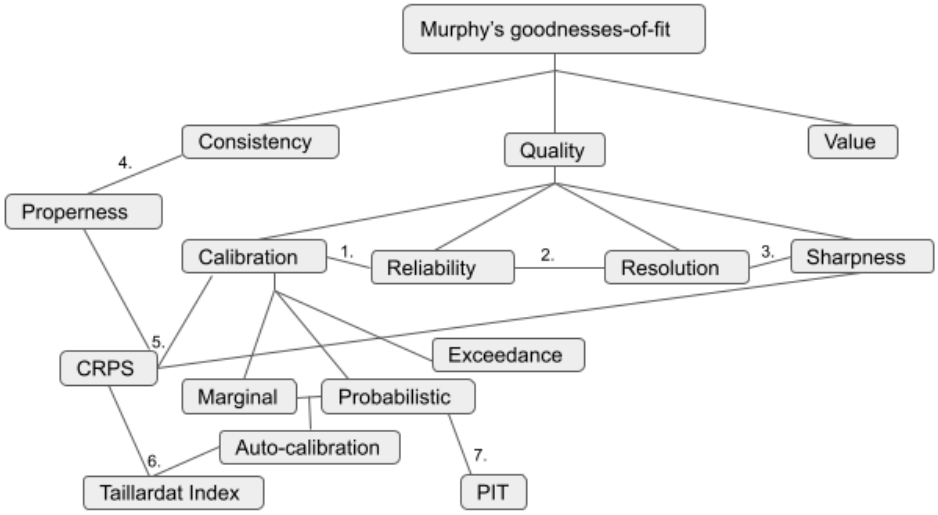

Figure 1.3: Flowchart of the research.

1. In most papers, calibration is used synonymously with reliability.
2. Reliability and resolution are coined by Stevenson et al. (2006) as the most important aspects of quality.
3. Given perfect reliability or calibration, resolution is equal to sharpness.
4. Consistency gives rise to properness, an important aspect of scoring functions.
5. The CRPS is a proper scoring function, which means it can be trisected into calibration, sharpness and uncertainty [5].
6. The Taillardat index is a CRPS-based index which requires auto-calibration from the forecasters for them to be eligible to be ranked by the index.
7. The PIT can be used to verify if a forecaster is probabilistically calibrated.

[5] Proof for the claim that all proper scoring functions can be decomposed into a reliability, resolution and uncertainty term, and therefore a calibration, sharpness and uncertainty term can be found in Bröcker (2009).

Chapter 2

Key theoretical concepts

The theoretical basis of forecasting and forecaster validation is quite extensive. In this chapter we will start with the basic concepts of forecasting and forecaster validation. We use Murphy's broad goodnesses-of-fit as a starting point of "what makes a forecaster a good forecaster" and slowly dig deeper into the theory surrounding these goodnesses-of-fit. Over the course of the chapter we will conceptualise which aspects of the goodnesses-of-fit are the most important ones, how we can determine if forecasters possess enough of these aspects and how we measure these aspects using the verification and validation methods we employ in our research.

2.1 The basis of forecasting

Forecasting is the process of predicting a quantity of interest. Examples of these quantities are the amount of rainfall tomorrow (in mm) and the peak windspeed of tomorrow (in km/h).
We can observe these quantities and we denote these observations by y_t for $t = 1, \ldots, T$, where t is a specific point in time. We assume that the observations are realised values of random variables $Y_t, \quad t = 1, \ldots, T$. These random variables can either be discrete or continuous. In our research we assume that observations are realisations of continuous random variables.
The set of observations, $(y_1, \ldots, y_T)$, can be seen as a random draw from the distribution of random variable Y, where Y is distributed $Y \sim G$, where G is a cumulative distribution function (cdf). G is known as the climatological distribution.
In practice, G is not the best prediction we can make. Let us re-use the example of the average daily temperature of an arbitrary day in June in the Netherlands which we used in the introduction. G is here the cdf corresponding to the climatological probability density visualised in Figure 1.1. Now let us assume that there is a heat wave in the Netherlands and that the past days had an average daily temperature of around $30°C$ and the meteorological status in and around the Netherlands does not seem to change in the upcoming days. One would assume that given this information, the forecast for the average daily temperature of an arbitrary day in June which is preceded by a heat wave would be closer to forecast C of forecaster Y as visualised in Figure 1.2.
Thus, given other quantities which we deem to have effect on the quantity of interest[1], we can make more accurate predictions than G. We call these 'other quantities' *information*, which we represent by the random variable Δ_t, for $t = 1, \ldots, T$, which we assume to be independent and identically distributed from each other. We assume that $Y_t|\Delta_t$ has a certain distribution, $Y_t|\Delta_t \sim H_t$, where H_t is a cdf and Δ_t is the full information that one can receive at t.
Note that G is not based upon the individual information Δ_t, $t = 1, \ldots, T$, but on the dis-

[1] Such as 'the amount and thickness of the clouds at night', which has effect on 'the temperature of tomorrow'.

tribution of the information. Therefore, the distribution G does not change depending on the available information at a given point in time. In our research we assume that the distribution of the information remains unchanged over time (a stable climate). Therefore, G has no subscript.

The observation y_t is a realisation of Y_t, which means that the most accurate prediction of the quantity of interest y_t is the distribution of Y_t, but the distribution of Y_t is unknown in practice. Therefore we predict the distribution of Y_t using a forecaster, $\mathcal{F}$. $\mathcal{F}$ is a mathematical model that, given the information Δ_t, issues a forecast f_t which predicts the distribution of Y_t to quantify a future observation y_t. There exists point forecasters, whose forecasts are a singular value, and probabilistic forecasters, whose output is a probability density function (pdf). In our study, we will only use probabilistic forecasters. f_t is therefore a pdf and we denote its cdf by F_t.
In order to showcase the differences between a (generic) forecaster, the perfect forecaster and the climatological forecaster, we have constructed the following example:

Example: a forecaster, the perfect forecaster and the climatological forecaster.
For this example we want to predict the temperature of tomorrow, which is an arbitrary day in January in the Netherlands. Let's assume that the pdf of the climate of the temperature of tomorrow in the Netherlands, if tomorrow is an arbitrary day in January, is the following:

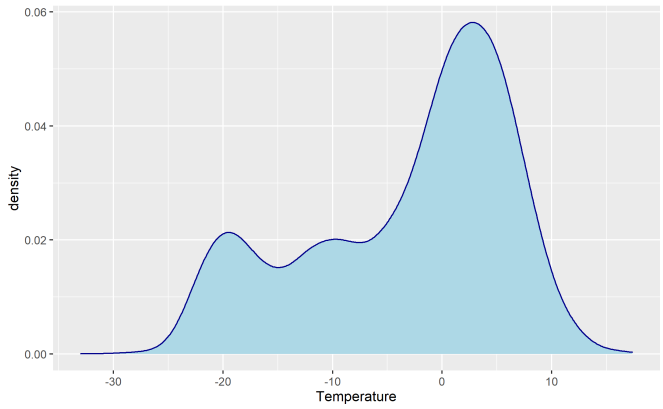

Figure 2.1: Climatological probability density of the average daily temperature in January in the Netherlands in $°C$ (fictional density for sake of easy illustration).

The cdf of the climatological probability density is equal to G and by definition the climatological forecaster forecasts the climatological probability density. Therefore we know the forecast of the climatological forecaster.
Let's assume that the previous couple of days were very cold, in the range of $-20°C$. Knowing this and more information, a forecaster $\mathcal{F}$ can forecast the average temperature of tomorrow with decreased uncertainty compared to the climatological forecaster. However, a forecaster $\mathcal{F}$ is not always able to perfectly forecast the temperature of tomorrow. For example, there might be information missing, there might be measuring errors in the information and/or the model might be slightly miscalibrated. Moreover, not every forecasting system will take into account the same information to issue its forecasts. Due to the imperfection of the information in practice, multiple forecasters can issue forecasts which can look very distinct. Take for example the following three forecasters:

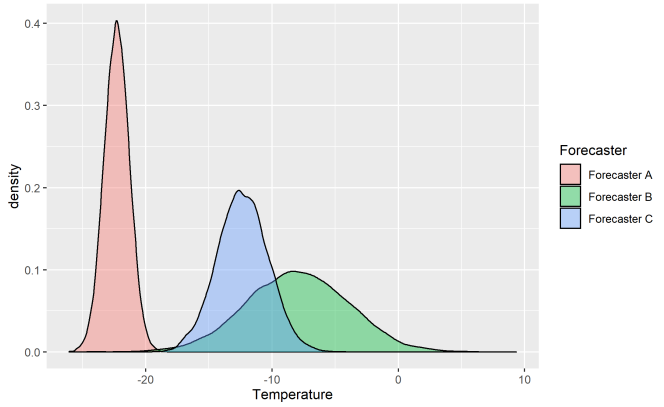

Figure 2.2: Three distinct forecasts of the average daily temperature of tomorrow in January in the Netherlands in $°C$ (fictional forecasts for sake of illustration).

In our example we assume that there exist three different forecasters, forecaster A, forecaster B and forecaster C, which all base their forecast on information.
One of these forecasts is based on the perfect information and uses the perfect model, which is $Y_t|\Delta_t \sim H_t$, while the others have some mistakes in either the information and/or the model. The information of forecaster A concludes that the wind direction of tomorrow will be east, which in the winter in the Netherlands will result in colder temperatures and little to no clouds. Given the information forecaster A has and given the forecasting system behind forecaster A, forecaster A is fairly certain about its temperature range and therefore has a small variance.
The information of forecaster B concludes that the wind direction could be either north or north-east, which in the winter in the Netherlands gives milder temperatures than eastern winds, since the wind comes over water bodies like the North Sea (northern wind) or the Baltic sea (north-eastern wind) before arriving in the Netherlands. Water bodies are generally warmer than land areas in the winter and the forming of clouds will work as an insulating layer, trapping the heat in the lower atmosphere. Due to the uncertainty of the wind direction, forecaster B has a higher variance than the other forecasters.
The information of forecaster C concludes that the wind direction will be north-east to east-north-east, therefore the wind partially flows over the Baltic sea before arriving in the Netherlands. The warmer Baltic sea air will be warmer than the forecast issued by forecaster A. On the contrary, a north-eastern wind will go over more land than a northern wind, therefore the forecast issued by forecaster C is on average colder than that of forecaster B.

In this example, forecaster C is the forecast which results from the perfect information and the perfect model. One peculiar thing to note is that the perfect forecaster does not have the lowest variance. This is possible, because sometimes another forecaster is overconfident and issues a forecast which is too narrow, like forecaster A in our case. Many forecasters exist and most of the time, like in our example, there are multiple forecasters available for a quantity of interest (forecaster A, forecaster B, forecaster C and the climatological forecaster in our example). However, how do we know which forecaster is the best forecaster?

Forecaster validation is the theory of ranking forecasters among each other. These rankings can be based on skills that forecasters possess, theoretical backing or results of simulations/real-world observations. These rankings are obtained by use of verification methods and validation methods. A verification method can be used to verify of a forecaster possesses a certain aspect or trait, whereas a validation method can validate how well a forecaster is in a certain attribute.

In our research we differentiate between these two types of methods because each type of method has its own advantages and disadvantages. In our research we will use two different methods, the Probability Integral Transform (PIT) (Dawid (1984), Rosenblatt (1952), Pearson (1933)) and the Continuous Ranked Probability Score (CRPS) (Matheson & Winkler (1976)). The PIT is usually displayed as a histogram, called the PIT-histogram, from which we can verify whether the forecaster is probabilistically calibrated (Gneiting et al. (2007)) or not. The PIT is used on a set of forecast-observation pairs and is a verification method. The CRPS is a scoring function, which is a function which measures the quality of a forecaster based on scores between individual forecast-observation pairs, $s(F_t, y_t)$, and is a validation method. The Mean Squared Error is an example of a scoring function. These scores have a numerical output and most scoring functions, including the CRPS and the MSE, award lower scores to better forecasters. In the following section we will go over the theoretical concepts which the validation methods need to measure in order to rank forecasters adequately.

2.2 What makes a forecaster a good forecaster?

From the introduction we recall that Murphy's goodnesses-of-fit are *consistency*, *quality* and *value*. These goodnesses-of-fit can be seen as a broad framework of forecaster validation. We will explore each of these goodnesses-of-fit in more depth in the subsections below.

2.2.1 Consistency

Consistency is the match between a forecaster's own best judgements and the issued forecasts. The definitions of and differences between the forecaster's own best judgements and the forecaster's forecasts are not clear in and of itself. Murphy defines them as follows:
He assumes that a forecaster derives its forecast about a future event from a knowledge base which can consist of information from different sources. He furthermore assumes that the forecasting process ends up in the formulation of a judgement the future event. He defines these judgements as internal, "recorded only in the forecaster's mind", and that they are not yet an external statement, or forecast. These judgements are the result of the rational processing of the information contained in the knowledge base of the forecaster and therefore it seems reasonable to Murphy that the forecast should be consistent with the information that is contained in the judgement, with total consistency achieved when the issued forecast of a forecaster is always completely in line with its own best judgement. Otherwise, such a forecast would not properly reflect the forecaster's true state of knowledge.
The forecaster's own best judgement can vary from the forecaster's forecast when the entity in control of issuing the forecasts alters their forecasts to, for example, achieve a better evaluation of their forecaster or when the forecast's "spatial or temporal specificity" differs from that of the judgement of the forecaster. The phenomenon where the entity in control of issuing the forecasts alters the forecasts to achieve a better score is called *hedging* (Lerch et al. (2017)). We want our forecaster to be consistent, so we need a counter measure against hedging. A widely used counter measure against hedging is the use of *(strictly) proper* scoring functions. According to Bröcker & Smith (2007), proper scores are the only internally consistent scores, by which they mean that proper scores are the only scores who will give the optimal expected score to a forecast if the observation is drawn from the forecast distribution. We define the concept of properness as follows:

Definition (Properness of a scoring function):
Given y_t, the observation, f_t, the issued forecast of forecaster $\mathcal{F}$, j_t, forecaster $\mathcal{F}$'s own best judgement, and $s(f_t, y_t)$, the score of forecast f_t w.r.t. observation y_t.
Then, the score $s(.)$ is proper if:
Total equality between the issued forecast and the forecaster's own best judgement, $f_t \equiv j_t$, imply a minimisation of the expected score of the forecast-observation pair, $\min\{\mathbb{E}[s(f_t, y_t)]\}$.
Strict properness holds when this minimum of the expected score is unique and therefore the relation becomes 'if and only if'.

An example of how a specific proper scoring function, the Brier Score, will minimise the expected score of a forecaster when its forecasts are in line with its internal judgements can be examined in section 5a of Murphy (1993).

2.2.2 Quality

Quality is the match between the issued forecast and its corresponding observation.
In the beginning of forecaster validation, measure-oriented approaches were used to calculate a forecaster's quality. This approach focuses on the overall correspondence between forecasts and observations and only uses a few overall aspects of forecast quality. One example of a measure-oriented score function is the Mean Squared Error (MSE).
In the 1980's and 1990's a new approach became more prominent, the distribution-oriented approach. This 'new' approach uses the joint distribution of forecasts and observations, $p(\mathcal{F}, Y)$, at its base. Where $\mathcal{F}$ is the forecaster whose outcome, the forecast, depends on the information, which is a random variable, and where Y is a random variable from which the observation is a realised value. In his paper, Murphy (1993) states that all non-temporal information relevant to evaluating a forecaster's quality is captured by the joint distribution of forecasts and observations. The joint distribution $p(\mathcal{F}, Y)$ can be separated in a conditional distribution and a marginal distribution in two different ways, coined by Murphy & Winkler (1987) as the following:

The *calibration-refinement* factorisation: $p(\mathcal{F}, Y) = p(Y|\mathcal{F}) * p(\mathcal{F})$.
The *likelihood-base rate* factorisation: $p(\mathcal{F}, Y) = p(\mathcal{F}|Y) * p(Y)$.

Here, base rate is the name given to the characteristic which depends on the forecasting situation, but not on the forecaster. Another name for base rate is sample climatology or climatological density. With the use of these conditional and marginal distributions we can adequately measure a forecaster's quality. Murphy subdivided quality into a set of aspects of quality.
The calibration-refinement factorisation is linked to the aspects of quality named: *calibration*, *reliability*, *resolution* and *sharpness*, whereas the likelihood-base rate factorisation is linked to the aspects of quality named *discrimination* and *uncertainty*. In Murphy & Winkler (1987), they state that the factors of the likelihood-base rate factorisation can be used to indicate how we should predict the quantity of interest in absence of a forecast and the potential value contained in the forecasts over the climatological density, which is what we predict in absence of a forecast. For the factors of the calibration-refinement factorisation they state that these factors can be used to indicate the properties of the forecasts when taken at face value.
Most users in a practical setting will take forecasts at face value, in light of this information, the calibration-refinement factorisation seems the most appropriate factorisation to use when forecasting for the general public Murphy & Winkler (1987). In Stevenson et al. (2006), they state that *reliability* and *resolution* are the two most important aspects of quality when using the calibration-refinement factorisation. We will now introduce the quality aspects reliability and resolution, and their connection with calibration and sharpness. In order to show this connection more intuitively, we will introduce the definition of reliability and resolution for both

point forecasters and distributional forecasters.

Definition (Reliability of a scoring function, point forecaster):
Given $\mathcal{F}$, a point forecaster, f_i[2], a specific forecast of forecaster $\mathcal{F}$[3], and $\mathbb{E}[Y|\mathcal{F} = f_i]$, the expected value of the observations, given that the forecaster issued the forecast f_i.
A forecaster, $\mathcal{F}$, is reliable if and only if:

$$\mathbb{E}[Y|\mathcal{F} = f_i] = f_i \quad \forall f_i \in \mathcal{F}$$

This means that the expected value of the observations that occur given that the point forecast f_i is issued, is equal to f_i itself. In other words, the conditional bias of the observations given the forecasts should be zero (unbiased) and in the theory this is called the "conditional bias of the forecasts".

Definition (Reliability of a scoring function, probabilistic forecaster):
Given $\mathcal{F}$, a probabilistic forecaster, f_i[4], a specific forecast of forecaster $\mathcal{F}$, and $f(Y|\mathcal{F} = f_i)$, the probability density function of the observations, given that the forecaster issued the forecast f_i.
A forecaster, $\mathcal{F}$, is reliable if and only if:

$$f(Y|\mathcal{F} = f_i) = f_i \quad \forall f_i \in \mathcal{F} \qquad \text{Or in other words,}$$

$$P\left(Y \leq y|\mathcal{F} = f_i\right) = \int_{-\infty}^{y} f_i(x)dx \quad \forall f_i \in \mathcal{F}$$

This means that the pdf, and therefore the cdf, of the observations that occurs given the distributional forecast f_i is issued, is equal to f_i itself. In the theory of forecaster validation, preference is given to the first definition. However, this definition might seem unconventional for readers not familiar with the theory, which is why we added the second definition.

Definition (Resolution of a scoring function, point forecaster):
Given f_i, a point forecast, Y, the set of observations, and n, the number of distinct forecasts f_i.
The resolution of forecaster $\mathcal{F}$ is calculated by:

$$\text{Res}_{\mathcal{F}} = \frac{1}{n}\sum_{f_i} \underset{f_i}{\text{var}}\left(\mathbb{E}\left[Y|f_i\right]\right) = \frac{1}{n}\sum_{f_i} \mathbb{E}_{f_i}\left[\left(\mathbb{E}\left[Y|f_i\right] - \mathbb{E}\left[Y\right]\right)^2\right]$$

Note that the variance and the expected value have a subscript, meaning that they are conditional given their subscript.
Furthermore, most validation methods calculate the resolution (and reliability) aspect of a forecaster on forecast-observation pair basis, therefore the sum can be seen as a finite sum since there is never an infinite set of forecast-observation pairs available.

Resolution measures to what extent the issued forecasts of a forecaster correspond to different observations. If the forecaster has zero resolution, the expected value of the observations given the issued forecaster is equal to the marginal expected value of the observations, which is the

[2]Note that f_i is a value due to $\mathcal{F}$ being a point forecaster.

[3]Note that the concept of time is irrelevant in this definition, therefore we use f_i to refer to a specific forecast, rather than f_t. f_t could give the misconception that only y_t is taken into account. However, if $f_{t_1} = f_{t_2}$, for $t_1, t_2 \in \{1, \ldots, T\}$, the set $Y|\mathcal{F} = f_{t_1}$ includes both y_{t_1} and y_{t_2}. To avoid misconceptions, we use f_i for the forecasts rather than f, since $f(\cdot)$ indicates the distribution of $\cdot$ as well.

[4]Note that f_i is a pdf.

climatological occurrence chance for point forecasters.

Definition (Resolution of a scoring function, probabilistic forecaster):
Given $V_i(\cdot)$, where $V_i(\cdot)$ is the cdf of $Y|(\mathcal{F} = f_i)$, $G(\cdot)$, the climatological cdf of Y, and n, the number of distinct forecasts f_i. The resolution of forecaster $\mathcal{F}$ is calculated by:

$$\text{Res}_{\mathcal{F}} = \frac{1}{n} \sum_{f_i} \int_{-\infty}^{\infty} (V_i(x) - G(x))^2 \, dx$$

In the probabilistic case, the resolution measures the L_2-distance between the cdf of the observations given a certain forecast and the cdf of all observations, which is the climatological distribution G. If the forecaster has zero resolution, the distribution of the observations given a certain forecast is equal to the climatological distribution G for all forecasts that the forecaster can issue.

Calibration is a term which is not used in the subdivision of quality as coined by Murphy. Calibration is the $p(Y|F)$-part of the calibration-refinement factorisation and used mostly synonymously with the aspect reliability, in the sense that complete calibration is equal to complete reliability (Murphy & Winkler (1987)). In some recent papers, a bisection[5] or even a trisection[6] of calibration has been used to create a new angle of approach in quantifying forecaster quality and therefore we will use the term calibration in our research as well.

Sharpness is the $p(F)$-part of the calibration-refinement factorisation, which is named the refinement step. Sharpness is measured using only properties of the forecasts the forecaster, and can broadly be seen as a measure of the variance of the forecasts of forecaster $\mathcal{F}$. The lower the individual variances of the forecasts of $\mathcal{F}$, the higher the sharpness. A probabilistic forecaster can be overconfident and therefore be sharper than the perfect forecaster. This is different compared to calibration, where the perfect forecaster is the only forecaster which is completely calibrated.

Resolution and sharpness are related in the sense that if a forecaster is perfectly calibrated/reliable:

$$\text{Resolution} \; = \underset{f_i}{\text{var}} \left(\mathbb{E}\left[Y | f_i \right] \right) \underbrace{=}_{\substack{\text{perfectly reliable means that } \mathbb{E}[Y|\mathcal{F} = f_i] = f_i.}} \underset{f_i}{\text{var}} \left(f_i \right) = \; \text{Sharpness}$$

Meaning that for perfect reliable forecasters, its resolution is equal to its sharpness.
This result can be extended to probabilistic forecasts. Other than that, properness, reliability and resolution are linked in the sense that proper scoring functions can be decomposed into a reliability, a resolution and an uncertainty term[7]. Uncertainty here can be described as the variability of the observations, $var(Y)$, which is intrinsic with respect to the quantity of interest and therefore cannot be eliminated by a forecaster.
Using the knowledge above, it is also possible to decompose a proper scoring function into a calibration, sharpness and uncertainty term if perfectly reliability is assumed.

2.2.3 Value

Value is the benefit that a forecaster brings to its user, compared to the scenario where there is no forecaster available. This means that the value of a forecaster differs per user. Value is

[5] Czado et al. (2009), Gneiting et al. (2008) and Gneiting et al. (2013).
[6] Gneiting et al. (2007) and Taillardat et al. (2019).
[7] See Bröcker (2009) for this proof.

outside of the scope of our research. However, we will come back to this in our recommendations.

2.3 Continuous Ranked Probability Score

The first method we utilize is the Continuous Ranked Probability Score, or CRPS.

$$CRPS(F_t, y_t) = \int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 \ \mathrm{dx}$$

The CRPS calculates the distance between the cdf of the issued forecast and the cdf of the observation, which is a step function. The CRPS is a scoring function which is applied to a forecast-observation pair. No further information than the observation and the cdf of the forecast is needed, which means that the CRPS can be applied to any forecast-observation pair. The origin of the CRPS can be traced back to the Brier Score, also called the Probability Score or PS (Brier (1950)) and the Ranked Probability Score (RPS) (Epstein (1969), Murphy (1971)), where the CRPS is the integral over the Brier Score and the continuous version of the RPS. The Brier Score is a strictly proper score, and so is the CRPS (Matheson & Winkler (1976), Gneiting & Raftery (2007)). As mentioned in the previous section, strictly proper scores can be decomposed into a reliability, a resolution and an uncertainty component (see Hersbach (2000) for this decomposition for the CRPS).
We want to give a visual representation of what the CRPS calculates using the following figure:

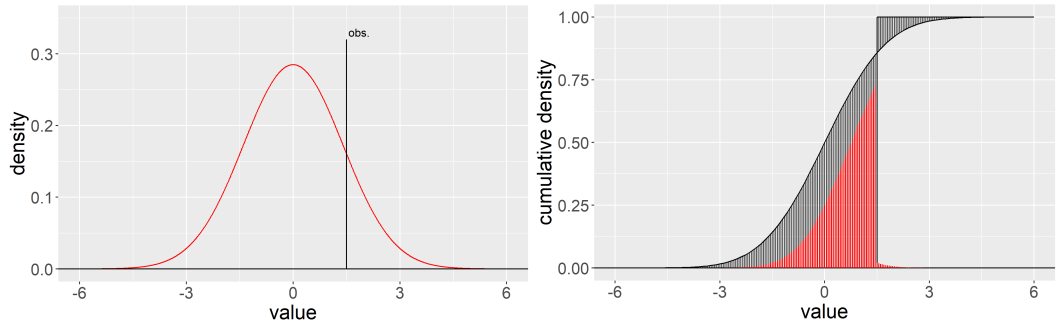

Figure 2.3: On the left we see the issued forecast and the observed observation, while on the right we see the calculated CRPS score, $(F_t(x) - \mathbb{1}\{x \geq y_t\})$ in black and $(F_t(x) - \mathbb{1}\{x \geq y_t\})^2$ in red, of the forecast-observation pair.

On the left side of the figure we see the issued forecast f_t, which is a pdf, and the observed observation, y_t. On the right side we see bars which represent the value of the integrand $\int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 \, dx$ at different values of x. From this figure, we see that the CRPS score[8] can be reduced if the mean of the issued forecast is closer to the observed observation and, if the observation is reasonably close to the mean of the forecast, that a reduction of the variance of the issued forecast will also result in a decrease of its CRPS score[9]. These reductions are in line with an increased calibration and an increased sharpness (or an increased reliability and an increased resolution). Keep in mind that a forecaster can have a too low of a variance, for a certain single observation this can result in a lower score, but over the average this will lead to a decreased calibration. Since perfect calibration means that the forecaster is unbiased, the minimisation problem of calibration and sharpness can be seen as a bias-variance trade-off.

