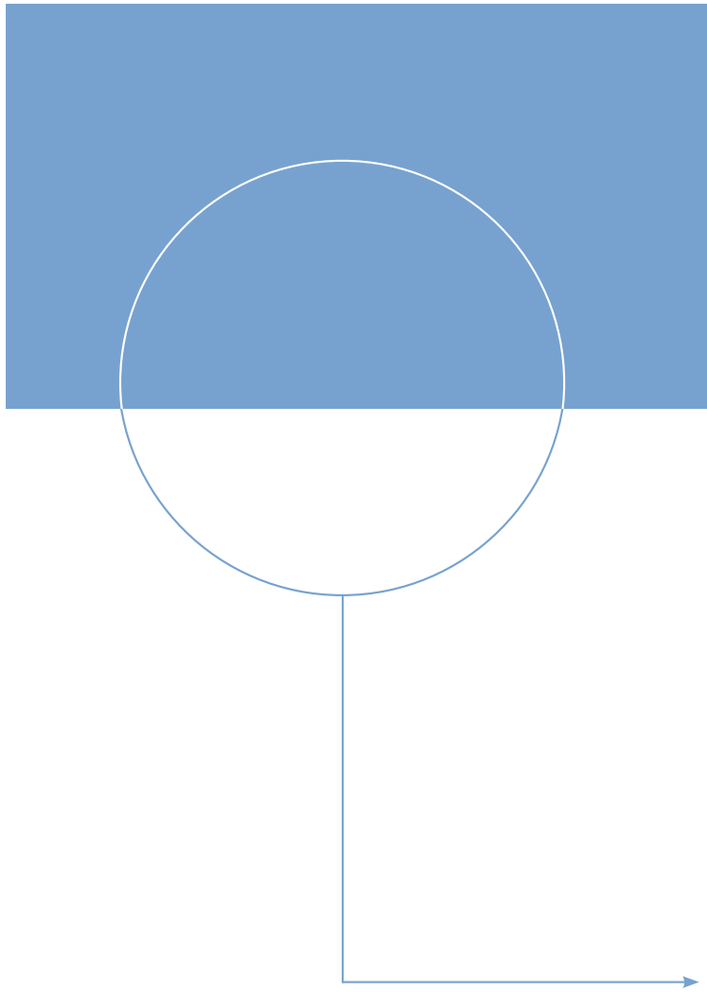

Master's thesis

2019

Silius Mortensønn Vandeskog

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

Silius Mortensønn Vandeskog

Modelling diurnal temperature range in Norway

Statistical modelling and spatial interpolation with the Five-Parameter Lambda Distribution

June 2019

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Modelling diurnal temperature range in Norway

Statistical modelling and spatial interpolation with the Five-Parameter Lambda Distribution

Silius Mortensønn Vandeskog

Applied Physics and Mathematics
Submission date: June 2019
Supervisor: Ingelin Steinsland, IMF
Co-supervisor: Thordis L. Thorarinsdottir, NR

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

We model the distribution of diurnal temperature range with the Five-Parameter Lambda Distribution (FPLD). Both local and spatial modelling with the FPLD is performed. For local parameter estimation of the FPLD we apply the method of quantiles, which estimates parameters by minimising the distance between two quantile functions. Quantile regression with explanatory variables is performed to model diurnal temperature range at locations without temperature observations. We introduce a new method for spatial interpolation of parametric distributions in order to perform spatial modelling of the FPLD. The new interpolation method combines quantile regression with the method of quantiles. Asymptotic conditions for consistency of the parameter estimators for the FPLD are presented. Additionally, simulation studies are performed for numerical evaluation of the proposed methods. The FPLD is fitted to 30 years of daily observations of diurnal temperature range from 55 weather stations in the southern parts of Norway. Modelling is performed independently for each season of the year. The FPLD shows much promise as a model for diurnal temperature range. The local parameter estimation, which uses the method of quantiles, is quite successful and a good model fit is observed for almost all the available data. The new interpolation method for spatial parameter estimation of the FPLD also shows much promise. Using this method, we are able to model the FPLD with a good fit to diurnal temperature range for winter, spring and autumn. During summer, the model fit is mediocre.

Sammendrag

Vi modellerer fordelingen av den døgnlige variasjonsbredden til temperatur med femparameter-lambdafordelingen (FPLF). Både lokal og regional modelltilpasning av en FPLF utføres. Lokal modelltilpasning blir utført ved hjelp av kvantiltilpasningsmetoden, som estimerer parametere ved å minimere avstanden mellom to kvantilfunksjoner. For å modellere variasjonsbredden til temperatur for beliggenheter uten tilgjengelig temperaturdata utfører vi kvantilregresjon med forklaringsvariable. Vi utvikler en ny metode for romlig interpolering av parametriske fordelinger, som anvendes for å utføre romlig modellering av en FPLF. Den nye metoden kombinerer kvantilregresjon med metoden for kvantiltilpasning. Asymptotiske betingelser for konsistente parameterestimatorer presenteres for metodene våre. I tillegg utføres flere simuleringsstudier for å numerisk evaluere de foreslåtte metodene. En FPLF tilpasses til 30 år med observasjoner av den døgnlige variasjonsbredden til temperatur. Observasjoner hentes fra 55 værstasjoner på Sør-, Øst- og Vestlandet, og modellering utføres uavhengig for vinter, vår, sommer og høst. Resultatene viser at FPLF-en er en svært lovende modell for variasjonsbredden til temperatur. Den lokale modelltilpasningen, som bruker metoden for kvantiltilpasning, viser gode resultater. Den nye interpoleringsmetoden viser også svært lovende resultater. Romlig modellering av variasjonsbredden til temperatur er suksessfull for vinter, vår og høst. Modellresultatene er ikke like gode for sommeren.

Preface

This thesis is a result of my work in the subject *TMA4900 – Industrial Mathematics, Master's Thesis* as a student at the Norwegian University for Science and Technology (NTNU). The project has been carried out in cooperation with the Norwegian Computing centre (NR), and it is a continuation of my work as a summer intern at NR the last two years. I would like to direct a huge thanks to Thordis L. Thorarinsdottir, from NR. She has provided me with a lot of help in finishing my thesis, through weekly meetings for almost a year. I would also like to thank Ingelin Steinsland, my supervisor from NTNU, for all the help she has given me this last year. Additionally, I am in debt to Marion Haugen from NR, for providing me with an extensive set of code that taught me how to work with such large amounts of data. Andreas Dobler at the Norwegian Meteorological Institute has also been helping me to get the necessary data, for which I am very grateful.

Silius M. Vandeskog
June 2019
Trondheim

Table of Contents

Notation

Table 1 displays notation used in this thesis.

Table 1: Variables and notation used in this thesis.

Symbols and abbreviations	Meaning
$\boldsymbol{y}$	An n-dimensional vector of observations.
$\boldsymbol{\beta}$	A k-dimensional regression coefficient vector.
$\mathbf{x}$	A k-dimensional vector of explanatory variables.
$\mathbf{X}$	Design matrix of dimension $n \times k$.
f, PDF	Probability density function.
F, CDF	Cumulative distribution function.
Q	Quantile function.
FPLD	Five-Parameter Lambda Distribution (3.6).
$\boldsymbol{\lambda}$	Parameters of the FPLD.
α, β, θ, etc.	Parameters.
$\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\theta}$, etc.	Estimators of parameters.
Any **bold** symbol	A vector or a matrix.
Any non-bold symbol	A scalar.

For any vector $\boldsymbol{v} \in \mathbb{R}^n$ the elements are indexed so that

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^T.$$

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ the elements are indexed so that

$$\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{pmatrix},$$

where $\boldsymbol{a}_i$ is a k-dimensional vector, $i = 1, \ldots, n$.

Chapter 1

Introduction

Our climate is constantly changing. In the last century, we have seen extraordinarily large and abrupt changes in global climate patterns. This may result in longer droughts, an increase in flooding and several other societal challenges (IPCC, 2014). Possible consequences of climate change are assessed in different impact studies, e.g. using hydrologic models for future flood assessment (Hanssen-Bauer et al., 2009). The hydrologic models are physically based numerical models, and deterministic given a certain input. They are driven by input variables describing local climatic conditions. These consist, among others, of daily precipitation, daily wind speed and intensity, daily minimum, maximum and mean temperature. For a future climate, these input variables are obtained from climate simulations, describing different possible scenarios of the future. In this way, we are able to describe possible consequences of a changing climate.

The climate is a global system. Consequently, in order to create models that can explain the climate to a satisfactory degree, one must start with a global model. Several global climate models, called global circulation models, have been established, which are based on physical models and numerical schemes. These are good at capturing large-scale climate features, but they typically have a grid cell resolution of 100–200 km due to computational limitations (Rummukainen, 2010). A scale of 100–200 km is too large when one aims to capture regional climate events, thus requiring dynamical or statistical downscaling of the global model output (Maraun and Widmann, 2018). Dynamical downscaling nests a regional climate model of higher resolution inside a global circulation model simulation, using simulations from a given global model as boundary conditions. Here, simulations are once again created using numerical solutions of physical laws. The grid cell resolution of a regional climate model can typically be found in the range of 10–50 km per grid cell. However, hydrologic climate models often require an input with resolution $\sim$1 km per grid cell. Additionally, it has been found that the regional climate model outputs contain several biases that should be corrected before the simulations can be applied in impact studies of climate change. Therefore, it is a common procedure to apply some kind of bias correction scheme to the output of a regional climate model (Vandeskog, Haugen, and Thorarinsdottir, 2018; Maraun and Widmann, 2018).

Statistical downscaling methods are data-driven methods that determine empirical links between large-scale and fine-scale climate model simulations. An advantage of statistical downscaling is computational efficiency, i.e. it is possible to downscale simulations onto a much finer grid than a dynamical approach is able to. Physically driven models are, however, able to describe interactions between climate variables much better than the purely data-driven statistical downscaling methods. Therefore, when one performs downscaling to high resolutions, it is common to first perform dynamical downscaling from a global circulation model simulation onto a regional scale model. Statistical bias correction can then be applied to the regional climate model output, followed by downscaling to the wanted resolution using statistical downscaling. The combination of statistical bias correction and downscaling is commonly referred to as post-processing (Maraun and Widmann, 2018).

Hydrologic models require minimum, maximum and mean daily temperature as input. These climate variables are clearly dependent. Consequently, multivariate post-processing of the variables should be performed, in order to correctly capture the dependencies. Multivariate post-processing is, however, difficult and highly computationally demanding, as the dependencies must be modelled in both time and space for large amounts of data. New results like those of Cannon (2018) have shown promise, but the common approach today is to perform univariate post-processing of each climate variable separately.

A major problem with the standard univariate approach is that separate post-processing can give inconsistencies in the constraint that daily minimum temperature must be lower than daily mean temperature, which again must be lower than daily maximum temperature. Inconsistent input in an impact study for climate change will subsequently result in inconsistent output for the given study. The Nordic Gridded Climate Data Set version 2 (Lussana, Saloranta, et al., 2018; Lussana, Tveito, and Uboldi, 2018) contains gridded historical climate data from 1950 to the present, for all of Finland, Sweden and Norway. The gridded data set is generated by applying a spatial interpolation approach to data from surface observation stations. Minimum, maximum and mean temperature were considered independently in the creation of this data set. During winter, 0.02% of all data contains daily maximum temperatures that are smaller than the daily minimum. Approximately 11% of all winter data contain daily mean temperatures outside the range of minimum and maximum temperature. This is a substantial problem that must be corrected before we can apply the data set in climate change assessment studies.

We have previously proposed that daily minimum, mean and maximum temperature can be modelled more consistently by transforming the variables into daily mean temperature, daily temperature skewness and diurnal temperature range (Vandeskog, Thorarinsdottir, and Steinsland, 2019). Di-

urnal temperature range is the difference between the daily minimum and maximum temperature. It is always bounded from below by zero. Temperature skewness is a number between 0 and 1, explaining the relative position of mean temperature between minimum and maximum temperature. We have shown that this transformation considerably reduces the correlation between the three climate variables. By enforcing the diurnal temperature range to stay positive, we can also ensure that minimum and maximum daily temperatures never cross each other. It is therefore of great interest to be able to model diurnal temperature range and temperature skewness, as these models might be used in improving post-processing schemes for temperature.

In this thesis, the Five-Parameter Lambda Distribution (FPLD) is presented as a model for diurnal temperature range. The distribution has previously been used to model diurnal temperature range in the EUSTACE research project (Lindgren, 2016). However, to our knowledge, the results have yet to be published. The FPLD was introduced by Gilchrist (2000), but has not obtained much attention in the statistical literature. We present some theoretical justification for the choice of model and explore some properties of the distribution. Methods for performing both local and spatial parameter estimation for the FPLD are developed. The local estimation method makes it possible to estimate parameters of the FPLD in areas where observations have been made available, while the spatial estimation method performs spatial interpolation for parameter estimation at locations without any temperature observations.