[8]Note that the 'S' in CRPS already means 'Score'. However, "CRPS score" is widely used in scientific papers.
[9]Keep in mind that the lower the score, the better the forecaster.

As mentioned in the previous section, it is also possible to decompose proper scoring functions in a calibration and sharpness component.
For a single forecast, the CRPS can be written as (Taillardat et al. (2019)):

$$CRPS(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(x) - \mathbf{1}\left\{x \geq y_t\right\}\right)^2 \, \mathrm{dx}$$

$$= \underbrace{\mathbb{E}_{F_t}\left|X_t - y_t\right|}_{calibration} - \underbrace{\frac{1}{2}\mathbb{E}_{F_t}\left|X_t - X_t'\right|}_{sharpness}$$

$$= y_t + 2\overline{F_t}(y_t)\mathbb{E}_{F_t}(X_t - y_t | X_t > y_t) - 2\mathbb{E}_{F_t}(X_t F_t(X_t))$$

where X_t and X_t' are two independent random copies generated from F_t, $\overline{F_t} = 1 - F_t$, and $\mathbb{E}_{F_t}$ is the conditional expected value over the issued forecast F_t.
In the following section we will take a closer look at calibration. We will trisect calibration to get a better understanding of what calibration can mean.

2.4 Calibration

The paper of Gneiting et al. (2007) suggests to "Maximise sharpness, given calibration". What level of calibration should we use as 'given calibration'? Gneiting et al. talk about three different types of calibration in their paper: probabilistic calibration, exceedance calibration and marginal calibration.
In the definitions below, $(H_t)_{t=1,2,...}$ are the true cdfs of $(Y_t)_{t=1,2,...}$ and $(F_t)_{t=1,2,...}$ are the cdfs of the probabilistic forecaster. These sequences are based on the information at each time step $(\Delta_t)_{t=1,2,...}$. $(H_t)_{t=1,2,...}$ and $(F_t)_{t=1,2,...}$ are assumed to be countable sequences of continuous and strictly increasing cdfs. Furthermore, the convergence is understood as almost sure convergence as $T \to \infty$.

(1) **Probabilistic calibration**:
$(F_t)_{t=1,2,...}$ is probabilistically calibrated relative to the sequence $(H_t)_{t=1,2,...}$ if:

$$\frac{1}{T}\sum_{t=1}^{T} H_t \circ F_t^{-1}(p) \xrightarrow{a.s.} p \qquad \forall p \in (0,1)$$

(2) **Exceedance calibration**:
$(F_t)_{t=1,2,...}$ is exceedance calibrated relative to the sequence $(H_t)_{t=1,2,...}$ if:

$$\frac{1}{T}\sum_{t=1}^{T} H_t^{-1} \circ F_t(x) \xrightarrow{a.s.} x \qquad \forall x \in \mathbb{R}$$

(3) **Marginal calibration**:
$(F_t)_{t=1,2,...}$ is marginally calibrated relative to the sequence $(H_t)_{t=1,2,...}$ if the limits:

$$\overline{H}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} H_t(x)\right\}$$

and

$$\overline{F}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} F_t(x)\right\}$$

exist, if $\overline{H}(x) \equiv \overline{F}(x) \quad \forall x \in \mathbb{R}$ and if the limit distribution places all mass on finite values.
$\overline{H}(x)$, as defined in marginal calibration, is equal to the climatological forecaster, which is $G(x)$.

Exceedance calibration is outside the scope of our research, we will not be exploring this topic in this thesis.
Marginal calibration can be understood as an equality between the limit of the marginal distribution of the forecasts and the limit of the marginal distribution of the observations. In other words, it can be interpreted as the equality between the climatological distribution of the observations and the climatological distribution of the forecasts of a certain forecaster $\mathcal{F}$.
In the following section we will introduce the Probability Integral Transform (PIT), which has a close relationship with probabilistic calibration.

2.5 The Probability Integral Transform

The Probability Integral Transform, the PIT, is the second method we use in our research. The PIT can be used on a set of forecast-observation pairs $((F_t, y_t)_{t=1,2,\ldots,T})$. For each forecast-observation pair, the PIT value can be described as follows:

$$PIT\,(F_t, y_t) = (F_t(y_t)) \qquad\qquad t = 1, 2, \ldots, T$$

These PIT values are generally shown graphically in a histogram, the PIT-histogram.
The shape of this PIT-histogram can be reviewed and serve as a requirement for probabilistic calibration. To explain why the PIT-histogram can be used as a requirement for probabilistic calibration, we first need to show the following:

Lemma: Given L a random variable, if $M = F_L(L)$, with F_L the cdf of L and F_L being continuous and strictly monotone increasing, then M has a standard uniform distribution.

Proof: $M = F_L(L)$, such that,

$$F_M(m) = \mathbb{P}(F_L(L) \leq m) \underbrace{=}_{\text{Since } F_L(\cdot) \text{ is strictly monotone increasing.}} \mathbb{P}(F_L^{-1}(F_L(L)) \leq F_L^{-1}(m)) = \mathbb{P}(L \leq F_L^{-1}(m)) = F_L(F_L^{-1}(m)) = m$$

Which gives that $F_M(m) = m \quad \forall m \in [0,1]$[10], which is the distribution of the standard uniform distribution.

From this result, we gather that for F_t to be identical to the true distribution H_t, we require the PIT-histogram to resemble the pdf of a standard uniform distribution, which is a pdf where each value between 0 and 1 has the same density. Visually this pdf resembles a linear graph with slope 0. From this Lemma we can derive the following:
Given $Y_t|\Delta_t \sim H_t$, where Y_t is the random variable from which the observations are generated at time step t, given F_t, the cdf of the forecaster, and given $M = F_{Y_t}$, then:

$$\text{if } F_t \equiv H_t \Rightarrow F_M(m) = m \quad \forall m \in [0,1], \text{ then if } \exists m \in [0,1] \text{ for which } F_M(m) \neq m \Rightarrow F_t \not\equiv H_t$$

We have now established that a horizontal PIT-histogram is a requirement for a forecaster to be in correspondence with the distribution of the observations.
We want to objectively verify the uniformity of the PIT-histogram, which can be performed by use of existing goodness-of-fit tests such as the Kolmogorov-Smirnov test, the Cramér-von Mises

[10] $\forall m \in [0,1]$ since $M \in [0,1]$ by definition of M and m is restricted to the range of M.

test and/or the Anderson-Darling test. Due to the finiteness of the set of forecast-observation pairs, the goodness-of-fit tests mentioned above cannot perfectly gauge the uniformity of the PIT-histogram. The accuracy of the uniformity of the PIT-histogram depends on the difference between F_t and H_t, and depends on the number of forecast-observation pairs available. In the following chapter, the simulation chapter, we will show an example where the size of the available dataset has an impact on the test result of a forecaster.

We want to note that, given $X_i, i = 1, \ldots$, which are each a random draw from the standard uniform distribution and given the null hypothesis that X_i is standard uniform distributed, the distribution of the p-values generated from these hypothesis tests will by definition approach the standard uniform distribution when i is large. Furthermore, if the PIT-histogram is non-uniform, the shape of the PIT-histogram can attain many forms[11]. It's for the most part impossible to rank these differently shaped PIT-histograms accurately. Therefore the PIT can only be used as a requirement for a forecaster's probabilistic calibration, or merely separate the forecasters who are probabilistically calibrated from those who are not. In the following subsection we will explore the connection between probabilistic calibration and the PIT.

2.5.1 Probabilistic calibration and the PIT

The PIT has a common ground with probabilistic calibration. To show this, we define F_t and H_t as usual and assume a horizontal PIT-histogram.

$$\frac{1}{T}\sum_{t=1}^{T} H_t(F_t^{-1}(p)) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}(Y_t \leq F_t^{-1}(p))$$

$$\underbrace{=}_{\text{Since } F_t(\cdot) \text{ is strictly monotone increasing}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}\left(F_t(Y_t) \leq F_t(F_t^{-1}(p))\right) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}(\underbrace{F_t(Y_t)}_{\text{which is the PIT}} \leq p)$$

$$\approx \frac{1}{T}\sum_{t=1}^{T} \underbrace{\mathbb{1}(F_t(Y_t) \leq p)}_{\text{the histogram height}}$$

$$\downarrow \quad \text{by uniformity of the PIT}$$

$$\approx p \quad \forall p \in (0,1)$$

From this result it shows that a horizontal PIT-histogram implies that a forecaster is probabilistically calibrated. We want to stress that the uniformity of the PIT is dependent on the size of the dataset and the difference in distribution between H_t and F_t, meaning that small differences or small datasets might not be able to accurately depict if a forecaster is probabilistically calibrated.

There are some candidates to assess exceedance calibration and marginal calibration. These are outside of our scope and will therefore not be discussed in our research.
Now that we established what qualities we are looking for in a forecaster, we will show some practical results in the upcoming chapter.

[11] In the next chapter we will show a sloped, U-shaped and humped/inverse-U shaped PIT-histogram.

Chapter 3

Simulation studies

In the previous chapter we went over the theoretical basis of forecasting and forecaster validation. From the theory we have established that we can accurately gauge a forecaster's quality by the rule "Maximise sharpness, given calibration". We introduced the CRPS and the PIT, which can be applied to rank forecasters. The CRPS is a strictly proper scoring function (validation method) which can be decomposed into a calibration, a sharpness and an uncertainty component (similar to a bias-variance decomposition). The PIT is a verification method which can be used to verify if a forecaster is probabilistically calibrated.
In this chapter we will introduce a data generating model and we formulate several different forecasters which try to forecast the observations generated from this model. Using the observations and the forecasts, we will show some of the particularities of the CRPS and the PIT in a clear way. For the PIT we will show the impact that dataset sizes can have on the result of the PIT-histogram and explore the information that is contained in non-uniform PIT-histograms. For the CRPS we will show that the CRPS might rank forecasters with a lower calibration higher than forecasters which have a higher calibration. At the end of the chapter we will also look at the relationship between the PIT and the CRPS.

3.1 Introducing the model

We assume that nature comes about using a random process and that we can quantify a part of this process by the information given at a certain point.
In our data generating model, the information (Δ_t) are standard normal distributed, $N(0, 1)$, while the observations (y_t) are $N(\Delta_t, 1)$ distributed. Both of these random variables follow a normal distribution, that is why this model is called a Normal-Normal model. This model is chosen due to its simplicity without loss of generality and because various other papers use a Normal-Normal model as well[1]. We assume that Δ_t are independent and identically distributed and y_t are independent as well. The true distribution H_t, where $Y_t|\Delta_t \sim H_t$, is therefore equal to $N(\Delta_t, 1)$. We assume that full information is available for the forecasters, although some forecasters use this correct information to quantify an incorrect forecast. The forecasters will be introduced briefly here and can be viewed in Table 3.1.
The first forecaster we make use of is the perfect forecaster, which is identically distributed w.r.t. the observations. The second forecaster we use is the climatological forecaster which, unlike the other forecasters, does not use any information and always predicts the same pdf for the observations, which is equal to the marginal distribution of the observations. In the Normal-Normal

[1] such as Hamill (2001), Gneiting et al. (2013), Gneiting & Katzfuss (2014), Ehm et al. (2016), Lerch et al. (2017) and Taillardat et al. (2019).

model we use, this marginal distribution is equal to the $N(0, 2)$-distribution[2].
We use two different biased forecasters.
A biased forecaster, which has an upward bias of 1, $N(\Delta_t + 1, 1)$, compared to the perfect forecaster, and a sign-biased forecaster, which twists the sign of the information: $N(-\Delta_t, 1)$.
We use an overdispersed forecaster, which overestimates the variance of the observations: $N(\Delta_t, \frac{9}{4})$.
We use two different unfocused forecasters, the unfocused forecaster and one we call Hamill's unfocused forecaster, as coined in Hamill (2001). Our unfocused forecaster has a fifty-fifty chance to either use the perfect forecaster, $N(\Delta_y, 1)$, or a biased forecaster, $N(\Delta_t + \gamma, 1)$, where this bias, γ, in itself can either be 1 or -1 with equal chances. Hamill's unfocused forecaster is an average of a perfect forecaster and a biased forecaster, which is equal to the biased part in our own unfocused forecaster.
The last forecaster is called Hamill's forecaster. This forecaster is constructed by Hamill (2001) to show that even forecasters which are always biased, can still show a horizontal PIT-histogram. Hamill's forecaster has equal chance to either be $N(\Delta_t - \frac{1}{2}, 1)$, $N(\Delta_t + \frac{1}{2}, 1)$ or $N(\Delta_t, \frac{169}{100})$.

Forecaster	Distribution	Variable (if applicable)
The information	$\Delta_t \sim N(0, 1)$	
The observation	$y_t \sim N(\Delta_t, 1)$	
The perfect forecaster	$F_t^P \sim N(\Delta_t, 1)$	
The climatological forecaster	$F_t^C \sim N(0, 2)$	
The biased forecaster	$F_t^B \sim N(\Delta_t + b, 1)$	$b = 1$
The sign-biased forecaster	$F_t^S \sim N(-\Delta_t, 1)$	
The overdispersed forecaster	$F_t^O \sim N(\Delta_t, \frac{9}{4})$	
The unfocused forecaster	$F_t^U \sim \tau * N(\Delta_t, 1) + (1 - \tau) * N(\Delta_t + \gamma, 1)$	$\tau \in \{0, 1\}, \gamma \in \{-1, 1\}$
Hamill's unfocused forecaster	$F_t^{HU} \sim \frac{1}{2}\{N(\Delta_t, 1) + N(\Delta_t + \phi, 1)\}$	$\phi \in \{-1, 1\}$
Hamill's forecaster	$F_t^H \sim N(\Delta_t + \delta, \sigma^2)$	$(\delta, \sigma^2) \in \{(\frac{1}{2}, 1); (-\frac{1}{2}, 1); (0, \frac{169}{100})\}$

Table 3.1: The Normal-Normal model.

We use three different dataset sizes, t, in our research, which are 1 000, 10 000 and 100 000.

3.2 The PIT

In the previous chapter we have established that the PIT can show if a forecaster is reasonably probabilistically calibrated. We split this section into three parts. In the first subsection we will show the results of some of the forecasters introduced in the previous section. In these results we include the insight we gained from chapter 2 to explain some features of the PIT-histograms. The figures showing the PIT-histograms are based on a dataset size of 10 000 forecast-observation pairs. In the second subsection we will explore the claim that a U-shaped PIT-histogram can occur due to a conditional bias and that a U-shaped PIT-histogram does not imply that the forecaster is underdispersed, which naively could be expected. In the third and last subsection we will show that the size of the dataset can make a significant impact on the shape of the PIT-histogram.

[2]as has been used by Hamill (2001), Gneiting & Katzfuss (2014) and Taillardat et al. (2019).

3.2.1 PIT-histogram results

Sloped PIT-histogram:

The biased forecaster, $N(\Delta_t + 1, 1)$, will systematically overestimate the observations. Therefore $\{F_t(y_t)\}_{t=1,\dots,T}$, will systematically show an overabundance of low values, which results in the following, negatively sloped, PIT-histogram.

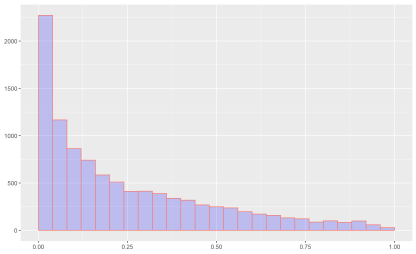

Figure 3.1: Biased PIT-histogram.

For a negatively sloped PIT-histogram in general, we can state that we encounter a lot of observations which return low values of the cdf of the forecaster. From this we can conclude that the cdf of the forecaster generally overvalues the observations. However, not much can be said for the dispersion of the cdf of the forecaster compared to the observations. For a positively sloped PIT-histogram, the results are vice-versa, although nothing can be stated about the dispersion either.

Humped, or inverse-U shaped, PIT-histogram:

The overdispersed forecaster, $N(\Delta_t, \frac{9}{4})$, overestimates the variance of the observations. It overestimates the standard deviation of the observations by 50%. Therefore we encounter less observations in the tails and more observations in the middle of the cdf of the forecaster than expected. This results in a humped or inverse-U shaped PIT-histogram.

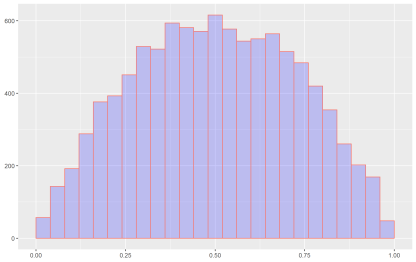

Figure 3.2: Overdispersed PIT-histogram.

For humped of inverse-U shaped PIT-histograms in general we can state that we encounter less observations in the tails and more observations in the middle of the cdf of the forecaster than expected. From this we conclude that the cdf of the forecaster is generally overdispersed compared to the observations.

U-shaped PIT-histogram:

The sign-biased forecaster, $N(-\Delta_t, 1)$, will perfectly forecast observations when the information is 0. However, the more the value of the information moves away from 0, either negatively or positively, the greater the mismatch between the forecaster's cdf and the observations. A negative information will result in a systematic overestimation and a positive information will result in a systematic underestimation. This results in a U-shaped PIT-histogram.

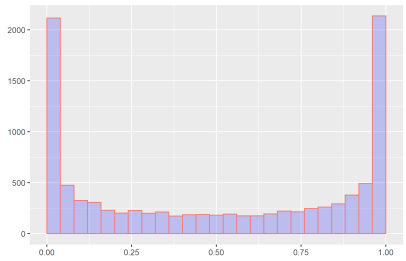

Figure 3.3: Sign-Biased PIT-histogram.

For a U-shaped PIT-histogram, we can state that we encounter a lot of observations in the tails of the cdf of the forecaster and less in the middle than expected. From this we could conclude that the cdf of the forecaster is generally underdispersed compared to observations. However, as shown by our sign-biased forecaster, this type of PIT-histogram can also emerge due to conditional biases and the forecaster can even be overdispersed[3] in combination with a conditional bias.

Horizontal PIT-histogram:

The perfect forecaster, the climatological forecaster, Hamill's unfocused forecaster and Hamill's forecaster all result in similar PIT-histograms. Where the perfect forecaster and the climatological forecaster are completely probabilistically calibrated and Hamill's forecaster has a slight mismatch which is small enough to remain unnoticed in small datasets. On the right you see the PIT-histogram for Hamill's unfocused forecaster. The PIT-histograms of all four forecasters can also be found in Appendix A.

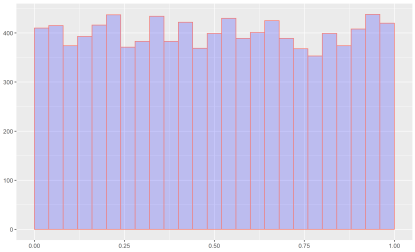

Figure 3.4: Hamill's Unfocused PIT-histogram.

In this subsection we have seen that if the PIT-histogram is non-horizontal, the PIT might not be totally useless. We have looked at different forms of non-horizontal PIT-histograms and what they can tell us about the dispersion and the localisation of the cdf of the forecaster compared to the observations.
Although the non-horizontal PIT-histograms might not always give a conclusive answer why the forecasts cannot accurately depict the observations, they do give a direction in which we can find improvements for the cdf of the forecaster.
However, one cannot state from the PIT-histogram that one forecaster is better than another, other than between a forecaster with a horizontal PIT-histogram and a forecaster with a non-horizontal PIT-histogram.

3.2.2 Creating a special overdispersed forecaster

In Hamill (2001), Hamill states that a U-shaped PIT-histogram can also emerge from a conditional bias. In the previous subsection we have shown that our sign-biased forecaster's PIT-

[3]We will back up this claim in the next subsection.

histogram is a U-shaped PIT-histogram, even though it is not underdispersed. In this subsection we will extrapolate on the claim made in Hamill (2001) and show that even an overdispersed forecaster can have a U-shaped PIT-histogram. For this claim we altered our sign-biased forecaster to make it overdispersed.
The overdispersed sign-biased forecaster, $F_t^{OS} \sim N(-\Delta_y, \frac{9}{4})$ $\left((\mu, \sigma^2)\right)$. The standard deviation of F_t^{OS} is overestimated by 50% compared to the standard deviation of the observations. Even though this new forecaster is highly overdispersed, the PIT-histogram of F_t^{OS} is still very much U-shaped which, naively speaking, is a sign of underdispersion. One can also see great similarities between the PIT-histograms of the sign-biased forecaster in the previous subsection and of F_t^{OS} in this subsection, even though these forecasters' distributions are very dissimilar.

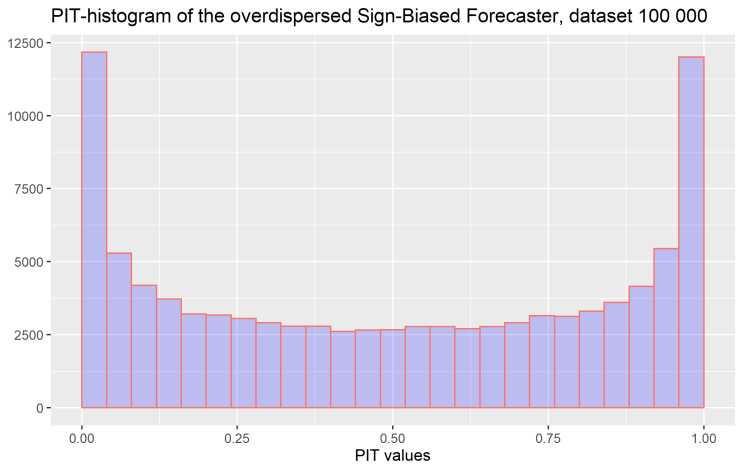

Figure 3.5: PIT-histogram of the overdispersed sign-biased forecaster, on a dataset of 100 000.

3.2.3 The impact of dataset sizes

In the previous chapter we showed that the uniformity of the PIT-histogram can be assessed using established goodness-of-fit (GoF) tests. Examples of these established goodness-of-fit tests are the Kolmogorov-Smirnov (KS) uniformity test, the Anderson-Darling (AD) uniformity test and the Cramér-von Mises (CvM) uniformity test. In this subsection we will show the impact of the size of the dataset on the shape of the PIT-histograms and on the rejection rate of the uniformity tests of the four forecasters with a uniform PIT-histogram (perfect (F_t^P), climatological (F_t^C), Hamill's unfocused (F_t^{HU}) and Hamill's forecaster (F_t^H)). From these four forecasters we know that, theoretically, the perfect forecaster and the climatological forecaster are probabilistically calibrated while Hamill's unfocused forecaster and Hamill's forecaster have a slight deviation from the correct distribution and are therefore not probabilistically calibrated. However, as shown in subsection 3.2.1 and in the paper of Hamill (2001), at certain smaller dataset sizes, this deviation is too small to be seen by inspection. In this subsection we conduct a research using the uniformity tests mentioned above to see if and when the slight deviation in Hamill's forecaster and Hamill's unfocused forecaster are picked up by the uniformity tests.

We researched the rejection rate and the distribution of p-values coming from the different uniformity tests as follows:

1. We simulate a set of information and a set of observations
2. We calculate the set of PIT-values of each the four forecasters with respect to the simulated set of observations and information

3. We derive a p-value from each of the three uniformity tests (KS, AD, CvM) mentioned above w.r.t. each forecaster's set of PIT values

To show the impact of size of the dataset, we simulated observation sets of sizes: 1 000, 10 000 and 100 000. The p-values are derived from hypothesis testing, where the null hypothesis states that the PIT values are standard uniform distributed, with the alternative hypothesis state that the PIT values are not standard uniform distributed. We use a significance level of 0.05. In order to show the average rejection rate of the uniformity tests, we need to simulate multiple datasets for each size. Therefore, we repeated this simulation 100 000 for each dataset size. In the table below, we calculated the percentage of cases where the p-value is below our significance level thus rejecting the null hypothesis hence we call this the rejection rate. These rejection rates are calculated for each of the four forecasters, for each of the three GoF tests and for each of the three datasets sizes. We chose our significance level of 0.05 based on scientific practice, however we can expect that other areas might use a different significance level. Therefore we included the distribution of the 100 000 p-values of each forecaster, for each dataset size, for each GoF test in Appendix B.

	1 000			10 000			100 000		
	KS	AD	CvM	KS	AD	CvM	KS	AD	CvM
F_t^H	5.21%	5.64%	5.08%	9.40%	13.31%	7.19%	73.36%	100.00%	80.67%
F_t^{HU}	4.89%	5.02%	4.99%	5.01%	5.12%	5.10%	4.99%	4.98%	4.99%
F_t^P	4.85%	5.04%	5.04%	4.93%	5.03%	5.02%	4.97%	4.99%	4.98%
F_t^C	4.92%	5.08%	5.09%	4.88%	4.92%	4.87%	4.92%	4.97%	4.99%

Table 3.2: Percentage of cases where the test statistic rejects the uniformity of the PIT-histogram (percentages are based on 100 000 simulations).

Theoretically, we expect to reject the uniformity of the PIT-histogram in 5% of the cases if the PIT-histogram is identical to a standard uniform distribution, as mentioned in section 2.5. In the row of Hamill's forecaster, we see an increased rejection rate commencing at a dataset size of 10 000. In the study of Hamill (2001), they used a dataset size of 10 000 and in this study they showed that Hamill's forecaster, although it's always biased, still shows a uniform PIT-histogram. In their study, they claim the uniformity of the PIT-histograms without providing any results of any uniformity test, although they do advocate the usage of uniformity tests to claim uniformity of the PIT-histogram. In our study we show that at a dataset of 10 000, there is an increased rate of rejection. However, in a practical environment where there is only one set of observations and thus one set of PIT values, which means that we only have one p-value, there is a reasonable chance that Hamill's forecaster would pass the uniformity test of the PIT-histogram. Therefore the warning of Hamill's paper still stands. Uniformity of the PIT-histogram should be used as a requirement only, and never be the only requirement. When the PIT-histogram is assumed to be uniform, this has to be taken with a pinch of salt, depending on the dataset size.

3.3 Results: CRPS score

Now that we have found which forecasters are, reasonably, probabilistically calibrated, we will gauge their performance using the CRPS score. We also want to see if a forecaster which is not probabilistically calibrated can still outperform some of the forecasters which are. In our research we are interested if this outperforming can happen with some consistency, so we simulated the

CRPS scores 100 times. It is quite cumbersome to display 100 CRPS scores per forecaster, hence we used a normality test to check if the set of CRPS scores are normally distributed. For the dataset of 1 000, we have also simulated this study 5 000 times to check for normality. For all forecasters, the set of 5 000 CRPS scores passes the Anderson-Darling normality test. We therefore assume that the CRPS scores are normally distributed for our given dataset sizes, so that we can represent the distributions of the CRPS scores of every forecaster by their mean and their standard deviation. This way, we can display the 100 CRPS scores of each forecaster by means of their distribution. Note that a lower CRPS score indicates a better forecaster.
This results in the following mean (rounded to 3 decimal places) and standard deviation (rounded to 4 decimal places) of the forecasters' CRPS score:

Forecaster	dataset size					
	1.000		10.000		100.000	
	mean	sd	mean	sd	mean	sd
F_t^P	0.566	0.0125	0.564	0.0041	0.564	0.0013
F_t^C	0.835	0.0189	0.835	0.0056	0.835	0.0019
F_t^B	0.700	0.0164	0.700	0.0061	0.700	0.0018
F_t^S	1.395	0.0398	1.390	0.0120	1.390	0.0039
F_t^O	0.593	0.0096	0.592	0.0032	0.592	0.0010
F_t^U	0.801	0.0187	0.797	0.0055	0.798	0.0019
F_t^{HU}	0.632	0.0148	0.632	0.0043	0.632	0.0015
F_t^H	0.615	0.0134	0.614	0.0040	0.614	0.0014

Table 3.3: Mean and standard deviation of the CRPS score of all forecasters of the Normal-Normal model (taken over 100 simulations).

In the table above, we see that the means of the CRPS scores of any specific forecaster do not change drastically between dataset sizes. We see here that the overdispersed forecaster (F_t^O), although this forecaster is not probabilistically calibrated, outperforms most of the forecasters which seem probabilistically calibrated (Hamill's forecaster, Hamill's unfocused forecaster and the climatological forecaster). We also see that the climatological forecaster, although probabilistically calibrated, performs worse than the biased forecaster and the unfocused forecaster. Bröcker (2009) concludes that all (strictly) proper scores have the downside that it is impossible to say how a strictly proper scoring rule will rank 2 not totally reliable forecasters. "The lack of reliability of one forecaster scheme might be out-balanced by the lack of resolution of the other". This result comes from the lack of sharpness of the climatological forecaster[4]. From this we conclude that the CRPS should not be used on its own when one would like to rank forecasters using the rule "Maximise sharpness, given calibration", which contradicts with the statement made by Tsyplakov (2013).
This brings us to a follow-up question, in the rule "Maximise sharpness, given calibration", what minimal level of calibration do we require forecasters to have?

In the upcoming chapter, we will study the steps undertaken in Taillardat et al. (2019). They introduce a different point of view towards the set of observations, which leads them towards a new index which makes use of the CRPS values of individual forecast-observation pairs. In

[4]Note that the CRPS score can be decomposed in calibration, sharpness and uncertainty.
Since uncertainty depends on the variance of the observations and since the climatological forecaster is more calibrated than the biased, unfocused and Hamill's forecaster, the worse CRPS score of the climatological forecaster has to come from its lack in sharpness.

their paper, they also reason what level of calibration a forecaster should possess in order to use it for their index. But before we look into that index, we conducted a small experiment in the last section of this chapter, which looks at the connection between p-values and CRPS scores.

3.4 Results: p-values, CRPS score -connection

In this section, we will look at possible connections between the CRPS and the p-values from the PIT-histogram. We can either look for connections between forecasters, or within a forecaster. Looking for a connection between forecasters based on the p-value and the CRPS score of a forecaster would make no sense. In short, we have only one p-value and one CRPS value in a practical setting, and even the perfect forecaster's p-values cover the whole range between 0 and 1 uniformly. Due to the volatility in the p-value, we cannot construct any meaningful relation between forecasters based on their combination of p-value and CRPS score. For that reason, we will restrict ourselves to a connection between the CRPS scores and p-values within a forecaster. Naively, one could say that the CRPS can be decomposed in a calibration term, a sharpness term and an uncertainty term. Therefore, a higher p-value of a forecaster should lead to a lower CRPS score.

For the perfect forecaster of the Normal-Normal model, the distribution of the CRPS scores based on the p-values of the Anderson-Darling uniformity test using 100 000 forecast-observation pairs looks as follows:

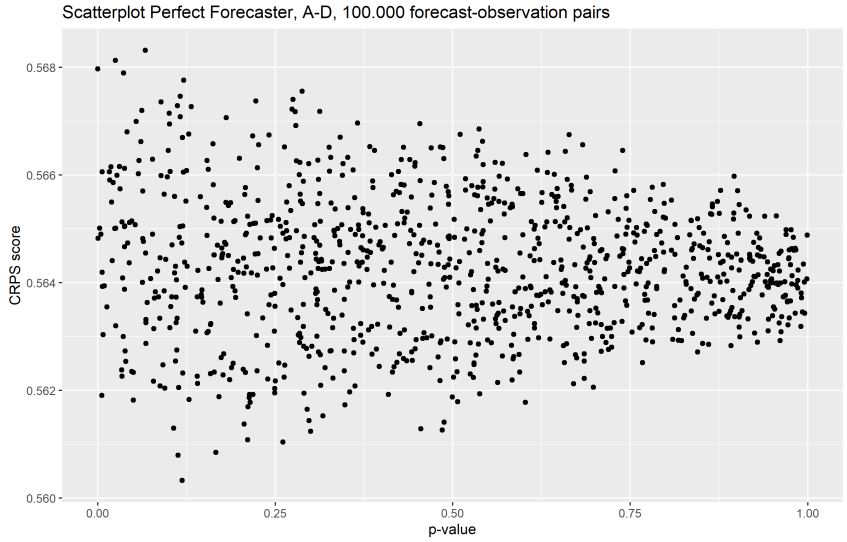

Figure 3.6: Distribution of the CRPS scores of the perfect forecaster, based on the p-values of the Anderson-Darling uniformity test.

From this cone-shaped figure we can see that a forecaster's smallest CRPS scores are located at lower p-values, and the same holds for a forecaster's biggest CRPS scores. Intuitively this makes sense. Given a set of information $\Delta_t, t = 1, \ldots, T$, and given the CRPS formula, the CRPS values of any forecaster are minimised if $F_t(y_t)|\Delta_t = 0.5$, with y_t the observation and Δ_t the information. Therefore, the lowest achievable CRPS score for a forecaster given a certain set of information, corresponds to a PIT-histogram where every PIT value is equal to 0.5. This results in a low p-value from any uniformity test on the PIT-histogram due to an over abundance of PIT values at 0.5. This CRPS score also undervalues the average CRPS score that

that forecaster would receive. The same holds for the higher CRPS scores, they occur when, given a set of information, more observations occur in the tails of the forecasts' distributions, $|(F_t(y_t)|\Delta_t) - 0.5| \to 0.5$, which has the same effect on the p-value as a low CRPS score. Due to the randomness of random variables, one could observe a situation where even a perfect forecaster has an overabundance of observations in the middle of its forecasters and a lack of observations in the tails of its forecasts, resulting in a below average CRPS score and a low p-value, rejecting the null hypothesis that the PIT-histogram is uniform.

The opposite is true for high p-values. A high p-value can only occur when the PIT-histogram is quite uniform, which requires a decent spread, where this spread depends on the forecasts' distributions, of PIT values. This puts contraints on the abundance of high and low CRPS values in the set of CRPS values, which results in the smaller range of CRPS scores that we see at higher p-values.

These conclusions seem straightforward for our model, where the spread of the forecast distribution of the perfect forecaster is always the same, no matter the information. If we used a different model where a change in information would lead to a different spread in the forecast distribution, would a cone shape still appear?
Yes, if we fix the value of information and generate a set of observations, we would end up with a cone shape due to the reasons given above. If we would go through all possible values of the information and generate a set of observations, we would end up with a set of cones. So if we do not fix the information but rerun the generation of information and observations enough, the end result would be a (convex) combination of the cone shapes from the set of cones. Since each of these cones is wider on the lower side of the p-values, their (convex) combination will also be wider, and so the cone shape will persist.

Note that this cone shape is more prominent when using the Anderson-Darling goodness-of-fit test, since this test puts more weight on deviations in the tails of a distribution (or PIT-histogram in this case). So does this cone shape persist when using the Cramér-von Mises or the Kolmogorov-Smirnov uniformity test?

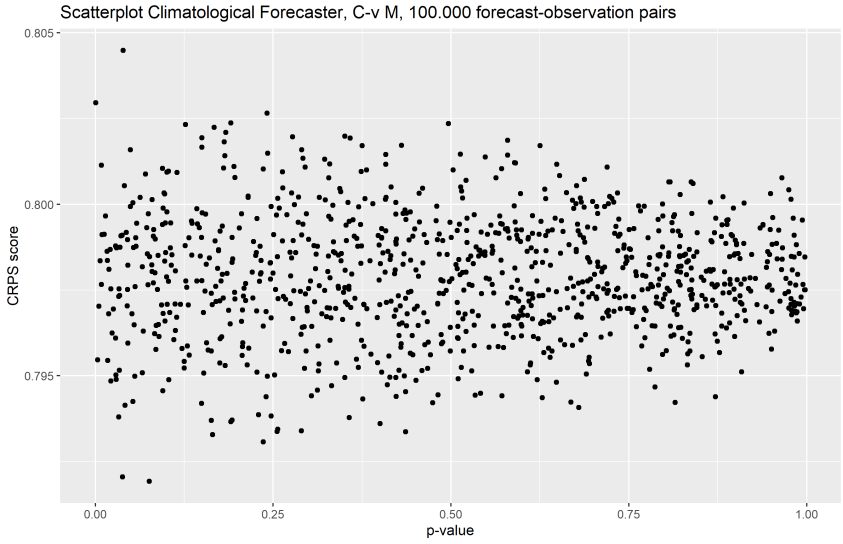

Figure 3.7: Distribution of the CRPS scores of the climatological forecaster, based on the p-values of the Cramér-von Mises uniformity test.

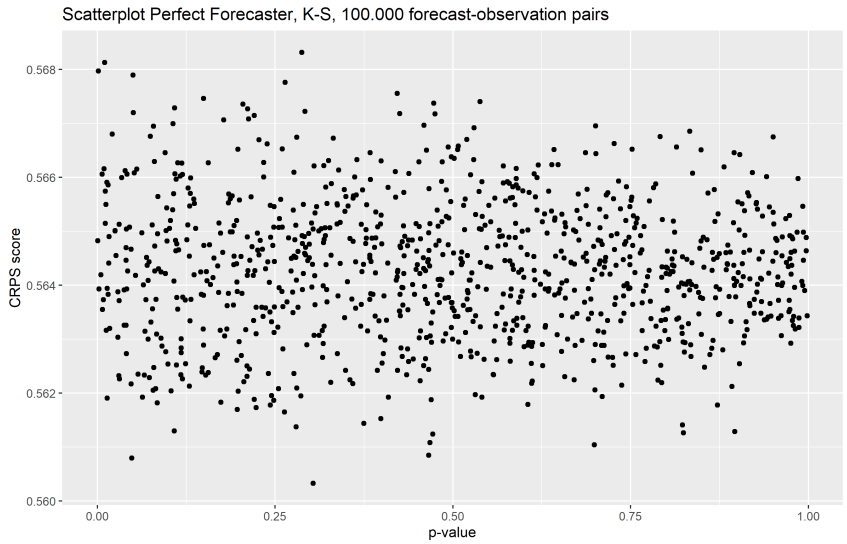

Figure 3.8: Distribution of the CRPS scores of the perfect forecaster, based on the p-values of the Kolmogorov-Smirnov uniformity test.

In our simulation we generated a scatterplot for each of the four forecasters which have a uniform PIT-histogram (perfect, climatological, Hamill's and Hamill's unfocused). In all of the cases the Anderson-Darling gave the strongest resemblence to a cone shape, followed by the Cramér-von Mises uniformity test and lastly the Kolmogorov-Smirnov uniformity test in some cases didn't really resemble a cone shape that well.
For Hamill's forecaster, the result for the 100 000 forecast-observation pairs is a bit different, since Hamill's forecaster only has low p-values.

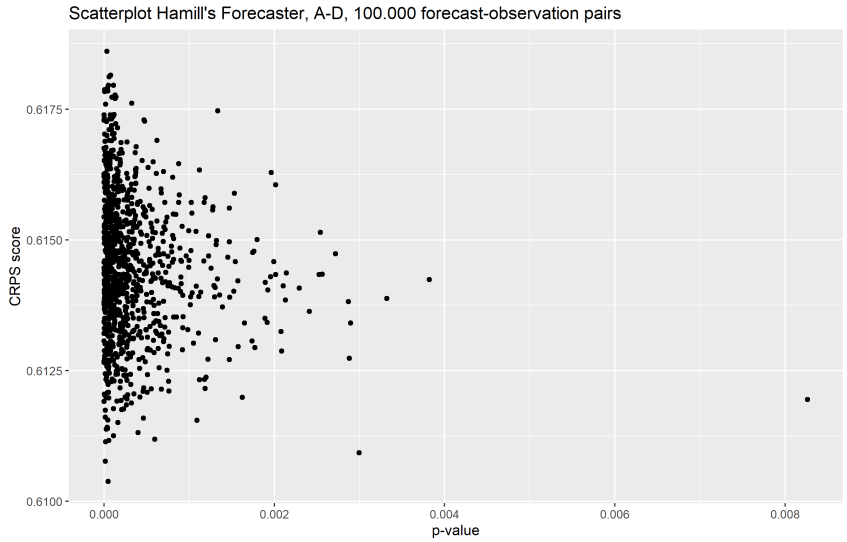

Figure 3.9: Distribution of the CRPS scores of Hamill's forecaster, based on the p-values of the Anderson-Darling uniformity test.

At the moment we are unsure if these cones can be used in any way. Although the introduction of these cones does give us cause for thought for an index which is based on on the difference between a forecaster's theoretical average CRPS score and the CRPS score calculated from the forecast-observation pairs.

Chapter 4

Extreme events and the CRPS

In this chapter, we will explore the validation of forecasters with an emphasis on extreme events. Much of the research conducted in this field makes use of CRPS-based validation methods such as the weighted CRPS and the "Taillardat index"[1]. To start off, we sketch why there is an increased interest in the forecasting of extreme events.

In general, the public and the media focus their attention on extreme events and the forecasts which preceded them due to their societal impact. In their lack of knowledge on the topics of forecasting and forecaster validation, they turn towards the forecasters which adequately forecasted these extreme event(s), regardless of their scientific backing. This can lead to problems when forecasters without good scientific backing, by which we mean forecasters which score poorly on their overall quality, gain an elevated status in the eyes of the public and consequently will be used to influence their decisions[2].

In order to solve this problem, we need to construct validation methods which can fairly score forecasters on their ability to forecast extreme events, without neglecting the overall quality of the forecasters.

The CRPS is a staple within the validation methods which score the overall quality of a forecaster. However, the CRPS itself can only be used on the entire set of available forecast-observation pairs. Due to the rarity of extreme events, the expected score deduction for a forecaster which is subpar in forecasting extreme events is minimal. To give an idea how minimal this change needs to be:

Given a model which is very heavy-tailed but still physically possible in the real world[3], the percentage of extreme events is about 0.1%. The weight of the CRPS values of these extreme events is between 2.5% and 3.5% of the sum of all CRPS values[4]. This means that a forecaster, forecaster A, whose CRPS values of the extreme events are twice as high as that of forecaster B, only needs its CRPS values of the regular events to be on average 3.7% lower than that of forecaster B in order to receive a better score.

The CRPS can be adjusted to put more emphasis on specific, hence also the extreme, events, this adjusted CRPS is called the *weighted CRPS*.

$$wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 \, w(x) \, \mathrm{dx}$$

where $w(x)$ is a weight function.

[1] This index is constructed by Taillardat et al. (2019) and has no official name as of yet.

[2] We will view this as the initial problem, which we will come back to in subsection 4.3.3.

[3] More information on heavy-tailed models can be found in De Haan & Ferreira (2007).

[4] We calculated these percentages using the Gamma-Exponential model which we will introduce later on in this chapter, using 100 000 forecast-observation pairs.

The addition of a weight function in the CRPS does cause some drawbacks. We will only feature one drawback here, because the wCRPS is out of the scope for our research.
One of the drawbacks of the wCRPS has to do with tail equivalence.

A forecast F is tail equivalent to the true cdf H if their upper end points are the same, $x_F = x_H$ and if the limit of the quotient of their survival functions is finite, $\lim_{x->x_F} \frac{\overline{F}}{\overline{H}} = c \in (0, \infty)$.

Tail equivalence is important, since it is needed for a forecaster to be marginally calibrated. Other than that, a forecaster who isn't tail equivalent to the real distribution of the observations will either forecast significantly more or significantly less extreme value than observed. Unfortunately for the wCRPS; $\forall\ \epsilon > 0$, we can construct a forecaster $\mathcal{F}$ which can produce forecasts F which are not tail equivalent to the true forecasters H, such that the difference between the expected wCRPS scores of H and F is less than ϵ.
$|\mathbb{E}_H[wCRPS(H, Y)] - \mathbb{E}_H[wCRPS(F, Y)]| \leq \epsilon$.[5]

This drawback causes forecasters which are not tail equivalent, which means that they either have a heavier or a lighter tail than the true cdf, to perform almost as well as the true cdf. A heavier tail means that the forecaster is more likely to forecast extreme events than the occurrence rate, while the opposite is true for forecasters with a lighter tail. Since we are interested in forecasters which are good at forecasting the extreme events, we want forecasters which are at least tail equivalent to the true cdf. This equivalence does not ensure a perfect forecaster, however it is one of the requirements which we deem important. The weighted CRPS is therefore not a good candidate to use for quality assessment of forecasters with an emphasis on extreme events.

In this chapter we will take a look at the Taillardat index. This index is constructed in an inspiring paper, which at the moment is only available in pre-release status. The paper establishes a new point of view towards the set of observations and in the discussion it calls upon the scientific community to "study the specific properties of this, their index (ed.), CRPS-based tool and its potential paths and pitfalls". Their index focuses on the right end tail behaviour of forecasters and is therefore more suited to assess the quality of forecasters with an emphasis on extreme events. In the following section we will give the theoretical basis of their index, after which we will investigate their simulation and give a reflection on the potential paths and pitfalls of their index. In the reflection we will also give a reflection on the broader shift towards an emphasis on forecasting extreme events. After the reflection we will investigate some potential research paths which make use of the new point of view coined in the paper by Taillardat et al..

4.1 Theoretical basis of the Taillardat index

4.1.1 A CRPS-based tool

In general forecast validation, the CRPS score is a single value by which we rank forecasters. This single value CRPS score is the mean of the individual CRPS values $CRPS(F_t, y_t)_{t=1,...,T}$.

$$CRPS(F, Y) = \frac{1}{T} \sum_{t=1}^{T} \int_{-\infty}^{\infty} \left(F_t(x) - \mathbb{1}\{x \geq y_t\}\right)^2 \ \mathrm{dx}$$

where T is the number of forecast-observation pairs.

Taillardat et al. note that the set of CRPS values holds more information than just this mean score. By approaching the set of observations as a random variable with its own extreme value

[5] A proof of this can be found in Appendix 6.2 in Taillardat et al. (2019).

behaviour, it naturally follows that the set of $CRPS(F_t, y_t)_{t=1,...,T}$ values also has a certain distribution and so its own extreme value behaviour. They build upon this framework of extreme value behaviour of observations and CRPS values, which culminates into the Taillardat index. This index uses the difference in extreme value behaviour of the CRPS values of the forecasters, to rank these forecasters among each other.

Before we go into more detail about the extreme value behaviour of the set of CRPS values, we go over the information which the set of CRPS values holds and the availability of empirical distributions in a practical setting. In a practical setting, we have the observations $y_t \ \forall t$, the forecasts $F_t \ \forall t$ and the CRPS values $CRPS(F_t, y_t)_{t=1,...,T}$.

We can shuffle the observations, creating $y_{\pi(t)} \ \forall t$, and calculate $CRPS(F_t, y_{\pi(t)})_{t=1,...,T}$. By shuffling the observations, we break the dependence between the forecast and the observation in a forecast-observation pair. This dependence is created by the usage of information, Δ_t, by the forecaster.
We note here that the shuffling of observations has no effect on the CRPS values of the climatological forecaster.
To elaborate on this, we firstly reiterate that we assume a stable climate. By assuming a stable climate, we know that the climatological forecaster will always yield exactly the same forecast $G \quad \forall t$, where t is the time variable.
Let's take the first forecast-observation pair of the climatological forecaster, (f_1, y_1), we know that the CRPS value for this pair is $CRPS(f_1, y_1) = \int_{-\infty}^{\infty} (G(x) - \mathbb{1}\{x \geq y_1\})^2 \ \mathrm{dx}$[6]. For a forecast-observation pair of the climatological forecaster at time t, we know that the CRPS value is $CRPS(f_t, y_t) = \int_{-\infty}^{\infty} (G(x) - \mathbb{1}\{x \geq y_t\})^2 \ \mathrm{dx}$.
Here we see that the forecast is still G, since we assumed a stable climate. We also see that if $f_1 \equiv f_t$, $CRPS(f_1, y_1) = CRPS(f_t, y_t)$. If we shuffle the observations, we uncouple the observations from their forecast-observation pair, mix them around by applying a random shuffle operator, π, and then couple them to a forecast to make new forecast-observation pairs. We note that this shuffling does not add, subtract nor changes any observation values. The only difference is that the observations are shuffled and coupled to a different, or sometimes the same, forecast. Since all forecasts are identical due to the stable climate, the *set* of CRPS values before the shuffling is identical to the *set* of CRPS values after the shuffling. Only the order of the CRPS values is shuffled due to the shuffling of the observations.
Mean (CRPS) scores, sets of (CRPS) values and distributional characteristics of sets do not change if the order of the values in a set are shuffled, hence the shuffling of the observations has no effect on the aforementioned characteristics of the set of CRPS values of the climatological forecaster.
Even if we would take a subset of the set of CRPS values, say $CRPS(F, Y)|Y > u$, where u is the 0.75-quantile of the distribution of Y, this subset would still contain the same CRPS values and its distributional characteristics remain unchanged after the shuffling of the observations.
To recap, by shuffling the observations, we break the dependence between the forecast and the observation in a forecast-observation pair. This dependence is created by the usage of information, Δ_t, by the forecaster. Since the shuffling has no effect on the climatological forecaster, we can conclude that the climatological forecaster has zero information.

The CRPS score is a proper score, which means that it can be trisected into a calibration, a sharpness and an uncertainty component. The uncertainty component is a trait which is only

[6]Note that because of a stable climate, the climatological forecaster always yields exactly the same forecast, whose cdf is equal to $G(x)$. Therefore the cdf of f_1 is equal to $G(x)$.

influenced by the observations. This means that by assuming a high level of calibration, the difference in CRPS score between forecasters is mostly due to the difference in expected sharpness. This is also strongly related to the level in which information is used[7].
We saw that the climatological forecaster uses zero information and can therefore be used as a benchmark for the amount of information used by forecasters which have a high level of calibration. Taillardat et al. will use this in an extreme value behaviour setting. However, we think that there is more use for this information, which we will present in our potential paths at the end of this chapter.

For the construction of the Taillardat index, they assume both probabilistic calibration and marginal calibration as their "high level of calibration" and they call this combination 'auto-calibration'.

To summarize, if we view the set of observations as a random variable with a certain extreme value behaviour, rather than as a realisation of random variables[8], it naturally follows that we can view the CRPS values as a random variable with its own extreme value behaviour. When assuming a high level of calibration, the difference in CRPS score or distribution of the CRPS values between forecasters can be attributed to the difference in information usage between these forecasters. We have established that the climatological forecaster uses no information, which means that we can use this forecaster as a benchmark for the expected sharpness of forecasters which have a high level of calibration. For the Taillardat index, we have to assume both probabilistic calibration and marginal calibration.

4.1.2 The extreme value behaviour of the CRPS

To study the extreme value behaviour of the CRPS values, we utilise the field of Extreme Value Theory (EVT). From the field of EVT, "excesses over large thresholds"[9] can be used to study the distribution of the right end tail of a random variable. Taillardat et al. start off their paper with the Normal-Normal (NN) model, which is almost identical to our NN-model used in the previous chapter. However, due to the light tails and the slow convergence in large values experienced in normal distributions, this model can be quite limiting when looking at the extreme value behaviour of forecasters. For this reason they create a new model, the Gamma-Exponential (GE) model. The GE-model is constructed as follows:

Forecaster	distribution	variable (if applicable)
The information	$\Delta_t \sim \Gamma(4,4)$	
The observation	$Y \sim \mathcal{E}xp(\Delta_t)$	
The perfect forecaster	$PF \sim \mathcal{E}xp(\Delta_t)$	
The climatological forecaster	$CF \sim$ GPD cdf with $\sigma = 1$ and $\gamma = \frac{1}{4}$	
The unfocused forecaster	$UF \sim \mathcal{E}xp(\frac{\Delta_t}{\tau_t})$, with $\tau_t = \frac{2}{3}U_1 + \frac{1}{3}U_2$	$U_1 \sim \mathcal{U}[\frac{1}{2}, 1], U_2 \sim \mathcal{U}[1,2]$
The extremist forecaster	$EF \sim \mathcal{E}xp(\frac{\Delta_t}{1.5})$	

Table 4.1: The Gamma-Exponential model.

Here, GPD means Generalised Pareto Distribution.

[7] For more background, see section 4.1 in Taillardat et al. (2019).
[8] See the introduction of Taillardat et al. (2019).
[9] De Haan & Ferreira (2007).

The marginal distribution of a Gamma-Exponential mixture model is equal to the Generalised Pareto Distribution (For reference see: the book Reiss & Thomas (2007) chapter 5.5, the article Wang (1998), the article Mert & Saykan (2005), the article Bopp & Shaby (2017) and the article Wong & Collins (2020)). The most interesting part of the GE-model is the fact that, without loss of generality, the perfect forecaster has a light tail, no matter the value of the information Δ_t, whilst the climatological forecaster has a heavy tail ($\gamma = \frac{1}{4}$)[10]. This will play a vital role in the difference in extreme value behaviour between forecasters and plays a vital role in the creation of their index.
To understand the extreme value behaviour of the CRPS, we reiterate the definition of the CRPS value and its decomposition.

$$CRPS(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(x) - \mathbb{1}\{x \geq y_t\}\right)^2 \, \mathrm{dx}$$

$$= \mathbb{E}_{F_t} |X_t - y_t| - \frac{1}{2}\mathbb{E}_{F_t} \left|X_t - X_t'\right|$$

$$= y_t + 2\overline{F_t}(y_t)\mathbb{E}_{F_t}(X_t - y_t | X_t > y_t) - 2\mathbb{E}_{F_t}(X_t F_t(X_t))$$

where X and X' are two independent random copies generated from the cdf F_t.
For large observations, $y_t > u$ with u large ($u >> 1$), we can rewrite its decomposition to

$$= y_t + 2\underbrace{\overline{F_t}(y_t)}_{\text{tends to 0 for } y_t > u} \mathbb{E}_{F_t}(X_t - y_t | X_t > y_t) - 2\mathbb{E}_{F_t}(X_t F_t(X_t))$$

$$= y_t - 2\mathbb{E}_{F_t}(X_t F_t(X_t))$$

This holds when $\mathbb{E}_{F_t}(X_t - y_t | X_t > y_t)$ is finite. If a forecaster is marginally calibrated, its total mass is on finite values. Furthermore the cdfs of its forecasts are continuous. Therefore marginal calibration functions as a requirement for the finiteness of $\mathbb{E}_{F_t}(X_t - y_t | X_t > y_t)$. We assume auto-calibration when using the Taillardat index, so marginal calibration is assumed by default and justifies our simplification.