The FPLD is applied for modelling of diurnal temperature range in the southern parts of Norway. The chosen data consist of daily temperature measurements for the years 1989 to 2018, for 55 different weather stations in Norway (Norwegian Meteorological Institute, 2019). Diurnal temperature range is clearly a phenomenon that varies both in space and time. However, in this thesis, we focus on modelling the spatial characteristics of diurnal temperature range. Consequently, when applying our method on real data, the temperature range is considered stationary in time within each of the four seasons of the year.

For local parameter estimation of the FPLD, the method of quantiles is presented as an alternative to the more standard methods like maximum likelihood estimation and the method of moments. This method attempts to minimise the difference between the quantile function of a parametric distribution and a set of estimated quantile from observed data. It is thus quite similar to the method of moments. Both methods perform pairing of statistics from a parametric distribution and sample statistics from observed data. They merely focus on different types of statistics. The method of quantiles also goes by the name percentile matching.

We fit the FPLD to diurnal temperature range observations, using the method of quantiles. The results are promising. The method is able to

find a good fit of the FPLD for almost all the available weather stations, although small errors do occur for some of the data.

A spatial regression framework with explanatory variables is developed in order to model the spatially varying distribution of diurnal temperature range. Standard linear regression only attempts to estimate the mean value of its response variable. However, we find that the higher order moments of the distribution of diurnal temperature range vary heavily in space. Standard linear regression is not able to model such behaviour. Accordingly, a Bayesian quantile regression framework is implemented for modelling the temperature range. In a quantile regression, a set of specific quantiles of a distribution is modelled, instead of the mean of the distribution. This leads to higher flexibility and the ability to model more complex distributions. A spatial random effect is not incorporated into the regression model.

Having developed a quantile regression model for diurnal temperature range, we perform interpolation on the parameters of the FPLD. Using the quantile regression model, we are able to estimate a set of quantiles in the distribution of diurnal temperature range at a given location. The method of quantiles is then performed on the available quantiles, for estimating the parameters of the FPLD at the given location. This approach is performed on the available data, and is able to successfully model diurnal temperature range, albeit not always with high performance.

The main contributions of this thesis can be divided into three parts. We perform modelling of diurnal temperature range with explanatory variables. Modelling of diurnal temperature range in time has been performed previously (e.g. Makowski, Wild, and Ohmura, 2008), and the annual trends of diurnal temperature range have also been modelled spatially (Zhou et al., 2009). However, to our knowledge, except from analyses of the effects of a single explanatory variable on mean diurnal temperature range (e.g. Gallo, Easterling, and Peterson, 1996; Waqas and Athar, 2018), no attempts at spatial modelling of diurnal temperature range have been published. Additionally, apart from the EUSTACE project, we are not aware of any usage of the FPLD within the climate sciences. Consequently, this thesis introduces the distribution into the field of climate science. In order to model the distribution of temperature range, we also develop a new method for interpolation of parametric distributions, which combines quantile regression and the method of quantiles. To our knowledge, this method has never before been published.

The remainder of the thesis is organised as follows: In Chapter 2, data are presented as a motivation and for understanding the problem of modelling diurnal temperature range. Following this, the FPLD is presented as a model for diurnal temperature range in Chapter 3. Some additional properties of the distribution are also presented. Inference for the FPLD is presented in Chapter 4. First, methods for local parameter estimation of the FPLD, including the method of quantiles, are discussed. Then, the

method of quantile regression is presented, and a method for spatial interpolation of the parameters of the FPLD is developed by combining quantile regression and the method of quantiles. Finally, we examine some asymptotic properties for our chosen methods and conditions for consistency of the developed estimators are presented and discussed. In Chapter 5, we evaluate the method of quantiles and the quantiles regression numerically in several different simulation studies. At last, in Chapter 6, we perform modelling of diurnal temperature range in Norway, with the FPLD. The results of this thesis are discussed in Chapter 7.

Chapter 2

Data

In this thesis, a methodology for local and regional modelling of the FPLD is developed and applied for modelling diurnal temperature range in the southern parts of Norway. The motivation behind modelling of the FPLD originates in the need for better models of diurnal temperature range. The distribution of diurnal temperature range is highly skewed and can take many forms, making it difficult to model with standard parametric distributions. We therefore first present our data, to motivate for the usage of the FPLD.

2.1 Data description

All the data applied in this thesis are freely available on the internet (Norwegian Meteorological Institute, 2019). The data consist of daily time series of temperature observations from 55 different weather stations in the southern parts of Norway. At all weather stations, daily minimum and maximum temperatures from the time interval 18–18 UTC are observed. We calculate diurnal temperature range as the difference between daily maximum and minimum temperature. Daily mean temperature is also made available at all weather stations. The mean temperature is extracted from the time interval 06–06 UTC, meaning that it does not coincide perfectly with the diurnal temperature range.

We are interested in the last 30 years of data. Consequently, all our data are collected from the time period 01/01/1989 – 31/12/2018. The available time series from the Norwegian Meteorological Institute, suffer from several occurrences of missing data. The 55 weather stations are selected because they contain less than 50% missing data. Longitude, latitude and altitude of each station are also available from the internet. The locations of all weather stations are found in Figure 2.1. The weather stations are not only located on the mainland, but also at lighthouses into the sea. We have been provided a map containing the shortest distance to the sea from any given location, created by the Norwegian Meteorological Institute (Dyrrdal et al., 2015). This is applied in obtaining the distance from the sea for all weather stations, which is used as an explanatory variable for spatial modelling of diurnal temperature range.

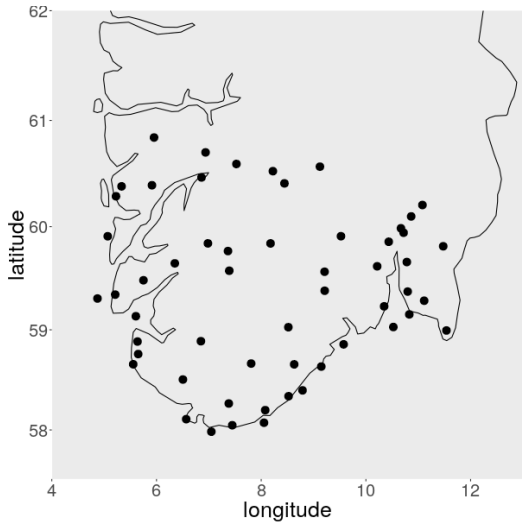

Figure 2.1: Locations of the 55 chosen weather stations in this thesis.

2.2 Data exploration

The distribution of diurnal temperature range is examined at all of the available weather stations. Some obvious errors are found in the available data, with selected days having higher minimum temperature than maximum temperature, i.e. negative diurnal temperature range. These observations are removed from the data. We assume that some positive temperature range values are faulty as well. However, these are difficult to detect without further knowledge of the data, so we are not able to remove or correct them. In total, the data set contains more than 450000 observations of diurnal temperature range. We must simply hope that the immense amount of data allows for some errors without significantly affecting our analysis.

It has been found that the distribution of diurnal temperature range in Norway is dependent upon the season of the year (Vandeskog, Thorarinsdottir, and Steinsland, 2019). In order to investigate whether this also holds for this data set, box-plots are created, containing 365 boxes, i.e. one for each day of the year. One such box-plot is created for each weather station and the results are examined. Examples of such plots are seen in Figure 2.2. It is clear from these plots that both the median and the range of diurnal temperature range differs greatly throughout the year. Consequently, all further data analysis is performed separately for each season of the year.

The observed densities of diurnal temperature range are examined. Fig-

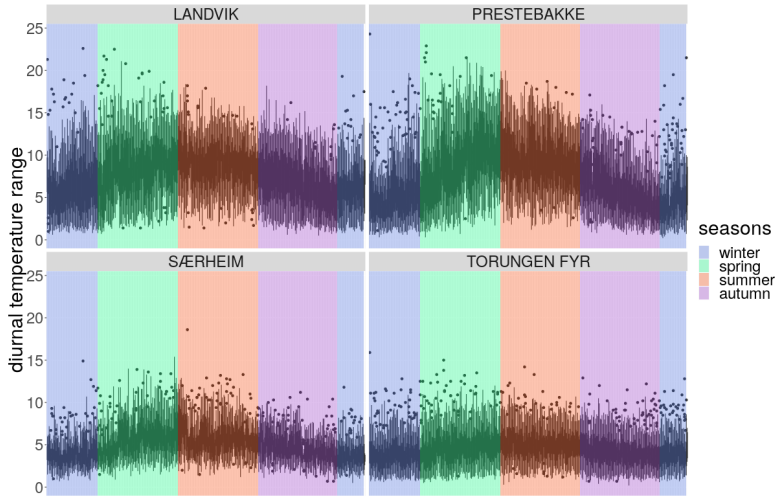

Figure 2.2: Representative box-plots displaying the daily median and variability of temperature range at given weather stations. Seasons are displayed using different background-colours.

ure 2.3 displays the observed densities of temperature range at all available weather stations for each season. It can be seen that most distributions are heavily skewed to the right. However, there are also large differences in the shape of the different distributions. The colour of each distribution represents the shortest distance to the sea. It seems that the distributions with modes to the left, i.e. mostly low values of diurnal temperature range, stem from weather stations close to the sea, and vice versa. Patterns like these can also be found if the distributions are coloured after their values of longitude, latitude or altitude. This is a strong indicator that the distribution of diurnal temperature range is somehow dependent upon its geographical location.

A more detailed display of different distributions of diurnal temperature range are found in Figure 2.4. Four histograms, containing diurnal temperature range data from different locations and seasons, are displayed. The shape and spread of these histograms vary heavily, and there does not seem to be any clear patterns or similarities between all four distributions. It is not obvious from these plots that any single parametric distribution is able to model all four distributions with a high level of success.

To investigate the spatial structure, we display key quantiles of diurnal temperature range at different weather station locations. For truncated and skewed distributions like those in Figure 2.3, there is not much information to be gained by examining the more standard statistics as sample mean and

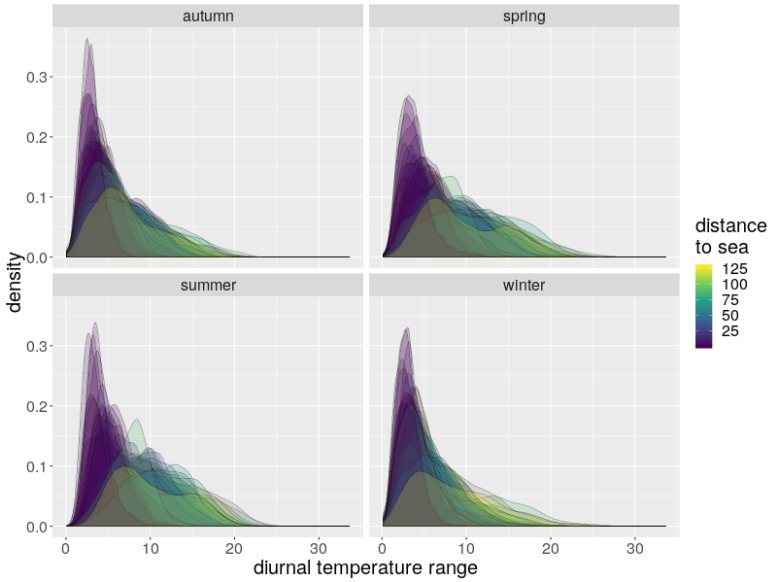

Figure 2.3: Observed densities of diurnal temperature range at all available weather stations. The shortest distance to the sea is represented in the colouring of each distribution.

variance. quantiles. Figure 2.5 displays the median of diurnal temperature range at all available weather stations for different seasons of the year. In all of the plots, one can see some kind of spatial pattern in the median. Weather stations close to the sea seem to obtain a lower value than that of stations further inland. One can also find some kind of drift in values as the longitude increases. The pattern seems to be quite similar for all four seasons. Interestingly, we find similar patterns as those in Figure 2.5 for all other tested quantiles. The interquartile range of diurnal range is also examined. The calculated values can be seen in Figure 2.6. Once again one can find similar patterns as those in Figure 2.5. This seems to indicate that geographical information should be included for modelling the quantiles of diurnal temperature range, spatially.