The idea of the Taillardat index stems from the difference in extreme value behaviour between the perfect forecaster and the climatological forecaster. For the perfect forecaster, they give the following:
Given X_t a random variable with continuous cdf F_t^P and Y_t a random variable with continuous cdf H_t. If F_t^P and H_t have the same right end point $x_{F_t^P} = x_{H_t}$, H_t belongs to the domain of attraction of L_{γ_t}[11], which is a GPD with $\gamma = \gamma_t$, and $c_{F_t^P} = 2\mathbb{E}_{F_t^P}(XF_t^P(X))$ is finite, then given Δ_t, we have for s such that $1 + \gamma_t s > 0$, as $u \to x_{H_t}$

$$\mathbb{P}\left(\frac{CRPS(F_t^P, Y_t) + c_{F_t^P} - u}{b(u)} > s | Y_t > u, \Delta_t\right) \to (1 + \gamma_t s)^{-\frac{1}{\gamma_t}}$$

This equations shows that given Δ_t and for extreme observations, the extreme value behaviour of the CRPS values is equivalent to the extreme value behaviour of the observations. For the climatological forecaster we arrive at a different tail behaviour.

[10] Short background in Extreme Value theory (De Haan & Ferreira (2007)):
A distribution is defined to have a heavy tail if its tail is heavier than the tail of the exponential distribution. The exponential distribution has a γ value of 0, where γ is known as the *tail index*.

[11] L_{γ_t} or L_{γ,σ_u} is defined as the extreme value distribution of the observations, meaning that: $Y|Y > u \sim L_{\gamma,\sigma_u}$ and: $Y_t|Y_t > u \sim L_{\gamma_t,\sigma_u}$. (more background can be found in De Haan & Ferreira (2007))

Namely, given that G belongs to the domain of attraction of L_γ with $\gamma > 0$, then, for s such that $1 + \gamma s > 0$, as $u \to x_G$

$$\mathbb{P}\left(\frac{CRPS(G,Y) - u}{b(u)} > s | Y > u\right) \to (1 + \gamma s)^{-\frac{1}{\gamma}}$$

The vanishing of the constant term is because of the sublinear/linear behaviour of $b(u)$ for $\gamma > 0$ (Taillardat et al. (2019) Appendix 6.6 for this proof, which also refers to Von Mises (1936)). This difference in extreme value behaviour should stem from the difference in sharpness if auto-calibration is assumed.

4.1.3 The index

The Taillardat index measures the difference in tail behaviour between the climatological forecaster, $CRPS(G,Y)$, and another forecaster, such as the perfect forecaster $CRPS(F_t^P, Y_t)$, by using the Cramér-von Mises test statistic for one sample, which is defined as[12]:

$$T_u = m \times \hat{\omega}_u^2 = \frac{1}{12m} + \sum_{i=1}^{m}\left[\frac{2i-1}{2m} - L_{\gamma,\sigma_u}(v_i)\right]^2$$

The Cramér-von Mises test statistic is a hypothesis test, which means we can derive a p-value from this test statistic. This test statistic is a hypothesis test which tests if, given extreme observations above a threshold u, the distribution of the CRPS values of forecaster F is identical to the distribution of the null hypothesis. The null hypothesis is the distribution of the CRPS values of the climatological forecaster G, given extreme observations above a threshold u. To create the Taillardat index, we calculate this p-value for the tested forecaster F and for the climatological forecaster G. By dividing the p-value from forecaster F, p_u^F, by the p-value of the climatological forecaster G, p_u^{clim}, it results in the following asymmetric index for forecaster F:

$$1 - \frac{p_u^F}{p_u^{clim}}$$

In short, they state with this index that: the more a forecaster's extreme value behaviour differs from the extreme value behaviour of the climatological forecaster, the better it is.

The idea that "the more the CRPS values of forecaster F differ in extreme value behaviour from the CRPS values of the climatological forecaster, the better it is" stems from the following:
In chapter 4 of Brehmer et al. (2019), they give the example that the tail index[13] of a forecaster is a max-functional.
A functional $T : \mathcal{F} \to \mathbb{R}$ is called a max-functional if

$$T(\lambda F_1 + (1-\lambda)F_0) = max(T(F_0), T(F_1))$$

holds for all $F_0, F_1 \in \mathcal{F}$ and for all $\lambda \in (0,1)$.

This means that if you take a convex combination of two distributions, given $\lambda \in (0,1)$, the tail index of that convex combination is the maximum of the tail indices of the individual distributions which make up that convex combination.
By assuming auto-calibration, all forecasters in question are marginally calibrated, meaning that

$$\overline{F}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} F_t(x)\right\}$$

[12] Cramér (1928), Von Mises (1928)
[13] The tail index is an indication for the thickness of the right end tail of a distribution, denoted by γ.

exists, and is equal to the climatological forecaster $G(x)$.
Intuitively, this would mean that if F is an auto-calibrated forecaster and $F_{\gamma_{max}}$ is the forecast for which F attains its maximum tail index γ_{max}, then if $F_{\gamma_{max}}$ has a positive probability of occurring, the climatological forecaster has to have a tail index of at least γ_{max}. Since this holds for all auto-calibrated forecasters, the climatological forecaster has the highest tail index of all auto-calibrated forecasters.
Let us assume the opposite. We state that we have a climatological forecaster F^G with tail index γ_G and an auto-calibrated forecaster F^Q, which has a non-zero, positive chance of issuing the forecast F_t^Q, which has a tail index of γ_Q, where $\gamma_Q > \gamma_G$. Since F^Q is auto-calibrated, the forecaster is by definition marginally calibrated. From marginal calibration we know that $\overline{F}(x) = \lim_{T\to\infty} \left\{ \frac{1}{T} \sum_{t=1}^{T} F_t(x) \right\}$ and $\overline{F}(x) \equiv \overline{H}(x) \equiv G(x)$. Since the tail index is a max-functional, we know that the tail index of the climatological forecaster is at least as large as the tail index of F^Q, which is in contradiction with $\gamma_Q > \gamma_G$.
Note that a higher tail index means a heavier right end tail, therefore the expected CRPS score of the forecaster with the higher tail index is usually higher than the expected CRPS score of a forecaster with a lower tail index. However, the statement that the climatological forecaster has the highest tail index of all auto-calibrated forecasters is not definitively proven in either Taillardat et al. (2019) or Brehmer et al. (2019). Therefore we do want to add some caution here.
The crux of this proof is that the definition of max-functionals takes the convex combination of a finite number of forecasts (two forecasts, to be exact), whereas the definition of the marginal calibration equates the climatological forecaster to the convex combination of an infinite number of forecasts.

Another point which remains unproven, but intuitively makes sense, is the assumption that the perfect forecaster has the highest sharpness of all auto-calibrated forecasters[14]. Although they do not mention this assumption explicitly, they do state that if high levels of calibration are required, the CRPS score boils down to a measure of expected sharpness. In combination with the sentence above, "the more your CRPS values differ from the extreme value behaviour of the CRPS values of the climatological forecaster, the better you are", they implicitly assume that the perfect forecaster has the highest sharpness of all auto-calibrated forecasters. If another forecaster has a higher sharpness, their index would prefer this forecaster over the perfect forecaster, which is impossible since the perfect forecaster is equivalent to the true cdf H_t and therefore has to receive the best score.

Now that we have seen the background of the index, we will give a run-down of our simulation of the Taillardat index on their GE-model in the upcoming section.

[14]Tsyplakov (2013) aims to give a proof for this statement.

4.2 Simulation of the Taillardat index

In this section we will conduct a simulation of the GE-model to show its results. In the next two sections we will use these results to give and investigate the potential paths and pitfalls of the Taillardat index. The steps undertaken to create the index are summarised in table 5 of their paper, which we will use as guidance throughout this simulation. At some of the steps, they leave room for interpretation, which occasionally makes it unclear which exact steps are taken in their simulation.
We try to follow their steps as closely as possible and at the points where it seems unclear, we give our best interpretations and possible other scenarios.
To recap the steps to create the Taillardat index:

1. Calculate the CRPS values of the available forecasters
2. Choose a threshold u, we will look at the forecast-observation pairs above this threshold
3. Estimate the Generalised Pareto (GP) distribution of the set of observations over u (or the set of CRPS values belonging to the observations over u)
4. Calculate the Cramér-von Mises statistic that the CRPS values of the forecaster belong to the estimated GP distribution
5. Calculate $1 - \frac{\text{p-value of the CvM statistic of the selected forecaster}}{\text{p-value of the CvM statistic of the climatological forecaster}}$
6. Repeat steps 1-5 for each threshold w over u

Without the loss of results, we can simplify the GE-model by only using the perfect forecaster and the climatological forecaster. The model will therefore look as follows:

Forecaster	Distribution
The information	$\Delta_t \sim \Gamma(4,4)$
The observation	$Y \sim \mathcal{E}xp(\Delta_t)$
The perfect forecaster	$PF \sim \mathcal{E}xp(\Delta_t)$
The climatological forecaster	$CF \sim$ GPD cdf with $\sigma = 1$ and $\gamma = \frac{1}{4}$

Table 4.2: The simplified Gamma-Exponential model.

The first step towards the Taillardat index is to calculate the CRPS values. In order to do this, we will look at both a *practical setting* and a more *theoretical setting*.
In a *practical setting*, the climatological distribution is generally calculated using the available observations (a climate is defined as the average over a long period of time) or a sample which is independent and identically distributed to the observations. Therefore, we will calculate the CRPS values of the climatological forecaster using a CRPS sample method. This method asks for a sample coming from the desired distribution. In order to reduce the bias in a model, a sample which is i.i.d. with regards to the observations can be used, which is often called *training data*.
In a more *theoretical setting*, we use the theoretical climatological distribution G and use the integral of the CRPS to calculate the CRPS values. In a practical setting this G is not known with absolute certainty. However, we are interested in both the theoretical and the practical solidity of the Taillardat index. Therefore we will show the results of both settings and make a clear distinction on which setting is which in our results.

After calculating the CRPS values, we have the following steps left:

1. A threshold u has to be chosen
2. A Pareto approximation is conducted using the threshold u, in order to estimate the γ and σ of L_{γ,σ_u}, the GP distribution
3. For each threshold over u, called w, a new scale parameter for σ, called σ_w, is estimated
4. Using the values for γ, σ and σ_w, the Taillardat index is calculated for each forecaster

The paper uses 1 000 000 forecast-observation pairs for their simulation. Practically this would equate to over 2 500 years of daily observations, or over 100 years of hourly observations. In most fields of research the datasets normally do not have this many forecast-observation pairs. Due to technical limitations of our available hardware and to accommodate a more practical setting, we will simulate using 1 000 and 10 000 forecast-observation pairs.

The threshold u is chosen to be the 0.75-quantile of the observations. This value was chosen for u because the results presented in Figure 5 of Taillardat et al. (2019) start at the 0.75-quantile. Other than these results the only boundary condition on u is that u has to be large enough such that $CRPS(G,Y)|Y > u \sim Y|Y > u$. We will go over all the remaining steps in the following subsection which displays our results.

4.2.1 Our results

In our results we will work through the calculation of γ, σ and σ_w, which have some difficulties to them, as well as the calculation of the Taillardat index. The index itself does not seem to come out as smooth as the paper announces it to be, but more on that we will show below.

The calculation of the γ and σ

Using the 0.75-quantile of the observations as threshold u, we estimate γ and σ. We can use the CRPS values of the climatological forecaster or the observations for the estimation, as equation (13) of the paper shows us that:

$$CRPS(G,Y)|Y > u \sim Y|Y > u \sim L_{\gamma,\sigma_u}$$

We have found that the 0.75-quantile is not sufficient to satisfy the equation above. As a result, u is not sufficiently large to satisfy the equation. In the following figure we show the results of a simulation which estimates the γ and σ using either $CRPS(G,Y)|Y > u$ or $Y|Y > u$. We have used the `fitgpd` formula in R for the 0.75-quantile and we reproduced this simulation 1 000 times. We have added the γ and σ estimates for the set of CRPS values of the perfect forecaster as well, which we use in our reflection in section 4.3. We have combined them into one figure to increase readability.

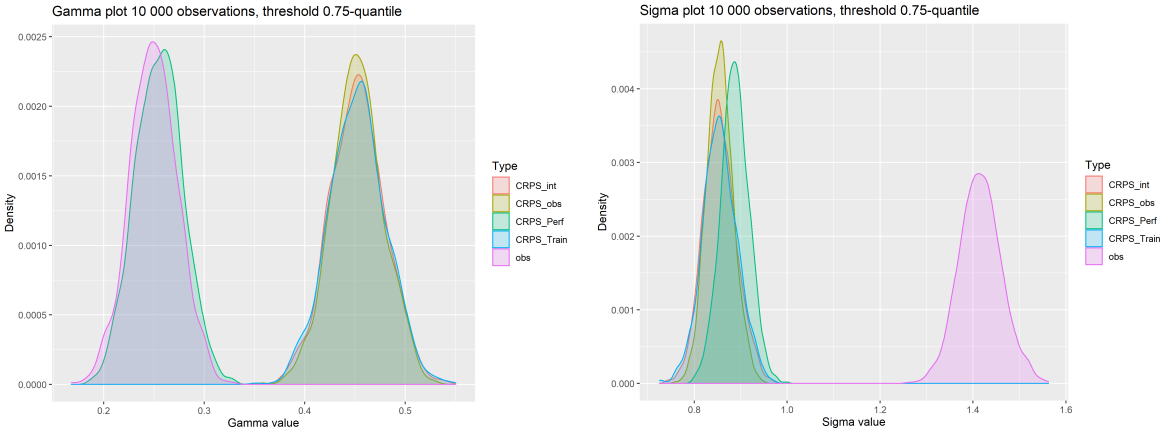

Figure 4.1: γ and σ estimations at the 0.75-quantile.

In the figure above we see three types of estimations which use the set of CRPS values of the climatological forecaster.
`CRPS_int` is the $CRPS(G, Y)|Y > u$ -estimation which uses the more theoretical setting we have mentioned in section 4.2. This version uses the integrand as presented in section 2.3 where we introduced the CRPS.
`CRPS_obs` is the $CRPS(G, Y)|Y > u$ -estimation which uses the more practical setting we have mentioned in section 4.2. This version uses the observations to construct the climatological cdf.
`CRPS_train` is similar to `CRPS_obs`, however `CRPS_train` uses training data to construct the climatological cdf.
On the left we see the γ-estimates for a set of 10 000 forecast-observation pairs, on the right we see the σ-estimates for a set of 10 000 forecast-observation pairs.
In this figure we see that the three estimations using the CRPS values are grouped closely together. The estimation using the tail of the observations is far removed from these estimations and so we have to look further for a better u.

In our search for a better u, we have recreated our simulations for u being the 0.90- and 0.95-quantile. These results, as well as the results for u being the 0.75-quantile, can be found in Appendix C. In these results we have added the results for 1 000 forecast-observations pairs and omitted the σ-estimates, because the γ-estimates have more importance since γ determines the heaviness of the tail. From these figures we observe the following:

1. The γ-estimates using the different sets of CRPS values stay closely together when using higher quantiles for u
2. The distance between the γ-estimates using the CRPS values and the γ-estimates using the observations decreases when using higher quantiles for u
3. The range of the γ-estimates widens when using higher quantiles for u

We want to see if point 2. and 3. do not cancel each other out, therefore we added three figures to Appendix C.1 which display the absolute differences in γ-estimates between the setting which uses the observation's right tail and the theoretical setting of the CRPS values. In these figures we do see that the absolute difference in γ-estimates decreases and using a 0.95-quantile for u instead of the 0.75-quantile decreases this difference on average by 85%, similar absolute differences occur when either of the 2 practical settings is used instead of the theoretical setting.

We are also interested if a lower γ-estimate using the observations corresponds to a lower γ-estimate when using the CRPS values. In the figure below we see the correlation between the two different sets. We only put one of the correlation options here, since all other options show a similar result (a full size version of this figure can be found in Appendix C.2).

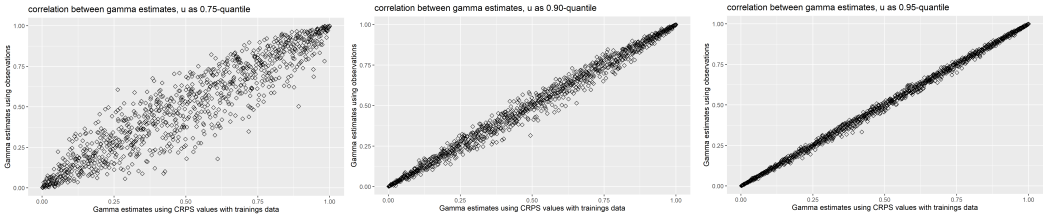

Figure 4.2: Correlation γ-estimates using obs. or CRPS values at the different values of u.

On the left side we see the results for u being the 0.75-quantile, in the middle u is the 0.90-quantile and on the right side, u is the 0.95-quantile. We saw the biggest deviation from the main diagonal when comparing the γ-estimates using $Y|Y > u$ with the γ-estimates using $CRPS(G, Y)|Y > u$ with training data, which we presented in our figure above. We see a near perfect straight line, which means the following:
Given a simulation in which we estimate a small γ compared to other simulations whilst using the observations, $Y|Y > u$, to conduct the γ-estimate, we can say that we are almost sure that the γ-estimate whilst using the CRPS values, $CRPS(G, Y)|Y > u$, will also result in a small γ compared to other simulations where we use the CRPS values to estimate γ.
This means that in our case the γ-estimates using the CRPS values will systematically overvalue the γ, we know this since the observations should have a tail index of 0.25. Combined with our previous results we see that this overvaluation decreases when u is chosen larger.
We take these results into account when trying to construct the Taillardat index. We want to remain true to the paper, so we will calculate the index with u equal to the 0.75-quantile. In addition we will also construct the index for u being the 0.90-quantile and the 0.95-quantile. The reason why we choose to do both the 0.90-quantile and the 0.95-quantile is similar to our reason to not look further past the 0.95-quantile. When we use the 0.95-quantile, our simulation of 1 000 forecast-observation pairs uses only 50 observations to calculate the tail index, this is not a lot of observations. When using even less, problems can arise due to a lack of information. Moreover, when we look at the γ-estimates in Appendix C, we see the simulation using 1 000 forecast-observation pairs already estimates γ to be negative in some of the simulations. Which brings other fundamental problems since the difference in extreme value behaviour between the perfect forecaster and the climatological forecaster is partly based on the fact that the climatological forecaster is heavy tailed[15].

The calculation of σ_w

The Taillardat index can be presented as a graph which features the index values between the threshold u and some high quantile. To calculate the index values over the threshold u, the paper suggests to keep the γ-estimate which is made at the threshold u, and to re-estimate the σ. This updated σ is computed via $\sigma_w = \sigma_u + \gamma_u \times w$.
However, if we use a threshold slightly higher than u, say $w = u + \epsilon$, where $\epsilon > 0$ but small, then σ_w can be much greater than σ_u, while the set of CRPS values which we fit this model to will hardly change or will not even change at all. σ_{w_2}, with $w_2 = u + 2 * \epsilon$, on the other hand,

[15]The constant term in the extreme value behaviour of the climatological forecaster vanishes due to the properties of $b(u)$ when $\gamma > 0$, as stated in section 4.1.2 .

is not much greater than σ_w, even though they also differ by one ϵ.
We therefore suggest to use $\sigma_w = \sigma_u + \gamma * (w - u)$.

The calculation of the Taillardat index

Now that we have estimated the γ and σ, the estimation of the Generalised Pareto (GP) distribution, for the whole range of the Taillardat index, we can compute the index. There are a lot of scenarios that we have covered, to list them here:

1. Estimation of the GPD using $Y|Y > u$ and calculating the CRPS values using the practical setting where the cdf of the climatological forecaster is calculated using a training set
2. Estimation of the GPD using $Y|Y > u$ and calculating the CRPS values using the theoretical setting where the cdf of the climatological forecaster is the theoretical cdf
3. Estimation of the GPD using $CRPS(G, Y)|Y > u$ and calculating the CRPS values using the practical setting where the cdf of the climatological forecaster is calculated using the set of observations Y
4. Estimation of the GPD using $CRPS(G, Y)|Y > u$ and calculating the CRPS values using the practical setting where the cdf of the climatological forecaster is calculated using a training set
5. Estimation of the GPD using $CRPS(G, Y)|Y > u$ and calculating the CRPS values using the theoretical setting where the cdf of the climatological forecaster is calculated using the theoretical cdf

Each of these 5 scenarios is tested for u being the 0.75-quantile, the 0.90-quantile and the 0.95-quantile. Each of these versions is then run for both a set of 1 000 forecast-observation pairs and a set of 10 000 forecast-observation pairs. This culminates into 30 different scenarios, which we deem to be exhaustive enough to test the stability of the Taillardat index.
In order to test the stability of each scenario, we have repeated the index calculation 100 times. In the figure below we show one of the most positive results we have found in our study.

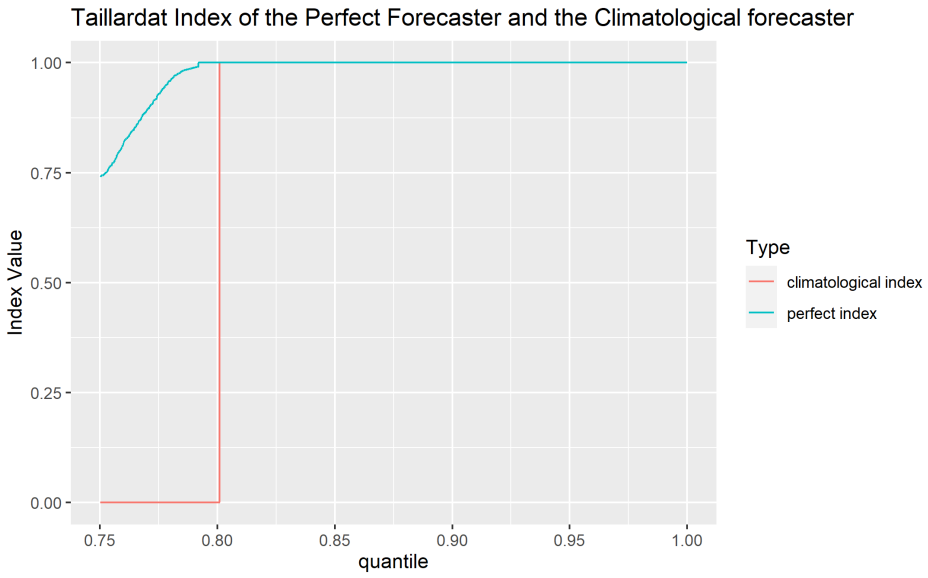

Figure 4.3: Index values of the perfect and climatological forecaster using u is the 0.75-quantile and 10 000 forecast-observation pairs.

From these results we see that the index does not have the expected result which is featured in Taillardat et al. (2019). For starters, we expect the index of the climatological forecaster to be 0 at all times, since $1 - \frac{\text{p-value from the CvM statistic of the climatological forecaster}}{\text{p-value from the CvM statistic of the climatological forecaster}} = 0$. However, after the threshold w gets further from u, there is an increased chance that the tail of the climatological CRPS values does not follow the earlier estimated GP distribution anymore, even when we update the σ-estimate. To show some intuition about the result, we have calculated the percentage of climatological index values which are equal to 0 for each of the 30 scenarios mentioned above.

		1.000			10.000		
u	GPD estimate ↓	CRPS obs.	CRPS train.	CRPS int.	CRPS obs.	CRPS train.	CRPS int.
0.75	Y obs.		47.2%	47.4%		0.0%	0.0%
	CRPS obs.	62.3%			22.0%		
	CRPS train.		62.2%			22.0%	
	CRPS int.			62.1%			22.0%
0.90	Y obs.		95.6%	95.5%		40.3%	40.0%
	CRPS obs.	94.8%			54.0%		
	CRPS train.		94.7%			53.9%	
	CRPS int.			94.7%			53.8%
0.95	Y obs.		97.2%	97.2%		72.1%	72.0%
	CRPS obs.	97.0%			73.4%		
	CRPS train.		97.0%			73.4%	
	CRPS int.			97.0%			73.3%

Table 4.3: Percentages of the index values of the climatological forecaster equal to 0.

We see that the percentages increase once a higher percentile is chosen for u. We think that these higher percentages occur because the 0.95-quantile is close enough to the high quantiles such that these smaller subsets are still reasonably equally distributed. Furthermore, every percentage equates to a bigger distance in quantiles when using the 0.75-quantile compared to the 0.95-quantile.

For the perfect forecaster we expect the index to exist in the range of $[0, 1]$, depending on its test statistic. The only scenarios where the index of the perfect forecaster stays within the range of $[0, 1]$ are the scenarios where u is the 0.75-quantile with 10 000 forecast-observation pairs and the GP estimation is done with either CRPS obs., CRPS train., or CRPS int.. Sadly these scenarios have the lowest percentage of correct index values for the climatological forecaster.

Although we are confident of our corrected σ_w calculation, we did recalculate all 30 scenarios using the original calculation of σ_w written in the paper, which uses $\sigma_w = \sigma_u + \gamma \times w$, instead of our corrected version of $\sigma_w = \sigma_u + \gamma \times (w - u)$. The results of this test return even lower percentages, for reference we can find this table in Appendix C.3.

In response to our disappointing results we have contacted the authors of the paper in order to track down where the process went wrong. In their response they included parts of their code, which will not be disclosed. From their code we saw similar steps to our undertaken steps and no solid reason why our results do not match the results mentioned in the paper.

4.3 Our reflection on the Taillardat index

In the previous section we have seen multiple scenarios in which the Taillardat index failed to deliver a solid result. In this section we will give our reflection on the results of the previous section, our reflection on the fundamentals of the Taillardat index and a reflection on the shift towards more emphasis on forecasting extreme events. In the next section we explore some potential paths for future research.