For spatial modelling of the diurnal temperature range, several explanatory variables are available. These consist of geographical information, in the form of longitude, latitude, altitude, and the distance to the sea for each station. In addition, information concerning the daily mean temperature is available from all weather stations. The historical mean and variance of daily mean temperature are added as explanatory variables for diurnal temperature range. We examine the dependencies between these variables and the quantiles of diurnal temperature range. The median of diurnal tem-

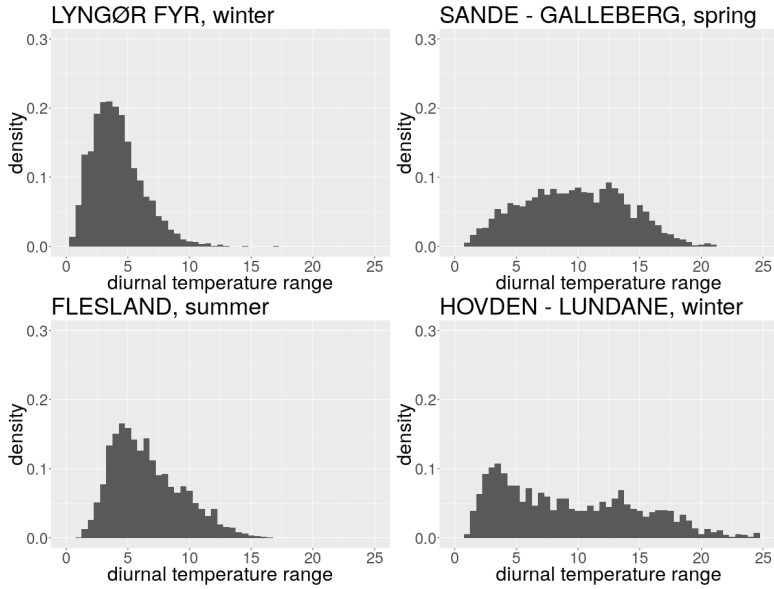

Figure 2.4: Histograms displaying observed diurnal temperature range. Data are collected from the period 01/01/1989 - 31/12/2018. Selected seasons and locations are indicated above each plot.

perature range is found for each weather station and each season. These values are plotted against the values of the different explanatory variables. The results are displayed in Figure 2.7. There seems to be some clear dependencies between the median of diurnal temperature range and all the possible explanatory variables. The dependencies are especially clear for the daily mean temperature observations. Other quantiles than the median of diurnal temperature range are plotted against the explanatory variables, and similar patterns are found for all of them.

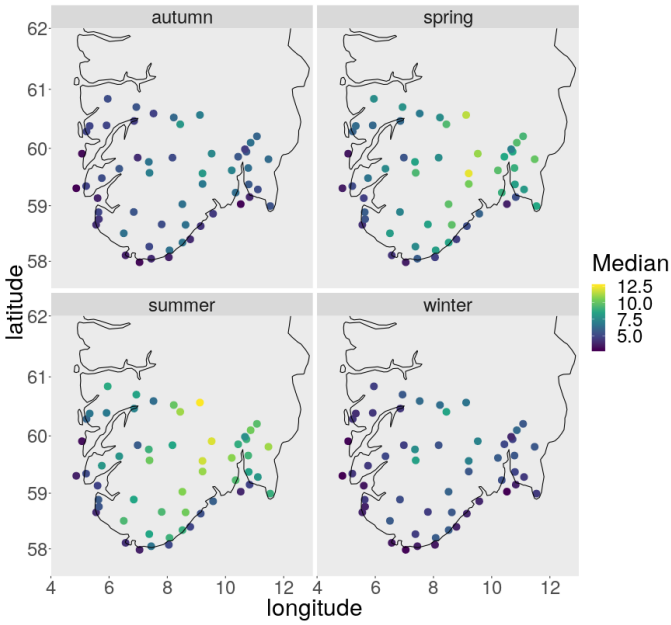

Figure 2.5: Empirical median of diurnal temperature range in Norway for all weather stations and seasons.

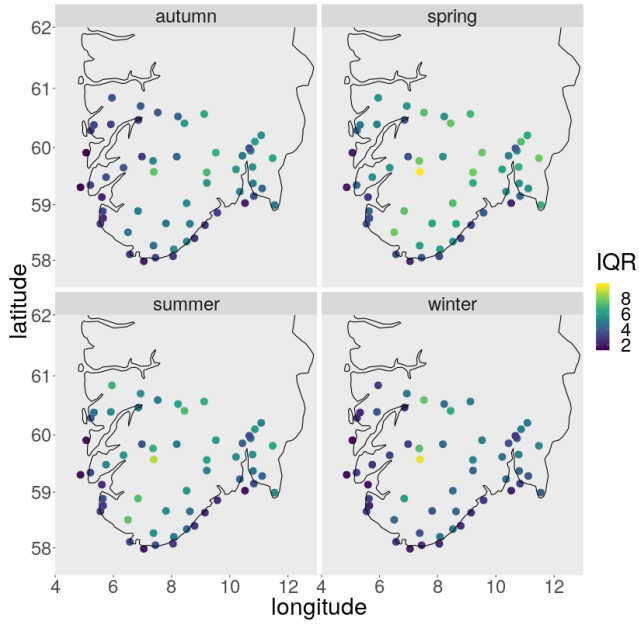

Figure 2.6: Interquartile range (IQR) of diurnal temperature range in Norway for all weather stations and seasons.

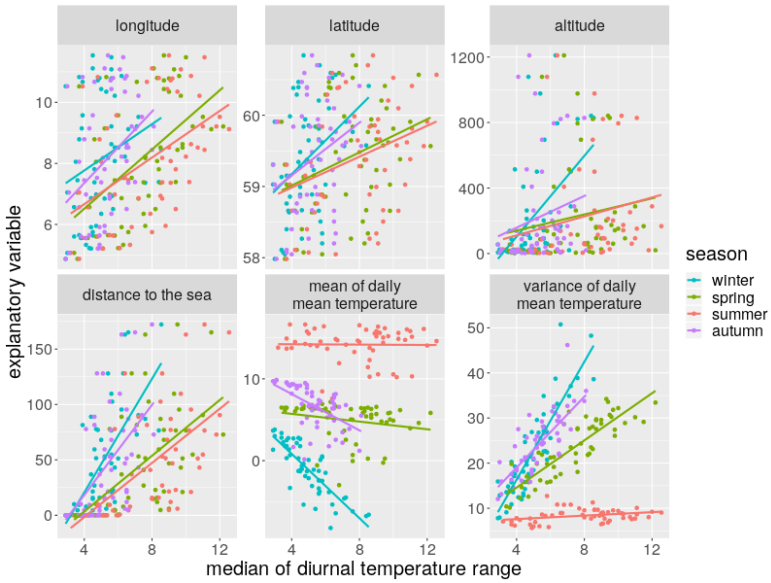

Figure 2.7: The median of diurnal temperature range is plotted against different explanatory variables at each of the 55 weather stations, for each season. Trends are estimated using Gaussian standard linear regression.

Chapter 3

Modelling diurnal temperature range

We wish to model the diurnal range of temperature using some parametric distribution. Parametric distributions for diurnal temperature range have not seen much interest within the fields of statistics and environmental sciences. However, one model for diurnal temperature range has been proposed by Lindgren (2016). Consider maximum and minimum daily temperature as two extreme quantiles in the distribution of temperature at a given day. Diurnal temperature range can then be modelled as the difference between these two quantiles. Some derivations and modelling techniques for the FPLD are developed by Vandeskog, Thorarinsdottir, and Steinsland (2019). The following chapter is inspired by that analysis.

3.1 Quantile functions

Quantile functions are frequently used in this thesis. Since these are used quite infrequently in modern statistics, a short introduction follows.

A quantile function $Q(p)$, associated with some CDF $F(y)$ specifies the value of the random variable y such that $F(y) = p$, i.e. $Q(p) = \inf\{y \in \mathbb{R} : p \leqslant F(y)\}$, $p \in [0, 1]$. If F is a strictly increasing function, then it is also both injective and surjective, and thus invertible. The quantile function Q is then equal to the inverse of F. A quantile function is always increasing in p, albeit not always strictly increasing. It is not necessary for the CDF to be analytically defined in order for the quantile function to exist, and vice versa. Consequently, the set of all possible quantile functions coincides with the set of all weakly increasing functions $Q(p)$, $p \in [0, 1]$. The set of all quantile functions is closed under addition, i.e. for any quantile functions Q_1 and Q_2, the sum $Q_3 = Q_1 + Q_2$ is a quantile function. This is a trivial result, as the sum of two increasing functions will be an increasing function. The set of all quantile functions with positive support is also closed under multiplication. These properties make the quantile function extremely flexible for statistical modelling. One can simply create new quantile functions by adding or multiplying known quantile functions with different preferable properties (Gilchrist, 2000).

3.2 Appearance of the FPLD

The generalised Pareto distribution (GPD) is commonly used to model the tails of other statistical distributions (e.g. Coles, 2001). The GPD has the cumulative distribution function

$$F(z,\xi) = \begin{cases} 1-(1+\xi z)^{-1/\xi}, & \xi \neq 0 \\ 1-e^{-z}, & \xi = 0 \end{cases} \tag{3.1}$$

and quantile function

$$Q(p;\xi) = \begin{cases} \left(1-(1-p)^{\xi}\right)/\xi, & \xi \neq 0 \\ -\log(1-p), & \xi = 0, \end{cases} \tag{3.2}$$

(e.g. Hosking and Wallis, 1987). More generally, one can add location and scale parameters

$$Q(p;\mu,\eta,\xi) = \mu + \eta\, Q(p;0,1,\xi),\ \ \eta > 0, \mu \in \mathbb{R}. \tag{3.3}$$

We model maximum daily temperature as some quantile of a GPD, while minimum daily temperature is modelled as some quantile of a reflected GPD. The original GPD has a range of $[0,\infty)$, giving a reflected GPD a range of $(-\infty,0]$. Gilchrist (2000) has shown that the distribution $-Q(1-p)$ is the reflection of the distribution $Q(p)$ across the line $y=0$. The quantile function of a reflected GPD is

$$Q^*(p;\mu,\eta,\xi) = \mu + \eta \begin{cases} \left(p^{\xi}-1\right)/\xi & \xi \neq 0 \\ \log p, & \xi = 0. \end{cases} \tag{3.4}$$

Diurnal temperature range is the difference between daily maximum and minimum temperature, i.e. the difference between two quantiles of a GPD. Since diurnal temperature range is the difference between two quantile functions, it is also a quantile of some quantile function, given that the resulting difference is an increasing function of p. We allow different behaviour in

each tail and calculate

$$\begin{aligned} y_{\text{range}} =& y_{\max} - y_{\min} \\ =& Q(p;\mu_1,\eta_1,\xi_1) - Q^*(p;\mu_2,\eta_2,\xi_2) \\ =& \mu_1 + \eta_1 \frac{1-(1-p)^{\xi_1}}{\xi_1} - \mu_2 - \eta_2 \frac{p^{\xi_2}-1}{\xi_2} \\ =& (\mu_1-\mu_2) + \frac{\eta_1-\eta_2}{2}\left\{\left(1-\frac{\eta_1+\eta_2}{\eta_1-\eta_2}\right)\frac{p^{\xi_2}-1}{\xi_2} - \right. \\ & \left.\left(1+\frac{\eta_1+\eta_2}{\eta_1-\eta_2}\right)\frac{(1-p)^{\xi_1}-1}{\xi_1}\right\}. \end{aligned} \tag{3.5}$$

This is the quantile function of a Five-Parameter Lambda Distribution (FPLD) (Gilchrist, 2000),

$$Q(p;\boldsymbol{\lambda}) = \lambda_1 + \frac{\lambda_2}{2}\left\{(1-\lambda_3)\frac{p^{\lambda_4}-1}{\lambda_4} - (1+\lambda_3)\frac{(1-p)^{\lambda_5}-1}{\lambda_5}\right\}. \tag{3.6}$$

The FPLD is not popular within the statistical literature. However, some areas of usage has been found for the distribution. Inference for the FPLD is presented by e.g. Tarsitano (2010) and Nair, Sankaran, and Balakrishnan (2013). Applications of the distribution have been found by Ahmadabadi, Farjami, and Moghadam (2012) and Noorian and Ahmadabadi (2018). They have applied the FPLD in statistical process control methods. Additionally, Movahedi et al. (2017) apply the FPLD in estimating industry component tolerances. To our knowledge, no publications have been made within the climate sciences that use the FPLD for statistical modelling.