4.3.1 Reflection on the results

In the previous section we have investigated the research done in Taillardat et al. (2019). In our investigation we have set up 30 different scenarios on which we tested the Taillardat index for the Gamma-Exponential model.
From our extensive results we have found that the Taillardat index is not able to create a stable result in any of the given scenarios. We define a stable result as a figure for which:

1. The index values of the perfect forecaster are limited to the $[0, 1]$ domain and
2. The index values of the climatological forecaster are 0, unless less than 30 data points remain in the set of CRPS values or a higher quantile than the 0.99-quantile is achieved

We note that the number of datapoints in our second point is taken arbitrarily, which is why we gave a quantile boundary as well. However these quantile boundaries might not suffice when a small dataset is used, because loss/lack of information happens earlier when smaller datasets are used.
Even though we duplicated our study for each scenario 100 times, we still couldn't find any simulation which gave us a stable result.

We saw that when w becomes significantly larger than u, the distribution of the set of CRPS values of the climatological forecaster will deviate from the original estimation of the GP distribution at u. This results in an index value of 1 for the climatological forecaster, instead of the expected 0. This is one of the instabilities of the Taillardat index.

Another instability of the Taillardat index stems from the index values of the perfect forecaster. We have stated in our results that in many of the scenarios, the index value of the perfect forecaster attained values outside of the expected range of index values, which is $[0, 1]$. The index values of the perfect forecaster only step outside the $[0, 1]$-range when the Cramér-von Mises statistic deems it more likely that the set of CRPS values of the perfect forecaster fit the estimated GP distribution than that the set of CRPS values of the climatological forecaster fit this distribution. The scenarios for which the perfect forecaster stayed within the $[0, 1]$-range are the scenarios which used the 0.75-quantile for u, a low quantile, with 10 000 forecast-observation pairs, more forecast-observation pairs, and with the GP distribution estimated using any of the three sets of CRPS values of the climatological forecaster we constructed in section 4.2. These results are not unexpected.
Firstly, when we look at the figures in our results section and Appendix C, we see that, given that u is the 0.75-quantile, the γ-estimates of the perfect forecaster are more distant from the γ-estimates of the climatological forecaster, when comparing to other values of u.
This will lead to a higher CvM statistic of the climatological forecaster compared to the CvM statistic of the perfect forecaster.
Secondly, when there are more forecast-observation pairs, the Cramér-von Mises statistic, and more broadly any goodness-of-fit statistic, can more accurately measure if a set of values is

distributed according to the given distribution.
Thirdly, when the GP distribution is estimated using any of the sets of CRPS values from the climatological forecaster, rather than the set of observations over u, the Cramér-von Mises statistic of the set of CRPS values of the climatological forecaster will most likely be higher than the set of CRPS values of the perfect forecaster.
Note that all of the three options will result in higher p-values using the CvM statistic for the climatological forecaster, and lower p-values using the CvM statistic for the perfect forecaster. Moreover, all of the options need to hold to keep the index values of the perfect forecaster within the $[0, 1]$-range, not just any subset of them. Note that this holds for the given Gamma-Exponential model and not necessarily has to hold for other models or when using other parameters in a GE-model.

The elaboration above clashes with a different condition of the Taillardat index.
In the paper they use the convergence provided in the extreme value behaviour of the set of CRPS values of the climatological forecaster to state that $CRPS(G, Y)|Y > u \sim Y|Y > u \sim H_{\gamma,\sigma_u}$. However, as we have seen above, the Taillardat index will be more instable when these sets are similar in distribution[16]. We therefore need more than 10 000 forecast-observation pairs to stabilise the Taillardat index for the higher quantiles of u for which we can assume that $CRPS(G, Y)|Y > u \sim Y|Y > u$. If we increase the number of forecast-observation pairs, the CvM statistic will be able to pick up on more subtle differences between the set of CRPS values of the climatological forecaster and the set of CRPS values of the perfect forecaster. We are unsure if this circumstance can be achieved, since this sounds to us like a contradiction. On the one hand we need that $CRPS(G, Y)|Y > u \sim Y|Y > u \sim L_{\gamma,\sigma_u}$, while on the other hand the CvM statistic needs to be able to detect differences between $CRPS(G, Y)|Y > u$ and $CRPS(F, Y)|Y > u$.

4.3.2 Reflection on the fundamentals of the Taillardat index

The Taillardat index is calculated using the Cramér-von Mises statistic between an estimated GP distribution and the distribution of a certain forecaster. The more the distribution of that certain forecaster differs from the estimated GP distribution, the better the forecaster is. If this difference becomes obvious, the CvM statistic will return a value 0, which is tied to the index value of 1[17]. However, if there is a big difference between:

- The extreme value behaviour of the CRPS values of the climatological forecaster and the CRPS values of the perfect forecaster, and a big difference between
- The extreme value behaviour of the CRPS values of the climatological forecaster and the CRPS values of one or more other, near-perfect, forecaster(s)

then their indices will all return the value 1. This means that, under the circumstances that the set of CRPS values of more than one forecaster is distributed significantly distinct from the set of CRPS values of the climatological forecaster, the Taillardat index cannot distinguish between these forecasters.

The Taillardat index therefore works best when the difference between the extreme value behaviour of the CRPS values of the climatological forecaster and that of the CRPS values of

[16]The instability arises from the higher chance that the CvM statistic will return a higher p-value for the set of CRPS values of the perfect forecaster, than the p-value of the set of CRPS values of the climatological forecaster.

[17]the Taillardat index is $1 - \frac{\text{p-value of the CvM statistic of a certain forecaster}}{\text{p-value of the CvM statistic of the climatological forecaster}}$, if a certain forecaster is very good, its CvM statistic is 0 and therefore the index value of that forecaster is $1 - 0 = 1$.

the perfect forecaster is neither too noticeable nor too insignificant[18]. The reason for this is that the values of the auto-calibrated forecasters other than the climatological forecaster need to return index values between 0 and 1, $(0, 1)$, rather than $[0, 1]$, in order to accurately rank them.

There is one extra fundamental issue of the Taillardat index, which makes it even harder for the Taillardat index of different forecasters to stay within the $(0, 1)$-range. Given that the CRPS values of the climatological forecaster are generated from the extreme value behaviour of L_{γ,σ_u}, then the p-values from the Cramér-von Mises test statistic are standard uniform distributed. We want the p-values from the Cramér-von Mises test statistics of the other forecasters to be lower than this p-value, in order to keep their Taillardat index within the $[0, 1]$-range. However, since the p-values of the climatological forecaster are standard uniform distributed, they can attain any value within the $[0, 1]$-range. The other forecasters therefore need to attain, for the most part, very low p-values. To give an example, Figure 3.9 shows 1 000 p-values coming from Hamill's forecaster using the Anderson-Darling test statistic. The highest p-value in this figure is less than 0.0085. However, Within these 1 000 measurements, there is still an instance where the p-value of the Anderson-Darling test statistic of Hamill's forecaster, is higher than that of the perfect forecaster.
We therefore need an even profounder difference in p-value distribution for the Taillardat index to work. However, in order to show a difference between different auto-calibrated forecasters, their p-values also need to be different from each other, which gets harder and harder when their p-values start to concentrate around 0^+.
Due to the range of p-values that the climatological forecaster can attain, even in the perfect case, the Taillardat index can almost never return a stable index.

The CRPS score does not suffer the same pitfall. The CRPS returns a value on a scale of $[0, \infty)$, for which the best achievable score is only attainable under supernatural circumstances such as a deterministic point forecaster which is always correct and observations whose random variables have zero uncertainty. In practice these circumstances are unlikely to occur, especially not for multiple different forecasters, therefore scores such as the CRPS do not suffer the same problems as indices based on p-values of goodness-of-fit tests like the Taillardat index.
The PIT-histogram does use p-values of goodness-of-fit tests, but the results of the PIT-histogram are used as a verification method for probabilistic calibration, not as a validation method which ranks forecasters.

4.3.3 The shift towards extreme events

In both the introduction and the start of this chapter we explained that there is a shift in the scientific community towards more emphasis on the forecasting of extreme events. The papers written in this area talk about a way to accurately score a forecaster's quality in forecasting extreme events, whilst not disregarding their overall quality. Most of these papers either do not mention Murphy's third goodness-of-fit, *value*, or only briefly mention it in their theory. They are focused on Murphy's second goodness-of-fit, *quality*. However, when we focus on extreme events, we do this because of the interest of the public towards extreme events and the impact of extreme events on the public. We therefore by definition tailor our scoring functions to the public, who value the forecasting of extreme events above that of regular events. This in turn brings a lot of ambiguity in the field of forecaster validation. For example, we do not know to

[18] If this difference is too insignificant, it could happen that the CvM statistic deems the set of CRPS values of the perfect forecaster closer to the estimated GP distribution than the set of CRPS values of the climatological forecaster, which will result in negative index values and thus instability, which we pointed out in the previous subsection.

what extent the public values extreme events over regular events. Neither do we know where the public draws the line between a regular event and an extreme event. Calculations can be done in which everything is expressed in monetary values, but not everything can be expressed in monetary values. Moreover, the public cannot be seen as one entity, but rather is a set of people who each have their own value system and risk tolerance. We believe that if we do not take proper care of the *value* component when conducting research on forecasters with an emphasis on extreme events, we will still endure the same initial problem. We are therefore of the opinion that more research effort should be put towards Murphy's third goodness-of-fit, *value*. As of yet, *quality* has received most of the research efforts and consequently is mapped out extensively. However, such mapping is non-existent for *value*, as far as we know. If a framework could be constructed for *value*, we can more accurately combine the field of *quality* with the field of *value*. Only when we can combine these frameworks, we can properly construct scoring functions which can score forecasters on their ability to accurately forecast extreme events whilst not disregarding their overall quality.

4.4 Future improvement paths of the Taillardat index

In the previous section we gave our reflections on the Taillardat index, which focused on the pitfalls of the Taillardat index. The most important pitfall for the Taillardat index is that the index is based on p-values of the Cramér-von Mises test statistic. In this section we will explore a path for the Taillardat index which is based on the "usage of information" of a forecaster. We will introduce a new index which tries to quantify this "usage of information" and we will address a couple of hurdles in the current path which require more attention.

4.4.1 The theory

In the paper of Taillardat et al. (2019), the authors state that given a high calibration, sharpness is almost identical to the level of information usage[19]. The authors show that the usage of information can be calculated by shuffling the observations. This shuffling, as mentioned in the theory section of the Taillardat index, breaks the conditional dependency between the forecasts and the observations. The more the forecaster's forecasts differ from one another when different information is given, the more different its set of CRPS values would be after the shuffling. This definition of "usage of information" can be named the variance of the forecasts, or the sharpness. If the forecaster uses the information in a good way, where the "good way" would be: "using the information in a similar way to how the real distribution H uses the information", the shuffling should produce a worse score than when no shuffling is applied. In the paper, the authors already produce a figure, figure 3 in their paper, in which they visualise the difference between the CRPS values based on the shuffled forecast-observation pairs and the CRPS values based on the normal forecast-observation pairs. A forecaster which follows the main diagonal undergoes no difference when the shuffling is applied.
Therefore the difference between the shuffled pairs and the normal pairs can be quantified by calculating the area between the curve and the main diagonal. This area is the integral of the difference between the empirical cdf of the CRPS values of the original forecast-observation pairs and the empirical cdf of the CRPS values of the shuffled forecast-observation pairs.

[19] "Indeed, assuming auto-calibration, evaluating the general amount of information brought by a forecast boils down to measure its expected sharpness. This is also equivalent to evaluate its expected score ...".

$$\int_{-\infty}^{\infty} ecdf_{CRPS(F,Y)}(x) - ecdf_{CRPS(F,Y_\pi)}(x)dx$$

where $\pi(\cdot)$ is the shuffling operator.
Note that if the original forecast-observation pairs result in smaller CRPS values, compared to the shuffled ones, the ecdf of the original forecast-observation pairs ($ecdf(CRPS(F,Y))(x)$) will return higher values than $ecdf(CRPS(F,Y_\pi))(x)$, which will result in a positive score. If the shuffled pairs, for some reason, perform better, their CRPS values will be smaller and the integral will return a negative value. This happens when the forecaster uses information, however, using the information actually results in a forecast which is further away from the actual observation than when a random value for the information is used.
Another observation is that the shuffling can result in lower values in some parts of the ecdf curve (lower/higher side of the curve or somewhere in the middle) and higher values in other parts of the ecdf curve. An example of this is a curve which looks like a sine-type wave over the main diagonal. This would result in both a part of the integral which is negative and a part of the integral which is positive, which could cancel each other out. Therefore the calculation of the area will only calculate the net area above the main diagonal, or as we would call it, "the net positive usage of information". By adding absolute signs in the integrand, we can calculate the total area between the main diagonal and the curve. However, we try to rank the forecasters based on how good they are, and using the information in a wrong way should not result in a better score, according to us.

This potential path uses the philosophy which inspired Taillardat et al.'s paper. However, this index is not focused on the right end tail behaviour. In the next subsection, we will firstly research if this new index works in general terms, before we try to specialise the index to work on the right end tail.

4.4.2 The research and results

In this section we will go through our research of the new index coined in the previous subsection and show the results. Our possible new index, which calculates the net positive usage of information of a forecaster, uses random shuffles to acquire $ecdf_{CRPS(F,Y_\pi)}$. We therefore have the following question regarding our new possible index:

1 Every shuffle is random, therefore one shuffle will give a different index value than another.
- 1.1 What is the variance of the index values calculated by different random shuffles?
- 1.2 To which extent does this variance decrease when we calculate the average index value over multiple random shuffles?
- 1.3 Can we achieve an acceptable variance without needing a high number of random shuffles?

To answer these questions we conducted the following simulation:

1. We generated a set of forecasts and observations, (F,Y).
2. We generate 100 shuffled sets of forecast-observation pairs, $(F,Y_{\pi_i})_{i=1,\ldots,100}$.
3. For each of the 100 shuffled sets, we calculate the integral:
   $\int_{-\infty}^{\infty} ecdf_{CRPS(F,Y)}(x) - ecdf_{CRPS(F,Y_\pi)}(x)dx$
   to calculate the area between the curve and the main diagonal, which is our index value.
4. We calculate the mean index value when using 1 shuffled set, 2 shuffled sets, ..., all the way to 100 shuffled sets.

5. We reproduce steps 2. till 4. 1 000 times to calculate the standard deviation of this mean index value.
6. We reproduce steps 1. till 5. 10 times to see if the standard deviation of the mean index value is similar when we use the same model but a different set of information and observations.

From this simulation we received the following result:

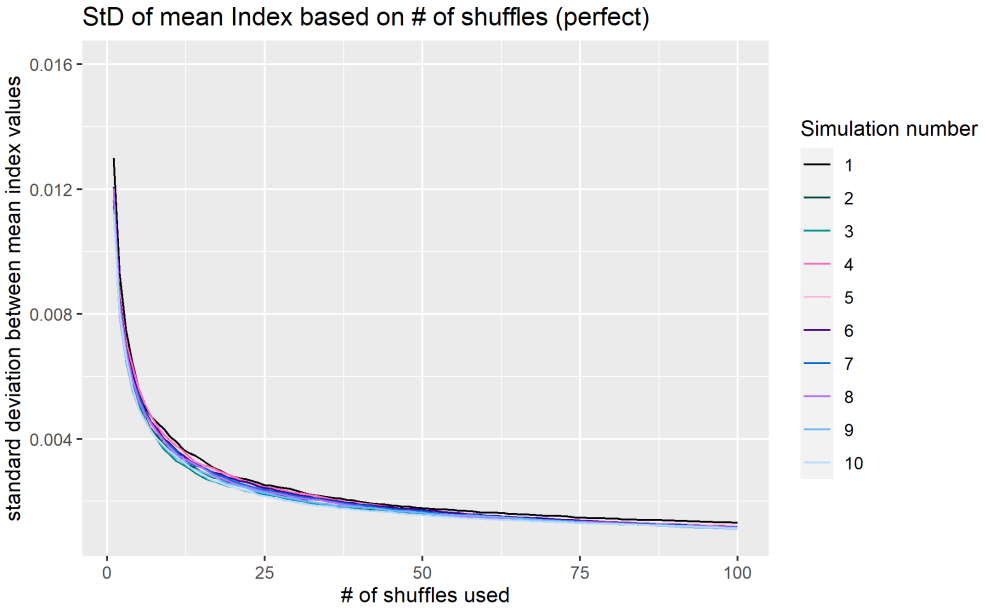

Figure 4.4: Course of the standard deviation of the mean index value, depending on the number of shuffles used (perfect forecaster).

We see that the decrease of the standard deviation of the mean index value is similar between different sets of information and observations which come from the same GE-model. Furthermore, we see that the steepest decline in standard deviation between mean index values occurs between the usage of 1 shuffle and about 15 shuffles. It therefore seems that using the mean index value from 15 shuffles gives a decent enough result, and using more than 15 shuffles does not seem to decrease the standard deviation of this mean index value by a lot more.
To set meaningful requirements for the number of shuffles, we need to know the range of the set of index values.
For example, if the index values are very large, a standard deviation of 0.012 is small and therefore a small number of shuffles would suffice. However, if the index values are smaller, a lower standard deviation might be necessary in order to accurately distinguish the forecasters. For the perfect forecaster, we observed index values between 0.1154 and 0.2982[20].
Therefore the standard deviation of the mean index value when using 15 shuffles, which is about 0.0033, is about $\frac{1}{35}$ of the minimal observed index value.

Another requirement we need to check is to see if the other forecasters' mean index values are distant enough from the mean index value of the perfect forecaster. If for example the index

[20]We want to note here that we simulated 10 different sets of information and we calculated 1 000 000 index values for each of these sets. The minimum and maximum index value come from different sets of information and observations and within one set of information and observations the range of the index values is lower than the minimum and maximum index value make it out to be.

value of the unfocused forecaster is within 0.0033 from the mean index value of the perfect forecaster, a standard deviation of 0.0033 does not seem sufficient enough to determine which forecaster has a higher index value.
In order to check this we first of all calculated the course of the standard deviation of the mean index values of the unfocused forecaster and the extremist forecaster between 1 shuffle and 100 shuffles (using the original GE-model as coined by Taillardat et al. (2019)). These figures can be found in Appendix C.4.
Second of all, we calculated the mean index value of the perfect forecaster, the unfocused forecaster and the extremist forecaster using the following steps:

1. We generated a set of forecasts and observations, (F, Y).
2. We generate 1 000 shuffled sets of forecast-observation pairs, $(F, Y_{\pi_i})_{i=1,\dots,1\,000}$.
3. We calculate the mean index value of these 1 000 index values.
4. We reproduce steps 1. till 3. 1 000 times to see how much the mean index values of the perfect forecaster, the unfocused forecaster and the extremist forecaster differ from each other when taking the average index value over 1 000 shuffles.

1 000 shuffles is far more than the 1 to 100 shuffles we calculated in the previous step. However, when we calculated the standard deviation of the mean index value when using 15 shuffles, we need multiple mean index values which used 15 shuffles. In order to not contaminate these results, we simulated 1 000 sets of 100 shuffles and took the standard deviation of 1 000 mean index values which used 15 shuffles, each of these mean index values used 15 independent shuffles such that no singular shuffle is used multiple times in the set of mean index values which each used 15 shuffles. Due to our hardware limitations, we could only calculate 3 000 000 index values within a reasonable time frame (1 000 000 index values for each forecaster). When we calculate the difference between the mean index values of the different forecasters, we compare the mean index value of one forecaster with another forecaster and therefore only need 1 000 shuffles per forecaster to achieve this comparison. This simulates 100 times quicker, which makes it possible to use 1 000 different sets of information and observations instead of 10.

At that point we found a disappointing result. In 571 of the 1 000 cases, the mean index values of the perfect forecaster and the unfocused forecaster differed less than 0.0033, which is equal to one standard deviation when using 15 shuffles. Moreover, the histogram of the difference shows that the difference between these mean index values is centred around 0, meaning that the perfect forecaster and the unfocused forecaster seem to use the same amount of information. Our index value would therefore rank the perfect forecaster and the unfocused forecaster as equally good. We have added these histograms to Appendix C.5 for reference. Other than that, in 1 000 out of 1 000 cases, the mean index values of the extremist forecaster are the highest of all the three forecasters. This means that our index would rank the extremist forecaster as the best forecaster out of the 3.

Our conclusion for this disappointing result is that we can only proclaim that the perfect forecaster has the highest sharpness of all the forecasters which are highly calibrated, which we did not assume as a requirement for our new index.
This is because the sharpness of a forecaster is calculated by the variance between the cdfs of the forecasts that this forecaster makes. For example, a forecaster which is marginally calibrated needs to comply with the requirement that:

$$\overline{F}(x) = \lim_{T\to\infty} \left\{ \frac{1}{T} \sum_{t=1}^{T} F_t(x) \right\} = G(x)$$

This requirement puts a cap in place on the variance of the cdfs of the forecasts of a certain forecaster, because the convex combination of the set of forecasts needs to equal $G(x)$. It is possible to create a forecaster F whose set of forecasts differ more from each other, but because of that the convex combination of the set of forecasts of F does not equal $G(x)$ anymore. Therefore we could create a forecaster F whose forecasts differ more from each other than that the forecasts of the perfect forecaster differ from each other, which would give this forecaster a higher sharpness. Our new index therefore needs extra requirements on the calibration side in order to work properly.
We are stuck at this point in our research for a possible new index which makes use of the new path that was undertaken in the theory of Taillardat et al. (2019). More research is required to determine if a viable path exists.

To elaborate on the requirement of more research:
In the results of Taillardat index in their paper, we see that the extremist forecaster will still outperform the perfect forecaster in the Taillardat index. The extremist forecaster is only disregarded because it is not auto-calibrated. This is a highly technical term, meaning that a non-expert in the field of forecasting and forecaster validation could overlook this fact and could therefore wrongly assume that the extremist forecaster is the best forecaster (which makes sense, because a visualisation of the Taillardat index shows that the extremist forecaster is the best forecaster).
The initial problem that we wanted to tackle is to create a new index which can accurately rank forecasters in their overall quality, with an emphasis on extreme events. The reason for the problem is because the media and the public would sometimes, in our eyes wrongfully, attribute positive recommendations to certain forecasters which have accurately predicted extreme events. Sometimes wrongfully because the scientific backing of some of these forecasters was lacking and therefore their overall quality was lacking. The Taillardat index only gets around this problem by exercising a technical detail which most of the media and public would not understand. Therefore any non-expert, like most of the media and the public, could overlook this detail and would therefore continue their recommendations based on which forecaster accurately predicted extreme events, so our initial problem still stands.
From an expert perspective, we could use the Taillardat index[21], or any equivalent index, to validate the forecasters that we use and see if we could switch out these forecasters for alternatives which do have a good overall quality, but have better forecasting capabilities for the extreme events. This results in more accurate forecasts of extreme events and therefore less occasions of extreme events which were not forecasted, which in turn would decrease the urge for the media and public to search for other forecasters which did accurately forecast these extreme events.
Note that if we find a forecaster which can more accurately forecast extreme events, it will have less overall quality than the forecaster with the highest overall quality, because if this forecaster also has the highest overall quality, we would have already used it to forecast the quantity of interest.

[21]We do note that the Taillardat index as is, is instable and therefore we do not recommend using it in its current form.

To conclude this section and to wrap up this chapter, we would propose the following potential paths with regards to the Taillardat index and the research of forecast validation in general:

1. An alteration of the Taillardat index which does not use p-values to rank forecasters, but a more exhaustive scale like the one that the CRPS uses.
2. The search and validation of verification methods which can verify if a forecaster is exceedance calibrated and/or marginally calibrated.
3. The construction of a framework on Murphy's third goodness-of-fit *value*.

Point number 2 is needed to satisfy the goal to "Maximise sharpness, given calibration". In order to exercise this guideline, we need an accurate way to verify each type of calibration.

Chapter 5

Conclusions and future research

5.1 Conclusions

From our research we conclude that:

1. The Taillardat index in its current form is instable and we should exercise caution when working with this index.
2. The PIT, due to its underlying nature of using p-values, might not give an accurate reading about whether or not a forecaster is probabilistically calibrated, depending on the dataset size and the distance between the cdf of the forecaster and the true cdf of the observations.
3. When using the rule "Maximise sharpness, given calibration", the CRPS should not be used on its own, but rather be an addition to a set of verification methods which verify which forecasters achieve the minimal calibration level(s) required.
4. Future research should be put towards a better understanding of Murphy's third goodness-of-fit, *value*, as well as towards verification methods which can verify if a forecaster is exceedance and/or marginally calibrated.

Alongside these conclusions, we would like to use this chapter to shed some light on several other research paths we have explored. These paths do not fit in our main line of research, however we do think that our insight is valuable for the readers of our research. In the next section we will discuss all of our secondary research paths and in the last section we will combine the recommendations from our main line of research with the recommendations of our secondary paths of research into one exhaustive set of recommendations.

5.2 Secondary research paths results

In this section we provide you with two secondary paths of research that we conducted at different points throughout our main research. The first path entails designing new models which contain a heavy-tailed climatological forecaster and a heavy-tailed perfect forecaster. The second path is an exploration of different weight functions for the CRPS, it includes weight functions which are not solely dependent on the value of the observation, which are traditionally used for the weighted CRPS.

5.2.1 Designing new models

The initial aim of our research was to build upon the newly established Taillardat index and see where we could expand upon this new index. We wanted to see how well the Taillardat index

would hold up when the perfect forecaster and the climatological forecaster were less dissimilar in their tail index. In the Gamma-Exponential model, the perfect forecaster is light tailed, while the climatological forecaster is heavy tailed, with a tail index of $\frac{1}{4}$. We therefore created 3 new models in which both the perfect forecaster and the climatological forecaster are heavy tailed. However, we encountered many hurdles while working with the Taillardat index, up to the point where we were unable to reproduce the rose-coloured results mentioned in its founding paper. Hence we saw no place in our main research to add these new models, since we couldn't conduct any meaningful experiments on them.

We designed three models in which the discrepancy between auto-calibrated forecasters such as the perfect forecaster and the climatological forecaster are less distinct in their tail behaviour. The fundamentals of the Taillardat index need a heavy tailed climatological forecaster, we therefore propose the following three models:

Forecaster	Distribution	Variable (if applicable)
The information	$\Delta_t \sim \mathcal{U}(0, \frac{1}{4})$	
The observation	$Y \sim \mathcal{P}ar(\Delta_t)$	
The perfect forecaster	$PF \sim \mathcal{P}ar(\Delta_t)$	
The climatological forecaster	$CF \sim$ heavy-tailed distribution ($\gamma = \frac{1}{4}$)	
The unfocused forecaster	$UF \sim \mathcal{P}ar(\tau_t \Delta_t)$, with $\tau_t = \mathcal{T}[T_1, T_1, T_2]$	$T_1 = \mathcal{T}[\frac{20}{21}, \frac{21}{20}]$, $T_2 = \mathcal{T}[\frac{10}{11}, \frac{11}{10}]$

Table 5.1: The Uniform-Pareto model.