The FPLD is an extension of the Generalised Lambda distribution (GLD) (Ramberg and Schmeiser, 1974)

$$Q(p;\boldsymbol{\lambda}) = \lambda_1 + \frac{p^{\lambda_3}-(1-p)^{\lambda_4}}{\lambda_2}, \tag{3.7}$$

which again is an extension of Tukeys Lambda distribution (Tukey, 1962)

$$Q(p;\lambda) = \begin{cases} \dfrac{p^{\lambda}-(1-p)^{\lambda}}{\lambda} & , \lambda \neq 0 \\ \dfrac{\log p}{1-p} & , \lambda = 0 \end{cases}. \tag{3.8}$$

While the FPLD is not commonly applied, the GLD has been used in several different studies, as it is deemed a highly flexible distribution for modelling data of many shapes. The GLD is applied within the field of meteorology,

where it is proposed as a model for solar radiation data (Öztürk and R. Dale, 1982). The GLD is also applied within finance (Marcondes, Peixoto, and Maia, 2018; Tarsitano, 2004), engineering (Upadhyay and Ezekoye, 2008), psychology (Delaney and Vargha, 2000), health and nutrition (Ejima et al., 2018) and many other fields.

3.3 Shape of the FPLD

The quantile function, and therefore also the CDF of the FPLD, is strictly increasing in p as long as $\lambda_2 > 0$ and $\lambda_3 \in (-1, 1)$. This can easily be seen as both p^{λ_4} and $-(1-p)^{\lambda_5}$ are strictly increasing functions of p for $p \in [0, 1]$. Thus, under some constraints on λ_2 and λ_3, the quantile function (3.6) of the FPLD is a valid quantile function. An examination of the Taylor expansion of $Q(p; \boldsymbol{\lambda})$ around $\lambda_4, \lambda_5 = 0$ show that the quantile function is continuous for all p and bounded away from infinity for $p \notin \{0, 1\}$. We get

$$\lim_{\lambda \to 0} \frac{p^\lambda - 1}{\lambda} = \ln p. \tag{3.9}$$

The quantile function of the FPLD (3.6) can be expressed as

$$Q(p; \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = \lambda_1 + \lambda_2 Q(p; 0, 1, \lambda_3, \lambda_4, \lambda_5). \tag{3.10}$$

From (3.10) it is clear that λ_1 and λ_2 act as location and scaling parameters for the distribution. It is harder to get an intuitive grasp of the influence of the remaining parameters at a first look. However, we find that the value of λ_4 is more important when p is close to zero, as $(p^{\lambda_4} - 1)$ is approximately equal to zero for all large values of p. The same can be said for λ_5 when p is close to one. This means that λ_4 mainly controls the left tail of the distribution and λ_5 mainly controls the right tail, while λ_3 acts as a weight between the two tails.

Plots displaying the effect of λ_3 can be seen in Figure 3.1. The upper right plot displays an FPLD with $\lambda_3 = 0$. As λ_3 increases towards 1, the left tail decreases while the right tail increases. A similar pattern can be seen when λ_3 decreases. Figure A.3 displays this behaviour in more details. The influence of λ_4 and λ_5 on the shape of an FPLD is also displayed in Figure A.4.

The support of an FPLD, given by $[Q(0, \boldsymbol{\lambda}), Q(1, \boldsymbol{\lambda})]$, can be both finite and infinite. If λ_4 is negative when p approaches 0, the quantile function will approach negative infinity. If, however, λ_4 is positive, the left tail is finite. Once again, similar properties hold for λ_5 when p approaches 1. The

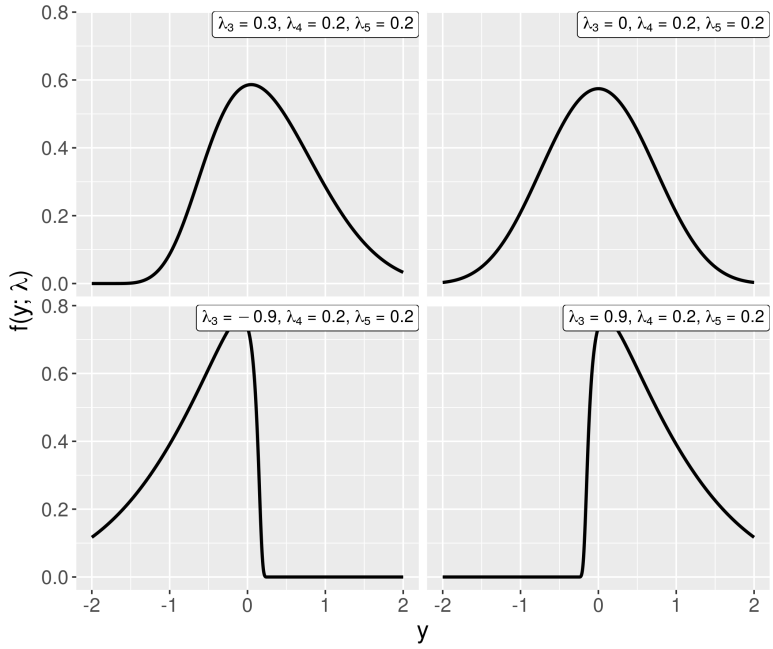

Figure 3.1: Probability density functions of the FPLD with different values of $\boldsymbol{\lambda}$. $\lambda_1 = 0,\ \lambda_2 = 1$ in all plots.

support of an FPLD is

$$[Q(0,\boldsymbol{\lambda}), Q(1,\boldsymbol{\lambda})] = \begin{cases} [\lambda_1 - \frac{\lambda_2(1-\lambda_3)}{2\lambda_4}, \lambda_1 + \frac{\lambda_2(1+\lambda_3)}{2\lambda_5}], & \lambda_4, \lambda_5 > 0 \\ [\lambda_1 - \frac{\lambda_2(1-\lambda_3)}{2\lambda_4}, \infty), & \lambda_4 > 0, \lambda_5 \leqslant 0 \\ (-\infty, \lambda_1 + \frac{\lambda_2(1+\lambda_3)}{2\lambda_5}], & \lambda_4 \leqslant 0, \lambda_5 > 0 \end{cases}. \tag{3.11}$$

The distribution is therefore capable of modelling data with several types of support. Diurnal temperature range is always bounded below by 0, making distributions with negative support a poor choice for modelling. This can be avoided using an FPLD with the inequality constraints

$$\lambda_1 - \frac{\lambda_2(1-\lambda_3)}{2\lambda_4} > 0,\ \lambda_4 > 0. \tag{3.12}$$

Examples of possible shapes of the FPLD, both with finite and infinite support, can be seen in Figures A.1 and A.2.