Forecaster	Distribution	Variable (if applicable)
The information	$\Delta_t \sim \mathcal{B}(5, 1)$	
The observation	$Y \sim \mathcal{P}ar(\frac{1}{4}\Delta_t)$	
The perfect forecaster	$PF \sim \mathcal{P}ar(\frac{1}{4}\Delta_t)$	
The climatological forecaster	$CF \sim$ heavy-tailed distribution ($\gamma = \frac{1}{4}$)	
The unfocused forecaster	$UF \sim \mathcal{P}ar(\tau_t \frac{1}{4}\Delta_t)$, with $\tau_t = \mathcal{T}[T_1, T_1, T_2]$	$T_1 = \mathcal{T}[\frac{20}{21}, \frac{21}{20}]$, $T_2 = \mathcal{T}[\frac{10}{11}, \frac{11}{10}]$

Table 5.2: The $B(5, 1)$-Pareto model.

Forecaster	Distribution	Variable (if applicable)
The information	$\Delta_t \sim \mathcal{B}(1, 5)$	
The observation	$Y \sim \mathcal{P}ar(\frac{1}{4}\Delta_t)$	
The perfect forecaster	$PF \sim \mathcal{P}ar(\frac{1}{4}\Delta_t)$	
The climatological forecaster	$CF \sim$ heavy-tailed distribution ($\gamma = \frac{1}{4}$)	
The unfocused forecaster	$UF \sim \mathcal{P}ar(\tau_t \frac{1}{4}\Delta_t)$, with $\tau_t = \mathcal{T}[T_1, T_1, T_2]$	$T_1 = \mathcal{T}[\frac{20}{21}, \frac{21}{20}]$, $T_2 = \mathcal{T}[\frac{10}{11}, \frac{11}{10}]$

Table 5.3: The $B(1, 5)$-Pareto model.

Here, the $\mathcal{T}$ is a two-point or three-point distribution with equal chances for each outcome.
In the aforementioned models, the tail index, γ, of the perfect forecaster is somewhere between 0 and $\frac{1}{4}$, depending on the value of the information, whereas the climatological forecaster has a

tail index of $\frac{1}{4}$.
We chose a tail index of $\frac{1}{4}$ for the climatological forecaster because this would make the climatological forecaster more similar to the climatological forecaster in the Gamma-Exponential model. Besides that, the natural quantities of interest are seldom very heavy tailed, therefore a low tail index would cater a more practical setting. In our models, the unfocused forecaster is perceived by goodness-of-fit tests like the Anderson-Darling test statistic as probabilistically calibrated up to at least a dataset size of 1 000 000. Note that the unfocused forecaster is not marginally calibrated, and therefore not auto-calibrated. However, in the paper of Taillardat et al. (2019), their unfocused forecaster is not marginally calibrated either, nor auto-calibrated, however it is presented as an auto-calibrated forecaster.
We tried to calculate what type of distribution the climatological forecaster theoretically follows, however we were unable to provide a definitive answer to this question. We do however know for certain that the climatological forecaster has a tail index of $\frac{1}{4}$. In Appendix D and Appendix E we provided the calculations for the tail index of the climatological forecaster of a general Uniform-Pareto distribution and a general Beta-Pareto distribution respectively.
In the figure below we can see the probability density function of the tail index of the perfect forecaster when using either the uniform distribution, the beta(1,5) distribution or the beta(5,1) distribution, rescaled to the [0,0.25]-range like we used in our models.

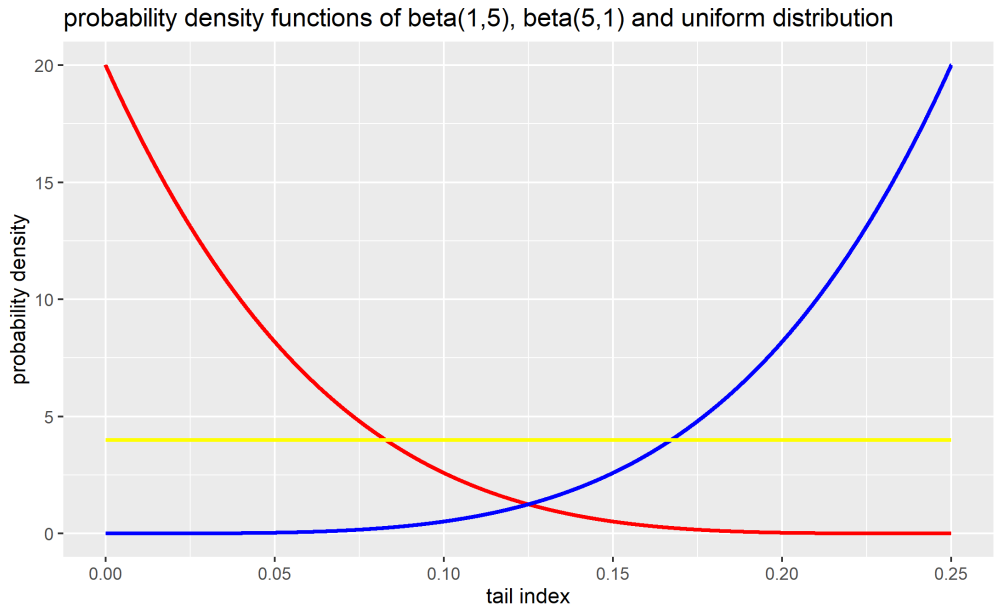

Figure 5.1: Probability density functions of the tail index of the perfect forecaster when using the (red) beta(1,5), (blue) beta(5,1) or the (yellow) uniform distribution model, rescaled to the [0,0.25]-range.

In the figure above, we see that the beta(1,5)-distribution will provide a perfect forecaster whose tail index is usually distant from the tail index of the climatological forecaster. The beta(5,1)-distribution will serve the opposite case where the perfect forecaster's tail index is, on average, close to the tail index of the climatological forecaster. The uniform distribution is chosen as a midway distribution between both extremes. Besides being a midway solution, the uniform distribution will produce a decent number of perfect forecasts with a low tail index, but also a decent number of perfect forecasts with a high tail index. By creating information evenly on the whole range, we may find something interesting regarding the right end tail behaviour of the perfect forecaster.

5.2.2 A new weight function for the CRPS

In the theory we see the weighted CRPS defined as follows:

$$wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(x) - \mathbb{1}\{x \geq y_t\}\right)^2 w(x)\ \mathrm{dx}$$

The weight function in this case is a function which is only dependent on x, or the distance between the cdf of the forecast and the cdf of the observation at a certain range denoted in $w(x)$. This emphasis can be made gradually but also abruptly, depending on the weight function.
For example, when using a quantile weight function, as used in section 2.2 of Taillardat et al. (2019), we can put an emphasis on the distance between the cdfs of a forecast-observation pair at extreme values of x. This weight function can therefore be used to evaluate how well a forecaster forecasts extreme events.
Let's assume that we use the quantile weight function for this exact purpose and let's assume that the quantile weight function is located exactly at the border between the regular events and the extreme events. If we look at this from the hypothesis testing point of view, the set of CRPS values using the quantile weight function will mainly consist of values acquired by false positives, false negatives and true positives. The true positives are the correctly forecasted extreme events, the false negatives are the extreme events which were forecasted as regular events and the false positives are the regular events which were forecasted as extreme events.
From a practical point of view, the most interesting forecast-observation pairs are the false positives and the false negatives[1]. Generally, these forecast-observation pairs have in common that the forecasted cdf is far removed from the cdf of the observation. A weight function which puts either more or less emphasis on this group is a weight function whose function is based on both the cdf of the forecaster and the cdf of the observation.

In the following part we will introduce three different weight functions which alter the CRPS in different ways. We will end this section with a notion on why most of the weight functions which put more emphasis on sharpness than the original CRPS seem to fail.

The original CRPS:

$$CRPS(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(x) - \mathbb{1}\{x \geq y_t\}\right)^2\ \mathrm{dx}$$

The CRPS, by definition, designates a higher value to a forecast-observation pair where the cdf of the forecast is further removed from the observation than a forecast-observation pair where the cdf of the forecast is closer to the observation. The magnitude of the score depends on the difference between $F_t(x)$, the cdf of the forecast, $\mathbb{1}\{x \geq y_t\}$, the cdf of the observation, and the power of the exponent to which this value is raised. In the CRPS, the exponent is 2, which converts the distance between the cdfs into an absolute, positive, distance. However, the choice of the value of the exponent can be altered to put emphasis on different areas. We can turn the exponent into a weight function by defining the weighted CRPS as:

$$wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} \left|F_t(x) - \mathbb{1}\{x \geq y_t\}\right|^{w(\cdot)}\ \mathrm{dx}$$

We note that the distances between the cdfs is a value between 0 and 1. Therefore a value of $0 < w < 2$, compared to the initial CRPS, has the following effect:

[1] Without loss of generality, we can separate the notion of extreme events and positives and regular events and negatives.
Even without these connections, the most interesting forecast-observation pairs in any case are false positives and false negatives, because these groups failed to forecast the correct characterisation of the quantity of interest.

A bigger percentage of the CRPS value can be derived from the areas where the distance between the cdfs is small and therefore ultimately diminishes the gravity of the areas where this distance is large. A value of $w > 2$ creates the opposite effect.
This weight function appears to be a good candidate, although a downside of this weight function is that these exponents do not alter the values 0 and 1 and as a result a high exponent will ultimately favour forecasts which have a high variance. The cause for this is that the situation where the distance between the cdfs is near 1 will hardly ever occur for a distribution with a high variance. Note that a forecaster which produces forecasts which have a higher variance compared to other forecasters have on average a lower sharpness. In Appendix F we will briefly show the results of the introduced weight functions on the NN-model. From Appendix F we conclude that an exponent smaller than 2 favours sharper forecasters over the perfect forecaster, whereas an exponent bigger than 2 favours the opposite type of forecasters.
A second weight function we would like to address is the following:

$$wCRPS(F_t, y_t) = \left(\int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 \ \mathrm{dx} \right)^{w(\cdot)}$$

Here the exponential weight function is taken outside of the integral. Here, a value of $w > 1$ puts more emphasis on the forecast-observation pairs whose integral is bigger than 1. This weight function will suffer a similar drawback.
If the exponent is increased by too much, the ranking of the forecasters will boil down to which forecaster can minimise the maximum CPRS value of its forecast-observation pairs. This drawback is similar to the drawback of the former weight function.
In Appendix F we can therefore see a similar shift in rankings between forecasters. As a result of dethroning the perfect forecaster, we deem both of these weight functions to be inadequate to serve as a weight function that we would endorse. Moreover, we can extent this conclusion without loss of generality and state that: if any alteration of the exponent in the two places mentioned above is made, it is possible to construct a forecaster $\mathcal{F}$ which can outperform the perfect forecaster.
The third and last weight function we introduce is the following:

$$wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 * |x - y_i| \ \mathrm{dx}$$

This weight function puts gradually more emphasis on the distance between the cdfs at values which are more distant from the observation value. The result is that, as long as the bulk of the pdf of a forecaster's forecast is located around the observation value, this weight function will decrease that forecaster's score. Which in turn results in more emphasis on the forecast-observation pairs which could be counted as false positives and false negatives. A drawback of this weight function is that near the observation value, the value goes to 0, which is why $(1 + |x - y_i|)$ might be a better weight function. We have included both weight function to Appendix F. By putting more emphasis on values further away from the observation value, these weight functions favour forecasters with a high sharpness and also in this case the perfect forecaster can be dethroned.

In conclusion, regarding the CRPS, we would like to note that the CRPS can be trisected into a calibration, sharpness and uncertainty part. Therefore the influence a change in sharpness has, is baked into the definition of the CRPS and so is the influence that a change of calibration has. Putting more emphasis on sharpness related conditions show adverse results. We therefore advise to conduct future research into a weight function which is derived from the calibration term. The perfect forecaster has the best calibration of all forecasters, and therefore such a weight function would benefit the score of the perfect forecaster the most of all forecasters.

5.3 Closure and recommendation for future research

This final section of our research is dedicated to a small closing piece followed by an exhaustive list of the recommendations we have given throughout our research for future studies.

Over the course of our research we established an understanding of the theoretical basis of forecasting and forecaster validation. By use of simulations we gained more insight in the practical workings of two prominent methods which can be used to validate forecasters, the PIT-histogram and the CRPS score. We combined this knowledge to understand the theoretical basis of the Taillardat index, to study its properties and to expand the results of the Taillardat index by giving its potential paths and pitfalls.

The initial aim of our research was to build upon the newly established Taillardat index and see where we could take this new index. We unfortunately encountered some hurdles on the way which resulted in a total shift in the perceived aim of our research. We underwent some hurdles personally as well, but ultimately we are content to finish our research with this document.

For future research we have the following recommendations:
A research towards altering the Taillardat index, such that its basis depends on a more continuous ranking scale than the current version, which is based upon p-values of goodness-of-fit tests.
A thorough exposition of Murphy's third goodness-of-fit *value*, to bridge the gap between the fields of *quality* and *value*.
A study towards a weight function for the CRPS which is solely based on calibration and not based on sharpness, including an exhibit of the changes such a weight function triggers.
A study towards verification methods which can verify either exceedance calibration or marginal calibration, in order to complete the list of verification methods for the trisection of calibration.
On the topic of calibration, we would like to add that for each type of calibration there also exists a "complete ... calibration", which requires calibration for every subsequence of the forecaster. We suspect that exploration of this area could lead to new research paths as well.

This encapsules our recommendations for future research.

Appendices

Appendix A

PIT-histograms simulation chapter

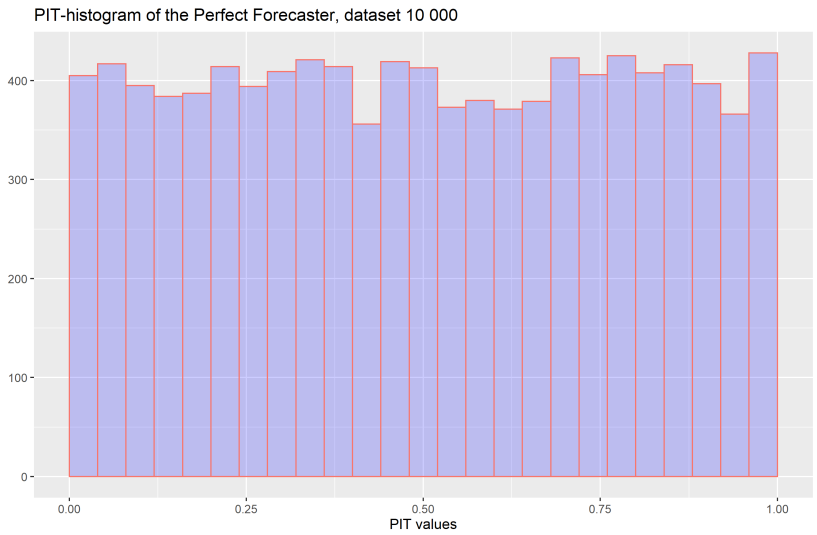

Figure A.1: Perfect PIT-histogram.

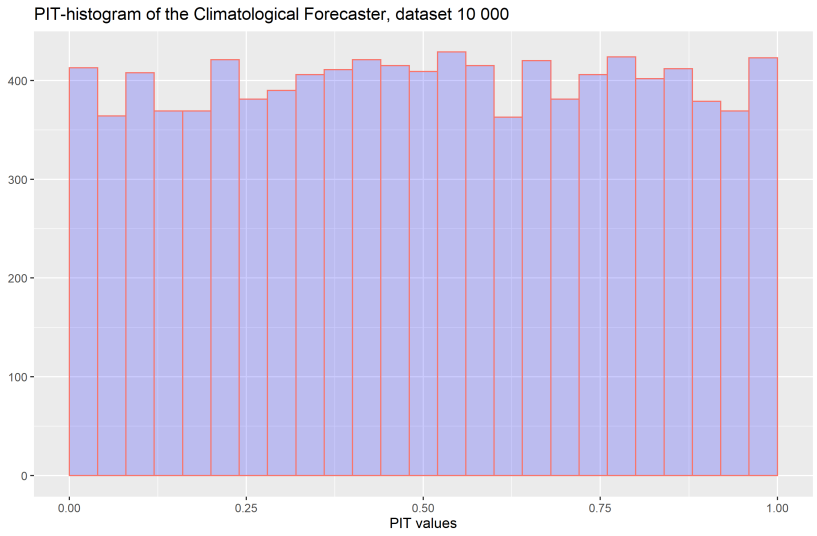

Figure A.2: Climatological PIT-histogram.

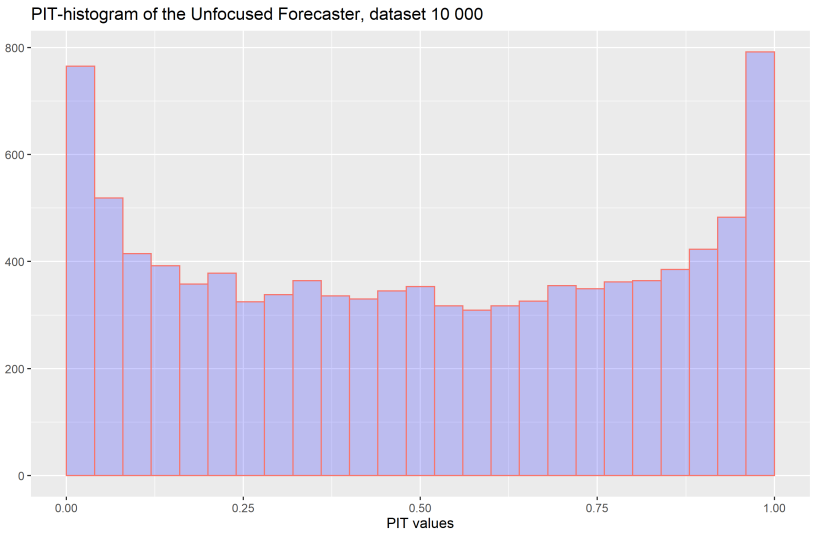

Figure A.3: Unfocused PIT-histogram.

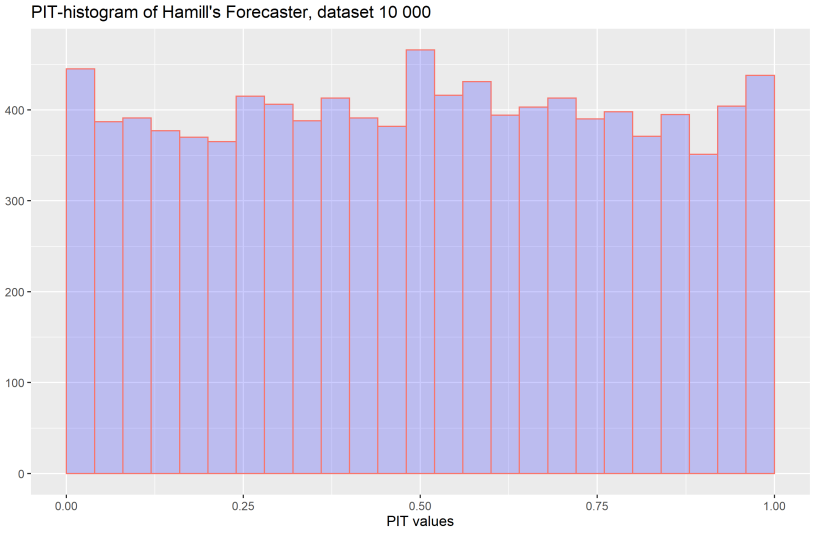

Figure A.4: Hamill PIT-histogram.

Appendix B

Histograms of uniformity p-values simulation chapter

In the following 4 pages we added the histograms which display the results of 100 000 simulations of uniformity tests for the perfect, climatological, Hamill's unfocused and Hamill's forecaster. In order to make them fit better, they are displayed sideways. Each title shows the test used, the forecaster used and the size of the dataset. Please note that the size of the dataset has nothing to do with the amount of values in the histogram. Each set produces one p-value, no matter the number of forecast-observation pairs and since we simulate each set 100 000 times, each histograms features 100 000 p-values no matter the dataset size, uniformity test and/or forecaster used.

Figure B.1: Histograms of p-values of the perfect forecaster.

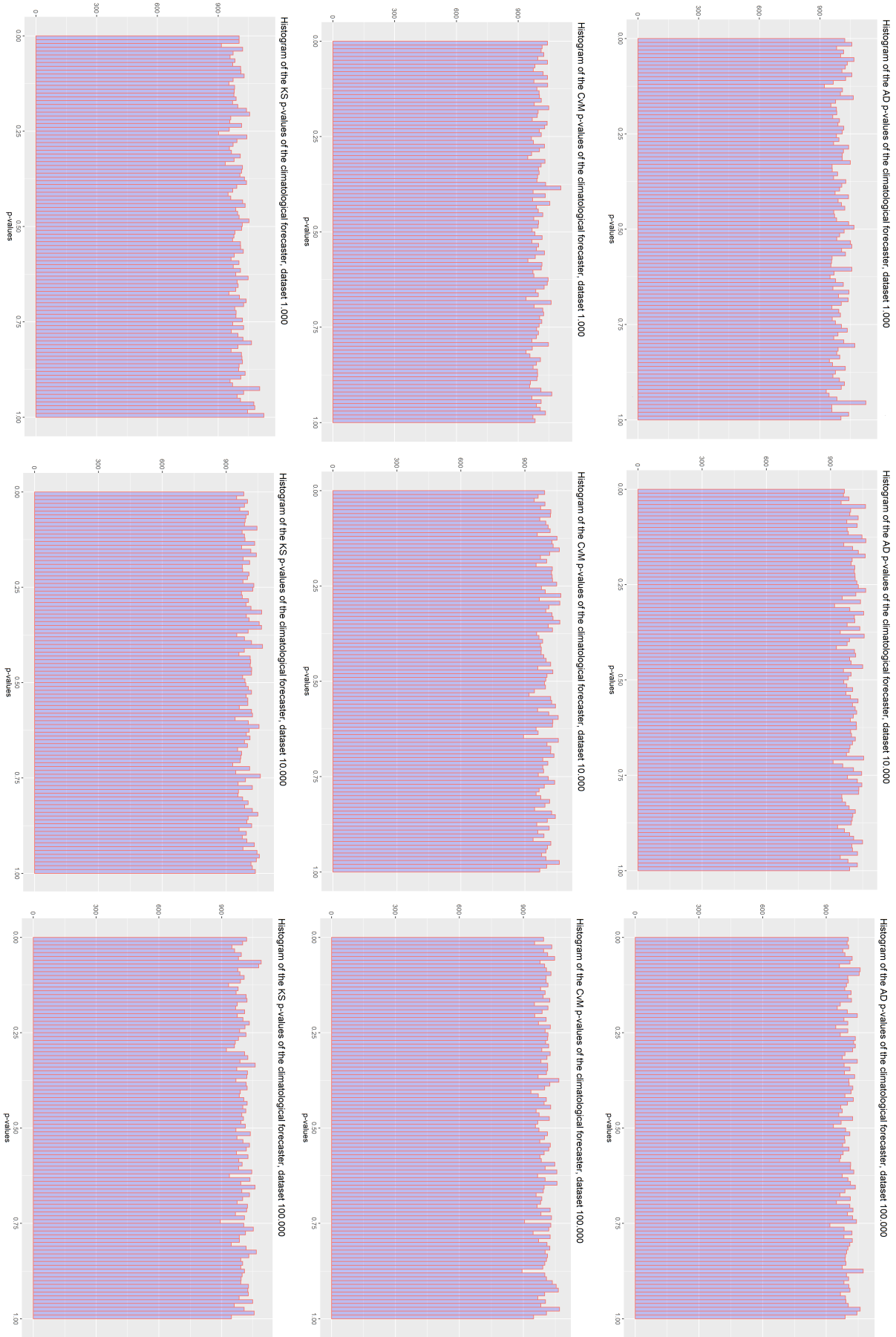

Figure B.2: Histograms of p-values of the climatological forecaster.

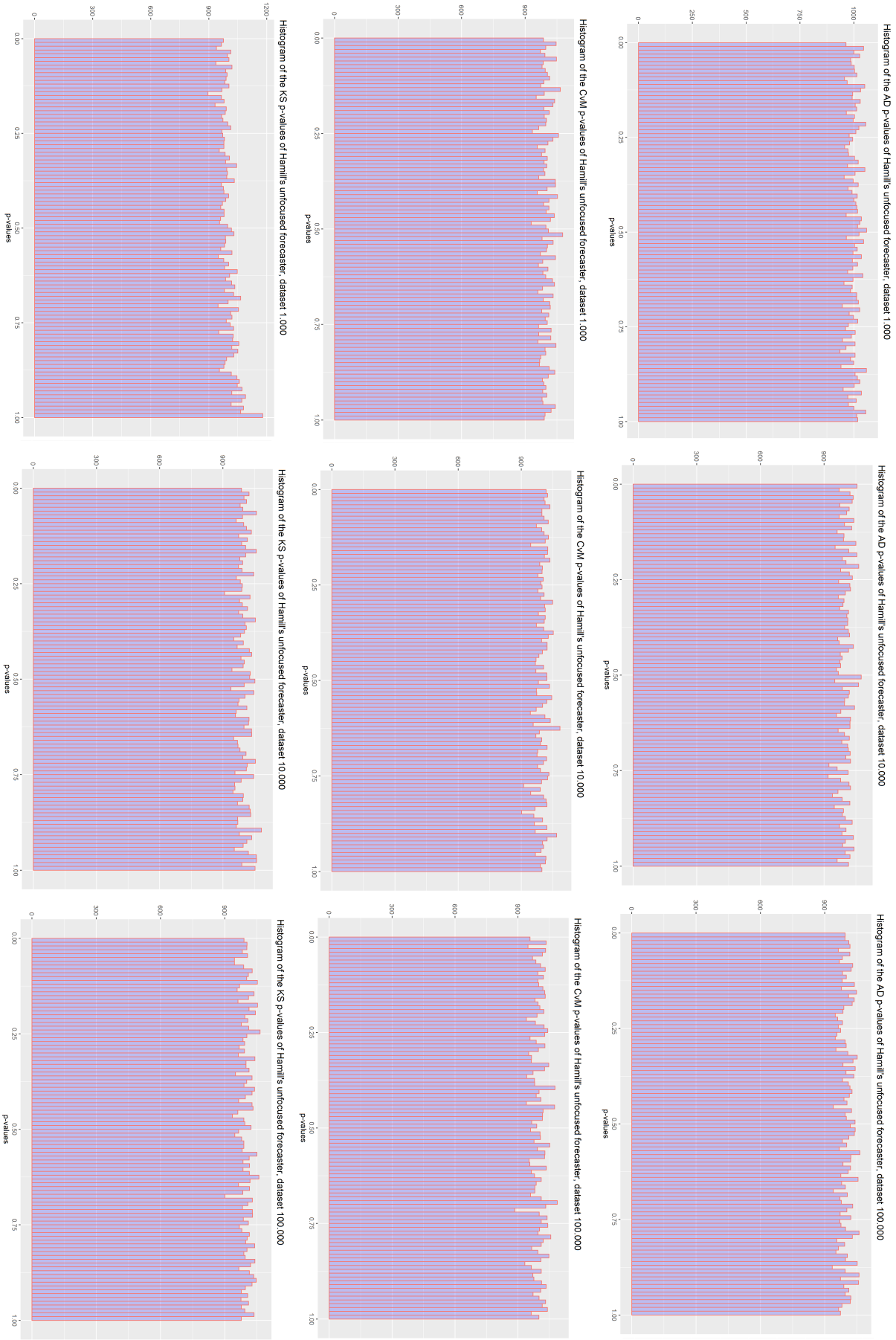

Figure B.3: Histograms of p-values of Hamill's unfocused forecaster.

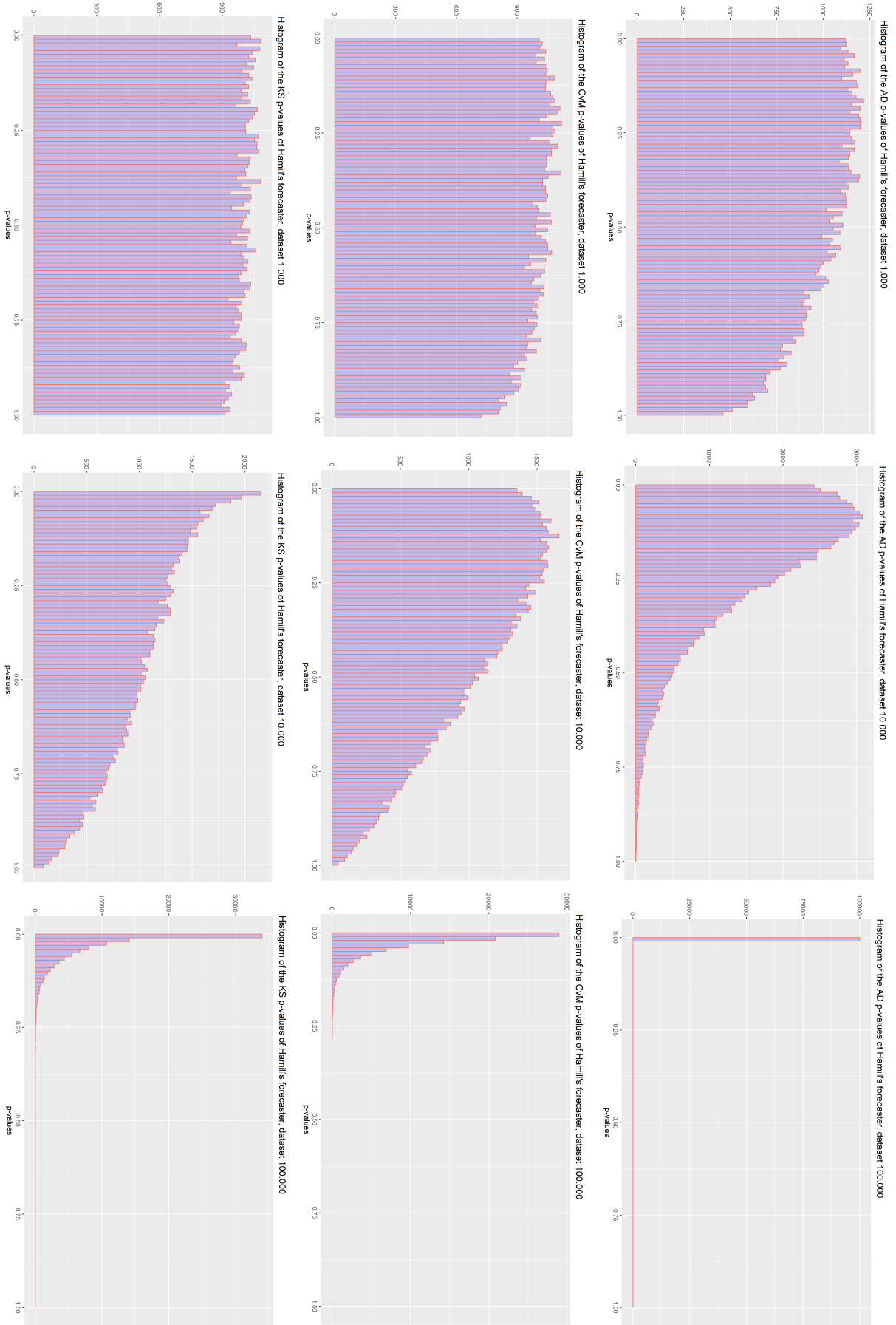

Figure B.4: Histograms of p-values of Hamill's forecaster.

Appendix C

γ- and σ-estimates

In section 4.2.1 we simulated γ-estimates for the 0.75-quantile using either the right tail of the observations or the CRPS values. The 0.75-quantile deemed to be not sufficient enough, that is why we reproduced the simulation using a 0.90-quantile and the 0.95-quantile for u. For these cases we have added the γ-estimates of the set of CRPS values of the perfect forecaster given $Y > u$. We will come back on this in our reflection in section 4.3.

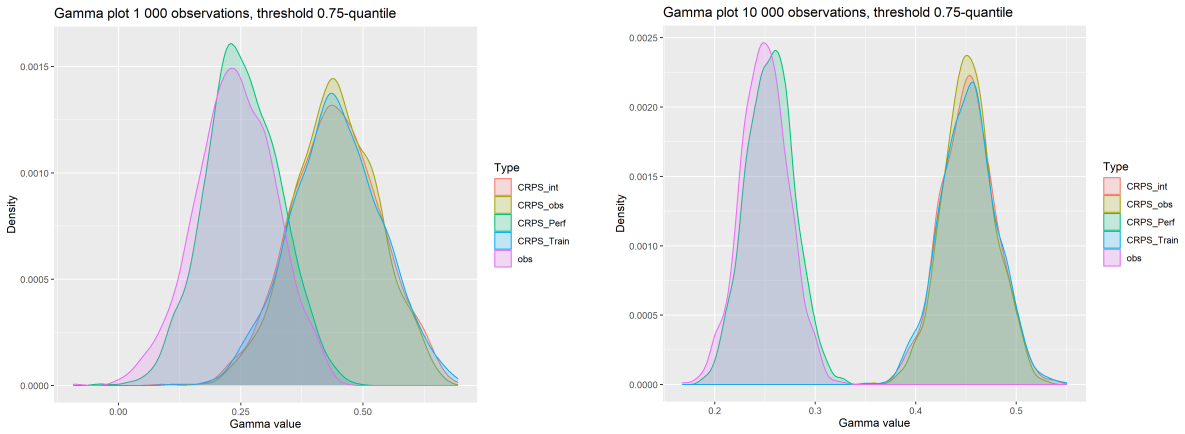

Figure C.1: γ-estimates using the 0.75-quantile for u.

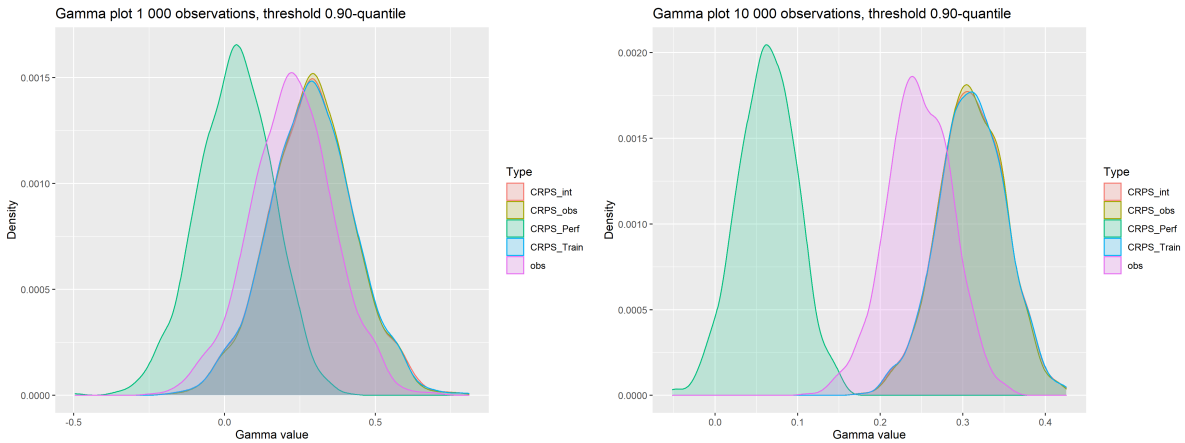

Figure C.2: γ-estimates using the 0.90-quantile for u.

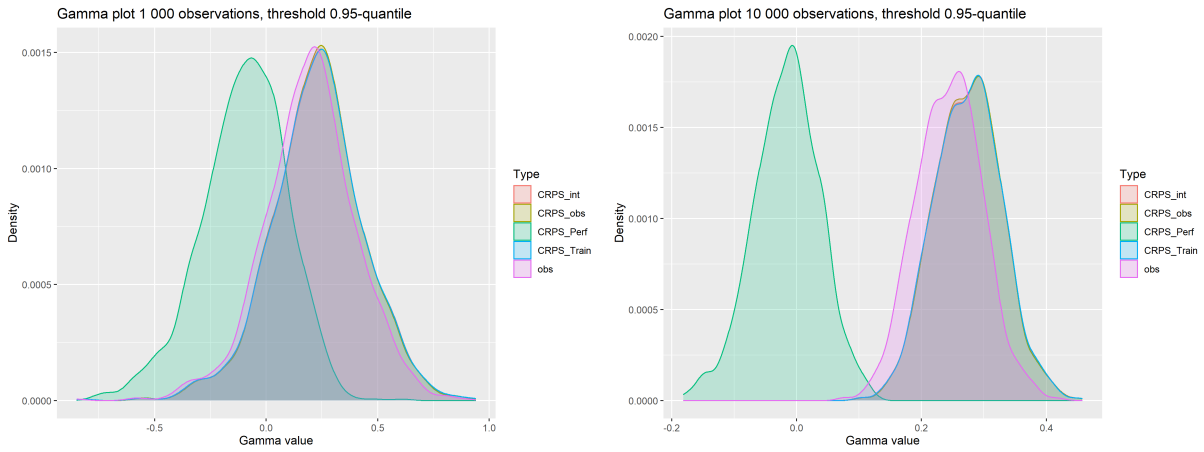

Figure C.3: γ-estimates using the 0.95-quantile for u.

C.1 Difference between γ-estimates

In this section we show the results of the differences in γ-estimates between the set of observations and the set of CRPS values with a more theoretical setting. These results are used in our results section 4.2.1.

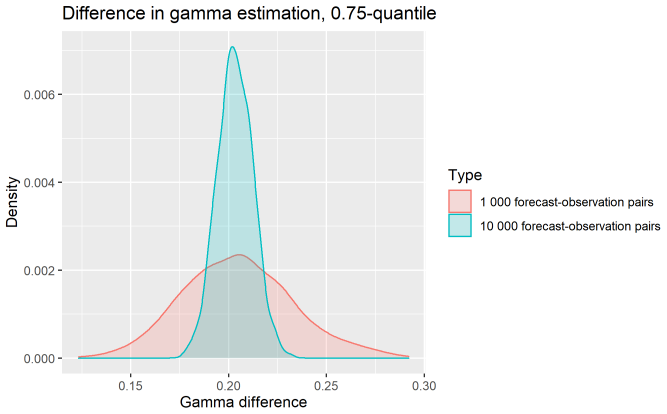

Figure C.4: Differences in γ-estimates at the 0.75-quantile.

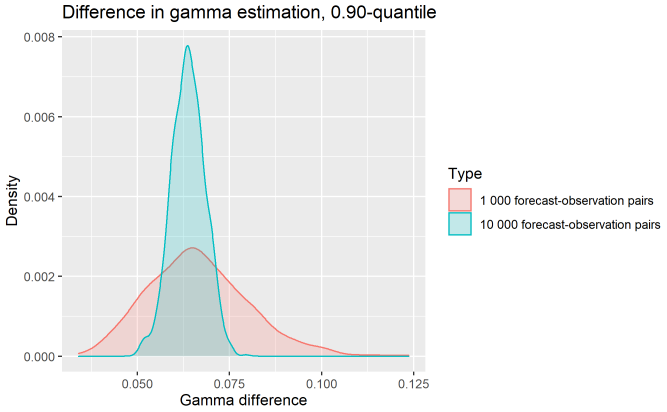

Figure C.5: Differences in γ-estimates at the 0.90-quantile.

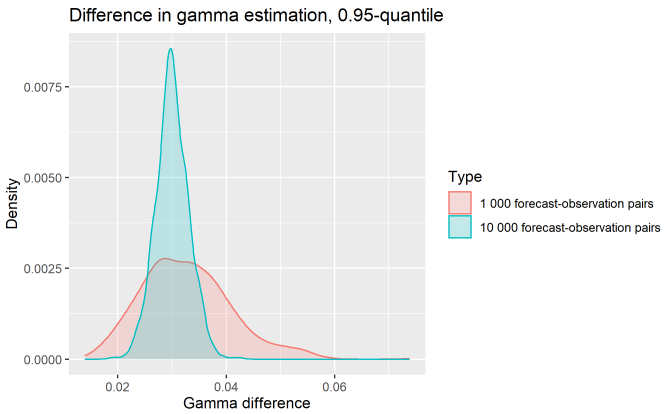

Figure C.6: Differences in γ-estimates at the 0.95-quantile.

C.2 Correlation figure full size

Full size figures of the correlation figure found in section 4.2.

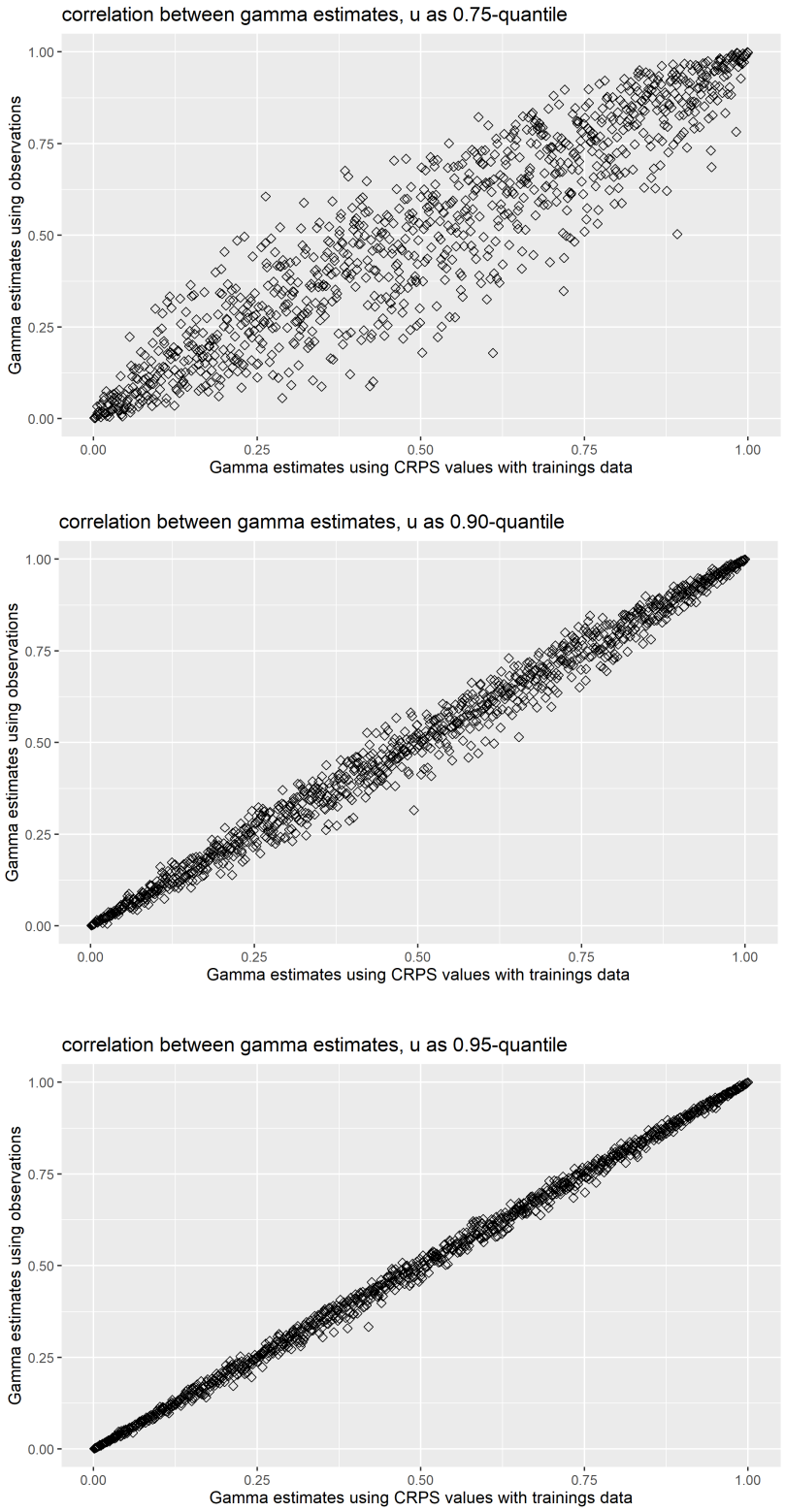

Figure C.7: Correlation γ-estimates using obs. or CRPS values at the different values of u.

C.3 Index percentages

In this section we put the table of index percentages when using $\sigma_w = \sigma_u + \gamma \times w$ instead of our corrected $\sigma_w = \sigma_u + \gamma \times (w - u)$ as a reference.

		1.000			10.000		
u	GP estimate ↓	CRPS obs.	CRPS train.	CRPS int.	CRPS obs.	CRPS train.	CRPS int.
0.75	Y obs.		1.52%	2.28%		0.00%	0.00%
	CRPS obs.	6.00%			0.04%		
	CRPS train.		6.37%			0.04%	
	CRPS int.			6.71%			0.04%
0.90	Y obs.		80.32%	79.63%		0.10%	0.10%
	CRPS obs.	75.63%			0.10%		
	CRPS train.		75.3%			0.10%	
	CRPS int.			75.3%			0.10%
0.95	Y obs.		93.73%	93.67%		2.26%	2.22%
	CRPS obs.	93.07%			2.19%		
	CRPS train.		92.63%			2.19%	
	CRPS int.			92.89%			2.23%

Table C.1: Percentages of the index values of the climatological forecaster equal to 0.

When we compare this table with the table which used the corrected calculation of σ_w, we see that the table which uses the corrected calculation of σ_w is more stable in any of the given scenarios.

C.4 Index percentages

In this section we show the figures of the standard deviation of the mean index value as presented in section 4.4.

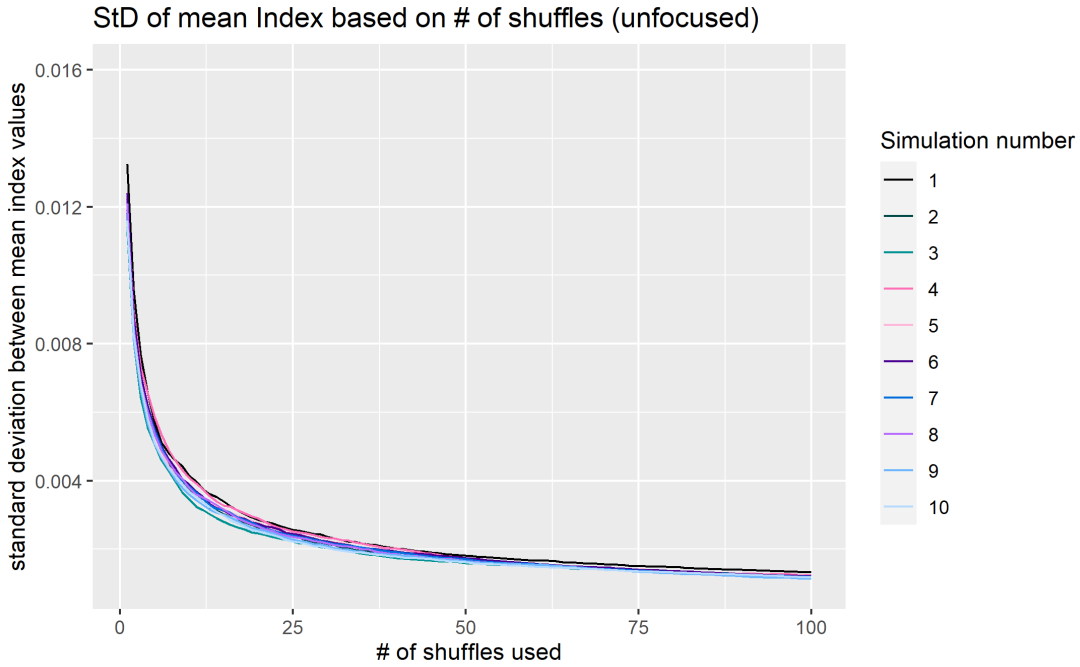

Figure C.8: Course of the standard deviation of the mean index value, depending on the number of shuffles used (unfocused forecaster).

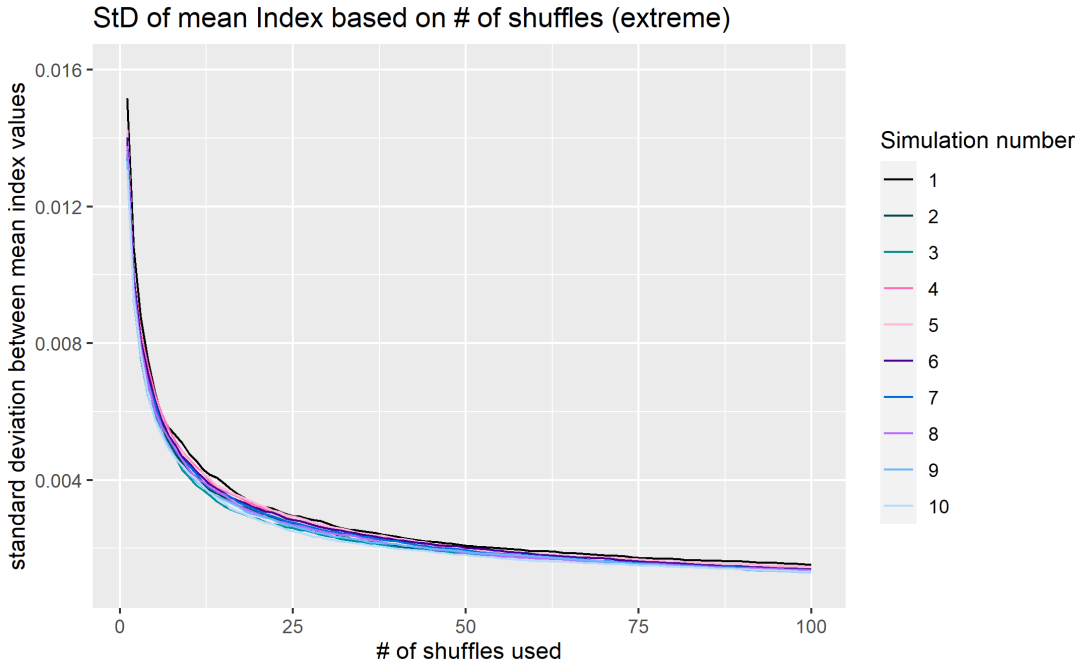

Figure C.9: Course of the standard deviation of the mean index value, depending on the number of shuffles used (extremist forecaster).

C.5 Index differences

In this section we show the histograms of the mean index values of the perfect forecaster, the unfocused forecaster and the extremist forecaster based on 1 000 shuffles.

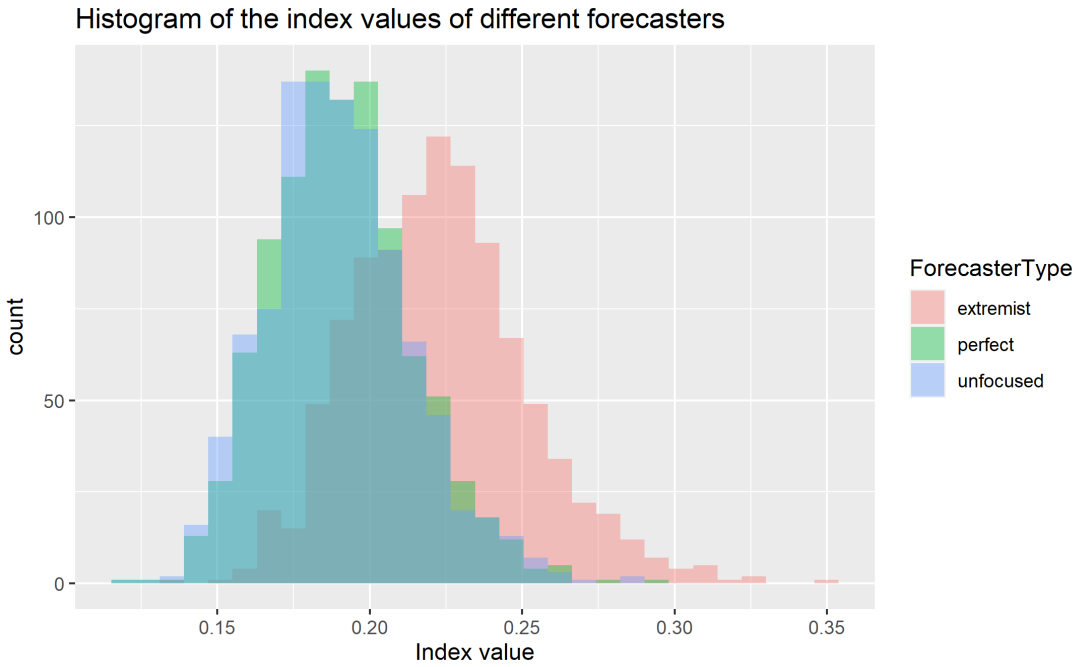

Figure C.10: The index values of the three forecasters.

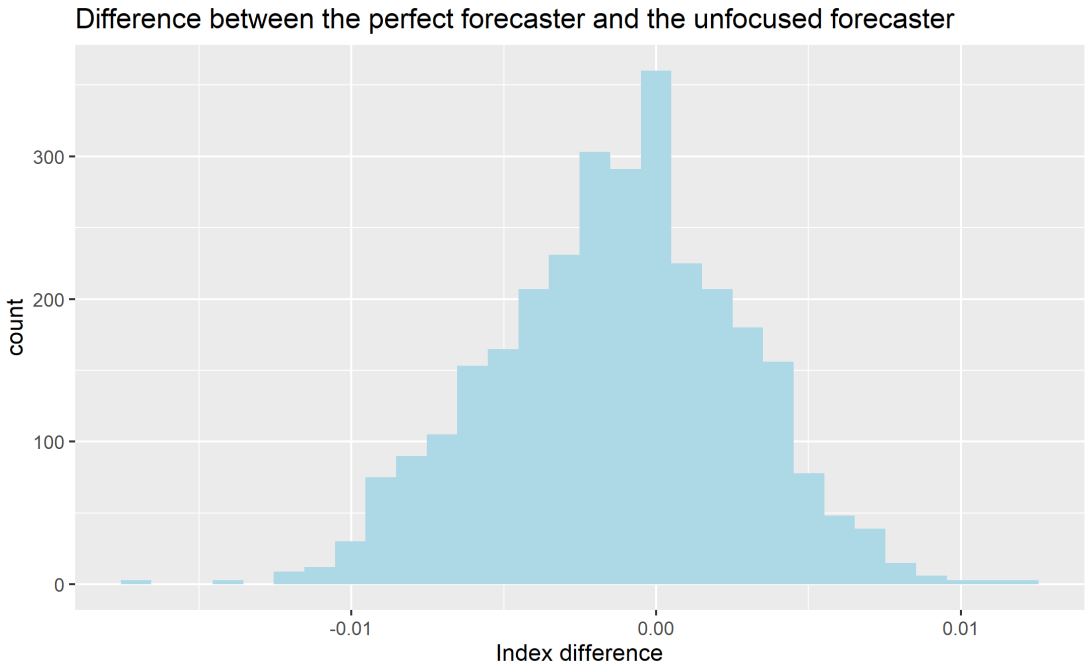

Figure C.11: Difference in index value between the perfect forecaster and the unfocused forecaster.

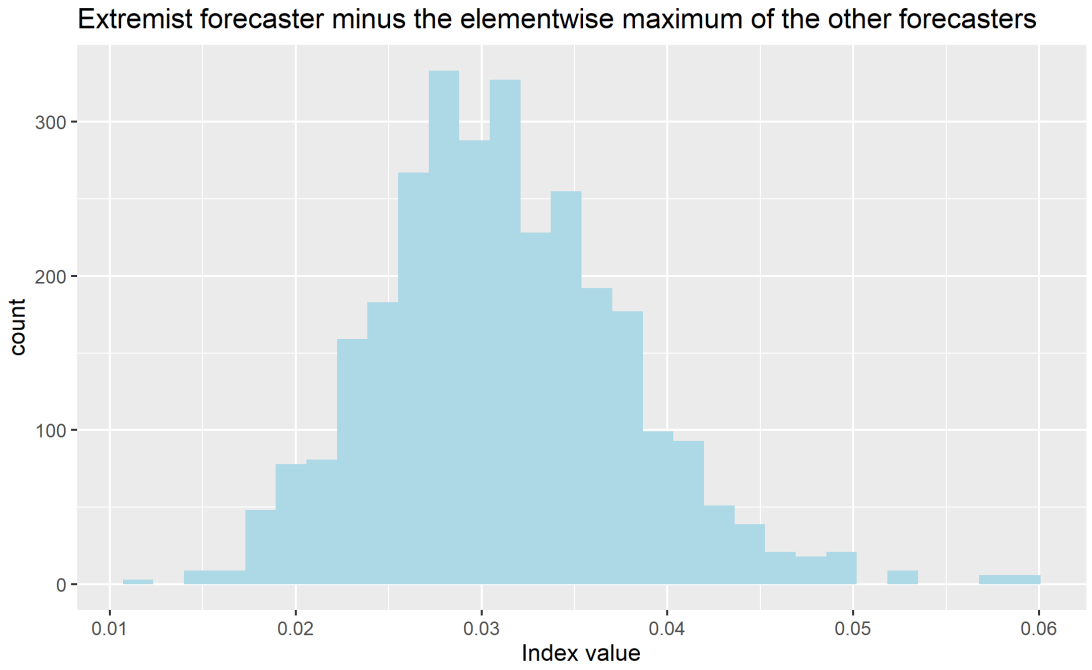

Figure C.12: Difference between the index value of the extremist forecaster and the element wise maximum of the index values of the perfect forecaster and the unfocused forecaster.

Appendix D

Uniform-Pareto model, climatological tail index

For the Uniform-Pareto model, the tail index of G is the following:

$$
\begin{aligned}
1 - G(y) &= \mathbb{P}(Y_t > y) \\
&= \int_0^C \mathbb{P}(Y_t > y | \gamma_t = \gamma) * f_{\gamma_t}(\gamma) d\gamma \\
&= \int_0^C y^{-\frac{1}{\gamma}} * \frac{1}{C} d\gamma \\
&\qquad \downarrow \text{here, } \frac{1}{\gamma} = U, \text{ so } \gamma = \frac{1}{U} \\
&= \int_{\frac{1}{C}}^{\infty} y^{-U} * \frac{1}{C} d\frac{1}{U} \\
&= \int_{\frac{1}{C}}^{\infty} y^{-U} * \frac{1}{C} * \frac{1}{U^2} dU
\end{aligned}
$$

We then get the following:

$$
\begin{aligned}
\lim_{t\to\infty} \frac{1 - G(yt)}{1 - G(t)} &= \lim_{t\to\infty} \frac{\int_{\frac{1}{C}}^{\infty} (ty)^{-U} * \cancel{\frac{1}{C}} * \frac{1}{U^2} dU}{\int_{\frac{1}{C}}^{\infty} t^{-U} * \cancel{\frac{1}{C}} * \frac{1}{U^2} dU}^{1} \\
&= \lim_{t\to\infty} \frac{\int_{\frac{1}{C}}^{\infty} (ty)^{-U} * \frac{1}{U^2} dU}{\int_{\frac{1}{C}}^{\infty} t^{-U} * \frac{1}{U^2} dU} \\
&\qquad \downarrow \text{ using the l'Hopital rule on } t \\
= \lim_{t\to\infty} \frac{\int_{\frac{1}{C}}^{\infty} -U * y * (ty)^{-U-1} * \frac{1}{U^2} dU}{\int_{\frac{1}{C}}^{\infty} -U * t^{-U-1} * \frac{1}{U^2} dU} &= \lim_{t\to\infty} \frac{-y \int_{\frac{1}{C}}^{\infty} (ty)^{-U-1} * \frac{1}{U} dU}{-1 * \int_{\frac{1}{C}}^{\infty} t^{-U-1} * \frac{1}{U} dU} \\
\lim_{t\to} \frac{\cancel{y} * \cancel{t^{-1}} * \cancel{y^{-1}} \int_{\frac{1}{C}}^{\infty} (ty)^{-U} * \frac{1}{U} dU}{\cancel{t^{-1}} \int_{\frac{1}{C}}^{\infty} t^{-U} * \frac{1}{U} dU} &= \lim_{t\to} \frac{\int_{\frac{1}{C}}^{\infty} (ty)^{-U} * \frac{1}{U} dU}{\int_{\frac{1}{C}}^{\infty} t^{-U} * \frac{1}{U} dU} \\
&\qquad \downarrow \text{ using the l'Hopital rule again } t
\end{aligned}
$$

[1] since $\frac{1}{C}$ is a constant, it can be crossed out

$$= \lim_{t\to\infty} \frac{\int_{\frac{1}{C}}^{\infty} -U * y * (ty)^{-U-1} * \frac{1}{U} dU}{\int_{\frac{1}{C}}^{\infty} -U * t^{-U-1} * \frac{1}{U} dU} = \lim_{t\to\infty} \frac{-y \int_{\frac{1}{C}}^{\infty} (ty)^{-U-1} dU}{-1 * \int_{\frac{1}{C}}^{\infty} t^{-U-1} dU}$$

$$\lim_{t\to\infty} \frac{\int_{\frac{1}{C}}^{\infty} (ty)^{-U} dU}{\int_{\frac{1}{C}}^{\infty} t^{-U} dU}$$

$$= \lim_{t\to\infty} \frac{\left[\frac{-(ty)^{-U}}{\ln(ty)}\right]_{\frac{1}{C}}^{\infty}}{\left[\frac{-(t)^{-U}}{\ln(t)}\right]_{\frac{1}{C}}^{\infty}}$$

$$\downarrow \text{ for } U = \infty, \text{ this goes to 0 so we get:}$$

$$= \lim_{t\to\infty} \frac{\frac{-(ty)^{-\frac{1}{C}}}{\ln(ty)}}{\frac{-(t)^{-\frac{1}{C}}}{\ln(t)}} = \lim_{t\to\infty} \frac{-(ty)^{-\frac{1}{C}} * \ln(t)}{-(t)^{-\frac{1}{C}} * \ln(ty)} = y^{-\frac{1}{C}}$$

Appendix E

Beta-Pareto model, climatological tail index

For the Beta-Pareto model, the tail index of G is the following:

$$
\begin{aligned}
1 - G(y) &= \mathbb{P}(Y_t > y) \\
&= \int_0^1 \mathbb{P}(Y_t > y | \gamma_t = \gamma) * f_{\gamma_t}(\gamma) d\gamma \\
&= \int_0^1 y^{-\frac{1}{\gamma}} * \frac{\gamma^{\alpha-1} * (1-\gamma)^{\beta-1}}{B(\alpha,\beta)} d\gamma \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{-\frac{1}{\gamma}} * \gamma^{\alpha-1} * (1-\gamma)^{\beta-1} d\gamma \\
&\qquad\qquad \downarrow \text{here, } \frac{1}{\gamma} = U, \text{ so } \gamma = \frac{1}{U} \\
&= \frac{1}{B(\alpha,\beta)} \int_1^\infty y^{-U} * \left(\frac{1}{U}\right)^{\alpha-1} * \left(1 - \frac{1}{U}\right)^{\beta-1} d\frac{1}{U} \\
&= \frac{1}{B(\alpha,\beta)} \int_1^\infty y^{-U} * \left(\frac{1}{U}\right)^{\alpha-1} * \left(\frac{U-1}{U}\right)^{\beta-1} * \left(\frac{1}{U}\right)^2 dU \\
&= \frac{1}{B(\alpha,\beta)} \int_1^\infty y^{-U} * \left(\frac{1}{U}\right)^{\alpha-1} * (U-1)^{\beta-1} * \left(\frac{1}{U}\right)^{\beta-1} * \left(\frac{1}{U}\right)^2 dU \\
&= \frac{1}{B(\alpha,\beta)} \int_1^\infty y^{-U} * \left(\frac{1}{U}\right)^{\alpha+\beta} * (U-1)^{\beta-1} dU
\end{aligned}
$$

We then get the following:

$$\lim_{t\to\infty}\frac{1-G(yt)}{1-G(t)}=\lim_{t\to\infty}\frac{\cancel{\frac{1}{B(\alpha,\beta)}}\int_1^\infty (yt)^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}{\cancel{\frac{1}{B(\alpha,\beta)}}\int_1^\infty t^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}$$

$$=\lim_{t\to\infty}\frac{\int_1^\infty (yt)^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}{\int_1^\infty t^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}$$

$\downarrow$ Let's see where the l'Hopital rule on t will take this:

$$=\lim_{t\to\infty}\frac{\int_1^\infty -U*y*(yt)^{-U-1}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}{\int_1^\infty -U*t^{-U-1}*\left(\frac{1}{U}\right)^{\alpha+\beta}*(U-1)^{\beta-1}dU}=\lim_{t\to\infty}\frac{-y\int_1^\infty (yt)^{-U-1}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}{-1*\int_1^\infty t^{-U-1}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}$$

$$\lim_{t\to\infty}\frac{\cancel{y}*\cancel{t^{-1}}*\cancel{y^{-1}}\int_1^\infty (yt)^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}{\cancel{t^{-1}}\int_1^\infty t^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}=\lim_{t\to\infty}\frac{\int_1^\infty (yt)^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}{\int_1^\infty t^{-U}*\left(\frac{1}{U}\right)^{\alpha+\beta-1}*(U-1)^{\beta-1}dU}$$

Here you see that the effect of the l'Hopital rule is the deceasing of the exponent in the term $\frac{1}{U}^{\alpha+\beta}$ by one.

So if we use the l'Hopital rule $\alpha+\beta$ times, which is possible since $\alpha,\beta\in\mathbb{N}_{>0}$, we will get:

$$\lim_{t\to\infty}\frac{\int_1^\infty (yt)^{-U}*(U-1)^{\beta-1}dU}{\int_1^\infty t^{-U}*(U-1)^{\beta-1}dU}$$

Here we see already that the α has no effect on the tail index.

Furthermore, if $\beta=1$, we get the same equation we saw in Appendix D,

which was already proved to be equal to y^{-1}.

For $\beta>1$ we get the following:

$$\lim_{t\to\infty}\frac{\int_1^\infty (yt)^{-U}*(U-1)^{\beta-1}dU}{\int_1^\infty t^{-U}*(U-1)^{\beta-1}dU}=\lim_{t\to\infty}\frac{\left[\frac{-(yt)^{-U}*\sum_{i=0}^{\beta-1}\frac{(\beta-1)!}{(\beta-1-i)!}*(U-1)^{\beta-i-1}*\ln(yt)^{\beta-1-i}}{\ln(yt)^\beta}\right]_1^\infty}{\left[\frac{-(t)^{-U}*\sum_{i=0}^{\beta-1}\frac{(\beta-1)!}{(\beta-1-i)!}*(U-1)^{\beta-i-1}*\ln(t)^{\beta-1-i}}{\ln(t)^\beta}\right]_1^\infty}$$

for $U=\infty$, this goes to 0 and for $U=1, i\neq\beta-1$ this also goes to 0 so we get: $\downarrow$

$$\lim_{t\to\infty}\frac{\frac{-(yt)^{-1}*\cancel{(\beta-1)!}}{\ln(yt)^\beta}}{\frac{-(t)^{-1}*\cancel{(\beta-1)!}}{\ln(t)^\beta}}=\lim_{t\to\infty}\frac{-(yt)^{-1}*\ln(t)^\beta}{-(t)^{-1}*\ln(yt)^\beta}=y^{-1}$$

Appendix F

Results new weight functions on NN-model

In this Appendix, we briefly mention the results of the effect of the new weight functions on the CRPS values of the forecasters of our NN-model as given in Chapter 3. We have added one more forecaster to our NN-model, the underdispersed forecaster.
The NN-model is given as:

Forecaster	Distribution	Variable (if applicable)
The information	$\Delta_t \sim N(0,1)$	
The observation	$y_t \sim N(\Delta_t, 1)$	
The perfect forecaster	$F_t^P \sim N(\Delta_t, 1)$	
The climatological forecaster	$F_t^C \sim N(0, 2)$	
The biased forecaster	$F_t^B \sim N(\Delta_t + b, 1)$	$b = 1$
The sign-biased forecaster	$F_t^S \sim N(-\Delta_t, 1)$	
The overdispersed forecaster	$F_t^O \sim N(\Delta_t, \frac{9}{4})$	
The unfocused forecaster	$F_t^U \sim \tau * N(\Delta_t, 1) + (1-\tau) * N(\Delta_t + \gamma, 1)$	$\tau \in \{0,1\}, \gamma \in \{-1,1\}$
Hamill's unfocused forecaster	$F_t^{HU} \sim \frac{1}{2}\{N(\Delta_t, 1) + N(\Delta_t + \phi, 1)\}$	$\phi \in \{-1,1\}$
Hamill's forecaster	$F_t^H \sim N(\Delta_t + \delta, \sigma^2)$	$(\delta, \sigma^2) \in \{(\frac{1}{2}, 1); (-\frac{1}{2}, 1); (0, \frac{169}{100})\}$
The overd. sign-b. forecaster	$F_t^{OS} \sim N(-\Delta_t, \frac{9}{4})$	
The underdispersed forecaster	$F_t^D \sim N(\Delta_t, \frac{7}{10})$	

Table F.1: The Normal-Normal model.

The weight functions we implemented for this simulation are the following:
Type I: $wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} |F_t(x) - \mathbb{1}\{x \geq y_t\}|^{w(\cdot)} \ \mathrm{dx}$,
with weight function $w(\cdot)$ equal to either $\frac{3}{2}$ or 4.
Type II: $wCRPS(F_t, y_t) = \left(\int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 \ \mathrm{dx}\right)^{w(\cdot)}$,
with weight function $w(\cdot)$ equal to either 2 or 4.
Type III: $wCRPS(F_t, y_t) = \int_{-\infty}^{\infty} (F_t(x) - \mathbb{1}\{x \geq y_t\})^2 * w(x, y_t) \ \mathrm{dx}$,
with weight function $w(\cdot)$ equal to either $|x - y_i|$ or $(1 + |x - y_i|)$.

We simulated the CRPS score for each of the 10 forecasters for 10 000 forecast-observation pairs and simulated these results 10 times. We note that 10 simulations might not be enough to simulate the small changes, however this simulation is only used to indicate a path of research.

Forecaster	original	Type I		Type II		Type III	
		$\frac{3}{2}$	4	2	4	$\lvert x-y_i\rvert$	$(1+\lvert x-y_i\rvert)$
perfect forecaster	0.5643	0.7584	0.2672	0.4807	0.8852	0.3909	0.9552
underdispersed forecaster	0.5792	0.7187	0.3463	0.5580	1.2494	0.3711	0.9502
overdispersed forecaster	0.5922	0.8755	0.2021	0.4490	0.5817	0.4820	1.0742
Hamill's forecaster	0.6134	0.8265	0.2897	0.5647	1.2146	0.4615	1.0749
Hamill's unfocused forecaster	0.6311	0.8492	0.2968	0.5945	1.3072	0.4875	1.1187
unfocused forecaster	0.7001	0.8974	0.3815	0.7774	2.2955	0.5614	1.2615
climatological forecaster	0.7976	1.0721	0.3778	0.9623	3.5785	0.7819	1.5795
biased forecaster	0.8356	1.0360	0.4956	1.0740	3.7132	0.7317	1.5673
overd. sign-b. forecaster	1.3008	1.6007	0.7961	2.8458	32.7295	1.8549	3.1557
sign-biased forecaster	1.3884	1.5939	1.0156	3.3605	43.2378	1.9126	3.3010

Table F.2: Mean scores of forecasters using different weight functions (the order of the forecasters in this table is according to their ranking using no weight function).

Forecaster	original	Type I		Type II		Type III	
		$\frac{3}{2}$	4	2	4	$\lvert x-y_i\rvert$	$(1+\lvert x-y_i\rvert)$
perfect forecaster	#1	**#2**	**#2**	**#2**	**#2**	**#2**	**#2**
underdispersed forecaster	#2	**#1**	**#5**	**#3**	**#4**	**#1**	**#1**
overdispersed forecaster	#3	**#5**	**#1**	**#1**	**#1**	#3	#3
Hamill's forecaster	#4	**#3**	**#3**	#4	**#3**	#4	#4
Hamill's unfocused forecaster	#5	**#4**	**#4**	#5	#5	#5	#5
unfocused forecaster	#6	#6	**#7**	#6	#6	#6	#6
climatological forecaster	#7	**#8**	**#6**	#7	#7	**#8**	**#8**
biased forecaster	#8	**#7**	#8	#8	#8	**#7**	**#7**
overd. sign-b. forecaster	#9	**#10**	#9	#9	#9	#9	#9
sign-biased forecaster	#10	**#9**	#10	#10	#10	#10	#10

Table F.3: Ranking order of forecasters using different weight functions.

Here, the bold rankings indicate a change of ranking compared to the original CRPS. The most important element to note is that the perfect forecaster is dethroned at every weight function. We want to add that the perfect forecaster placed second in all 10 simulations of each weight function, not only some of the simulations.
In the following table we have added the mean percentile change of the CRPS value of the forecasters when comparing a new weight function with the original CRPS.

Forecaster	Type I		Type II		Type III	
	$\frac{3}{2}$	4	2	4	$\lvert x - y_i \rvert$	$(1 + \lvert x - y_i \rvert)$
perfect forecaster	34.39%	−52.64%	−14.81%	56.87%	−30.73%	69.27%
underdispersed forecaster	**24.10**%	−40.21%	−3.65%	115.72%	**−35.93**%	**64.07**%
overdispersed forecaster	47.84%	**−65.87**%	**−24.18**%	**−1.76**%	−18.60%	81.40%
Hamill's forecaster	34.74%	**−52.77**%	−7.94%	98.00%	−24.76%	75.24%
Hamill's unfocused forecaster	34.55%	**−52.97**%	−5.81%	107.13%	−22.75%	77.25%
unfocused forecaster	**28.17**%	−45.51%	11.03%	227.87%	−19.82%	80.18%
climatological forecaster	34.41%	−52.64%	20.65%	348.64%	−1.97%	98.03%
biased forecaster	**23.98**%	−40.69%	28.53%	344.37%	−12.43%	87.57%
overd. sign-b. forecaster	**23.06**%	−40.21%	118.77%	2416.07%	42.59%	142.59%
sign-biased forecaster	**14.80**%	−26.85%	142.05%	3014.25%	37.76%	137.76%

Table F.4: Mean percentile change between the original CRPS value and the CRPS value using a different weight functions.

Here, the bold percentages depict the forecasters which decreased their CRPS value more, or increased it less, than the perfect forecaster. These results seems to show that skewing the emphasis towards distances of the cdfs between 0 and 1 favours forecasters with a higher sharpness, whereas skewing the emphasis away from distances of the cdfs between 0 and 1 favours forecasters with a lower sharpness. Furthermore, skewing the emphasis towards values further away from the observation value favours forecasters with a higher sharpness. We conclude from this that we cannot alter the CRPS by putting more or less emphasis on sharpness related conditions. We therefore argue for future research to research if it is possible to create a weight function which solely adds an extra calibration related weight to the CRPS and what the implications of such a weight function are.

Bibliography

Bopp, G. P. & Shaby, B. A. (2017), 'An exponential–gamma mixture model for extreme santa ana winds', *Environmetrics* **28**(8), e2476.

Brehmer, J. R., Strokorb, K. et al. (2019), 'Why scoring functions cannot assess tail properties', *Electronic Journal of Statistics* **13**(2), 4015–4034.

Brier, G. W. (1950), 'Verification of forecasts expressed in terms of probability', *Monthly weather review* **78**(1), 1–3.

Bröcker, J. (2009), 'Reliability, sufficiency, and the decomposition of proper scores', *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* **135**(643), 1512–1519.

Bröcker, J. & Smith, L. A. (2007), 'Scoring probabilistic forecasts: The importance of being proper', *Weather and Forecasting* **22**(2), 382–388.

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E., Brown, B. & Mason, S. (2008), 'Forecast verification: current status and future directions', *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* **15**(1), 3–18.

Cramér, H. (1928), 'On the composition of elementary errors: First paper: Mathematical deductions', *Scandinavian Actuarial Journal* **1928**(1), 13–74.

Czado, C., Gneiting, T. & Held, L. (2009), 'Predictive model assessment for count data', *Biometrics* **65**(4), 1254–1261.

Dawid, A. P. (1984), 'Present position and potential developments: Some personal views statistical theory the prequential approach', *Journal of the Royal Statistical Society: Series A (General)* **147**(2), 278–290.

De Haan, L. & Ferreira, A. (2007), *Extreme value theory: an introduction*, Springer Science & Business Media.

Ehm, W., Gneiting, T., Jordan, A. & Krüger, F. (2016), 'Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings', *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 505–562.

Epstein, E. S. (1969), 'A scoring system for probability forecasts of ranked categories', *Journal of Applied Meteorology (1962-1982)* **8**(6), 985–987.

Friederichs, P. & Thorarinsdottir, T. L. (2012), 'Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction', *Environmetrics* **23**(7), 579–594.

Gneiting, T. (2008), 'Probabilistic forecasting', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 319–321.

Gneiting, T. (2011), 'Making and evaluating point forecasts', *Journal of the American Statistical Association* **106**(494), 746–762.

Gneiting, T., Balabdaoui, F. & Raftery, A. E. (2007), 'Probabilistic forecasts, calibration and sharpness', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2), 243–268.

Gneiting, T. & Katzfuss, M. (2014), 'Probabilistic forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American statistical Association* **102**(477), 359–378.

Gneiting, T., Ranjan, R. et al. (2013), 'Combining predictive distributions', *Electronic Journal of Statistics* **7**, 1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. & Johnson, N. A. (2008), 'Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds', *Test* **17**(2), 211.

Hamill, T. M. (2001), 'Interpretation of rank histograms for verifying ensemble forecasts', *Monthly Weather Review* **129**(3), 550–560.

Hersbach, H. (2000), 'Decomposition of the continuous ranked probability score for ensemble prediction systems', *Weather and Forecasting* **15**(5), 559–570.

Jolliffe, I. T. (2008), 'The impenetrable hedge: A note on propriety, equitability and consistency', *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* **15**(1), 25–29.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., Gneiting, T. et al. (2017), 'Forecaster's dilemma: Extreme events and forecast evaluation', *Statistical Science* **32**(1), 106–127.

Matheson, J. E. & Winkler, R. L. (1976), 'Scoring rules for continuous probability distributions', *Management science* **22**(10), 1087–1096.

Mert, M. & Saykan, Y. (2005), 'On a bonus–malus system where the claim frequency distribution is geometric and the claim severity distribution is pareto', *Hacettepe Journal of Mathematics and Statistics* **34**, 75–81.

Moore, P. (2015), 'The birth of the weather forecast', *66c. com, April* **30**.

Murphy, A. H. (1971), 'A note on the ranked probability score', *Journal of Applied Meteorology* **10**(1), 155–156.

Murphy, A. H. (1993), 'What is a good forecast? an essay on the nature of goodness in weather forecasting', *Weather and forecasting* **8**(2), 281–293.

Murphy, A. H. & Winkler, R. L. (1984), 'Probability forecasting in meteorology', *Journal of the American Statistical Association* **79**(387), 489–500.

Murphy, A. H. & Winkler, R. L. (1987), 'A general framework for forecast verification', *Monthly weather review* **115**(7), 1330–1338.

Pearson, K. (1933), 'On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random', *Biometrika* pp. 379–410.

Reiss, R.-D. & Thomas, M. (2007), *Statistical Analysis of Extreme Values : With Applications to Insurance, Finance, Hydrology and Other Fields*, Springer Basel AG.

Rosenblatt, M. (1952), 'Remarks on a multivariate transformation', *The annals of mathematical statistics* **23**(3), 470–472.

Stevenson, M. et al. (2006), *Ian T. Jolliffe and David B. Stephenson, Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley and Sons, Chichester (2003) ISBN 0-471-49759-2*, Vol. 22, Elsevier.

Taillardat, M., Fougères, A.-L., Naveau, P. & de Fondeville, R. (2019), 'Extreme events evaluation using crps distributions', *arXiv preprint arXiv:1905.04022* .

Tsyplakov, A. (2013), 'Evaluation of probabilistic forecasts: proper scoring rules and moments', *Available at SSRN 2236605* .

Von Mises, R. (1928), 'Statistik und wahrheit', *Julius Springer* **20**.

Von Mises, R. (1936), 'La distribution de la plus grande de n valuers', *Rev. math. Union interbalcanique* **1**, 141–160.

Wang, S. (1998), 'Aggregation of correlated risk portfolios: models and algorithms', **85**(163), 848–939.

Wong, F. & Collins, J. J. (2020), 'Evidence that coronavirus superspreading is fat-tailed', *Proceedings of the National Academy of Sciences* **117**(47), 29416–29418.