3.4 Probability density function of the FPLD

The dependence between the probability density function f and quantile function Q of a random variable with a strictly increasing CDF F is given by

$$f(y)q(p) = 1, \text{ with } p = F(y), \tag{3.13}$$

where $q(p) = \frac{\mathrm{d}Q(p)}{\mathrm{d}p}$ is called the quantile density function (Gilchrist, 2000). We can prove this using the fact that

$$Q(F(y)) = \inf\{x \in \mathbb{R} : F(y) \leqslant F(x)\} = y, \forall\ y \in (Q(0), Q(1)), \tag{3.14}$$

if and only if F is a strictly increasing function. This leads to the equation

$$\frac{\mathrm{d}F(y)}{\mathrm{d}y}\frac{\mathrm{d}Q(p)}{\mathrm{d}p} = \frac{\mathrm{d}F(y)}{\mathrm{d}y}\frac{\mathrm{d}Q(F(y))}{\mathrm{d}F(y)} = \frac{\mathrm{d}F(y)}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}F(y)} = 1. \tag{3.15}$$

As discussed in Section 3.3, the CDF of the FPLD is strictly increasing in p as long as $\lambda_2 > 0$ and $\lambda_3 \in (-1, 1)$. Consequently, the result in (3.13) holds for the FPLD, for all legal values of $\boldsymbol{\lambda}$. The quantile density function of the FPLD equals

$$q(p; \boldsymbol{\lambda}) = \frac{\lambda_2}{2}\left\{(1-\lambda_3)p^{\lambda_4 - 1} + (1+\lambda_3)(1-p)^{\lambda_5 - 1}\right\}, \tag{3.16}$$

and the corresponding probability density function equals

$$f(y; \boldsymbol{\lambda}) = \frac{2}{\lambda_2}\left\{(1-\lambda_3)p^{\lambda_4 - 1} + (1+\lambda_3)(1-p)^{\lambda_5 - 1}\right\}^{-1}, \tag{3.17}$$

with $p = F(y; \boldsymbol{\lambda})$.

The quantile density function $q(p; \boldsymbol{\lambda})$ is strictly positive, as both terms in (3.16) are strictly positive for $\lambda_2 > 0, \lambda_3 \in (-1, 1)$. This ensures that the probability density function is positive and bound away from infinity.

3.5 Moments of the FPLD

The rth moment of a distribution with density function $f(y)$ is equal to

$$\mathbb{E}[y^r] = \int_{-\infty}^{\infty} y^r f(y)\mathrm{d}y. \tag{3.18}$$

For strictly increasing CDFs, one can substitute y with $p = F(y)$ to get the equation

$$\mathbb{E}[y^r] = \int_0^1 Q(p)^r \mathrm{d}p, \tag{3.19}$$

where $y = Q(p)$. Using 3.6 we reparameterise the quantile function of an FPLD to the compact form

$$Q(p; \boldsymbol{\lambda}) = a + bp^{\lambda_4} - c(1-p)^{\lambda_5}, \tag{3.20}$$

with

$$b = \tfrac{\lambda_2(1-\lambda_3)}{2\lambda_4}, \quad c = \tfrac{\lambda_2(1+\lambda_3)}{2\lambda_5}, \quad a = \lambda_1 - b + c. \tag{3.21}$$

Consequently, the rth power $y^r = Q(p; \boldsymbol{\lambda})^r$ equals

$$\begin{aligned} y^r &= \left(a + bp^{\lambda_4} - c(1-p)^{\lambda_5}\right)^r \\ &= \sum_{j=0}^{r} \binom{r}{j} a^{r-j} \cdot \left(bp^{\lambda_4} - c(1-p)^{\lambda_5}\right)^j \\ &= \sum_{j=0}^{r}\sum_{k=0}^{j} (-1)^k \binom{r}{j}a^{r-j}\binom{j}{k} b^{j-k}c^k p^{(j-k)\lambda_4}(1-p)^{k\lambda_5}, \end{aligned} \tag{3.22}$$

and the rth moment of the FPLD is equal to

$$\begin{aligned} \mathbb{E}[y^r] &= \sum_{j=0}^{r}\sum_{k=0}^{j} (-1)^k \binom{r}{j}\binom{j}{k} a^{r-j}b^{j-k}c^k \int_0^1 p^{(j-k)\lambda_4}(1-p)^{k\lambda_5}\mathrm{d}p \\ &= \sum_{j=0}^{r}\sum_{k=0}^{j} (-1)^k \binom{r}{j}\binom{j}{k} a^{r-j}b^{j-k}c^k \cdot B\left[1+(j-k)\lambda_4, 1+k\lambda_5\right], \end{aligned} \tag{3.23}$$

where $B[x,y] = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, $x, y > 0$ denotes the beta function. The rth moment of an FPLD therefore exists for all parameter values $\boldsymbol{\lambda}$ with $\lambda_4, \lambda_5 > -r^{-1}$, $\lambda_2 > 0$ and $\lambda_3 \in (0,1)$. From (3.23) and (3.20) it follows that the mean and variance of the FPLD equal

$$\mathbb{E}[y] = \lambda_1 + \frac{\lambda_2}{2}\left\{-\frac{1-\lambda_3}{1+\lambda_4} + \frac{1+\lambda_3}{1+\lambda_5}\right\}, \tag{3.24}$$

and

$$\begin{aligned} \mathrm{Var}[y] = &\frac{\lambda_2^2}{4}\left\{\frac{(1-\lambda_3)^2}{\lambda_4^2}\left(\frac{1}{1+2\lambda_4} - \frac{1}{(1+\lambda_4)^2}\right) + \right. \\ &\frac{(1+\lambda_3)^2}{\lambda_5^2}\left(\frac{1}{1+2\lambda_5} - \frac{1}{(1+\lambda_5)^2}\right) + \\ &\left. 2\frac{1-\lambda_3^2}{\lambda_4\lambda_5}\left(\frac{1}{(1+\lambda_4)(1+\lambda_5)} - B[1+\lambda_4, 1+\lambda_5]\right)\right\}. \end{aligned} \tag{3.25}$$

3.6 Parameter transformations

Many possible parameterisations of the quantile function of the FPLD are available. The advantage of the parameterisation in (3.6) is that it allows for an intuitive understanding of the FPLD and its shape. However, the parameter representation in (3.6) is not great for numerical applications. In order to ease estimation procedures for the parameters of an FPLD, we perform a transformation of variables. First, λ_1 is replaced by the median of an FPLD, $\lambda_1^* = Q\left(\frac{1}{2}; \boldsymbol{\lambda}\right)$. This is inspired by the reparameterisation of a general Pareto distribution, performed by Li, Reitan, and Stenius (2017). The median is a highly robust statistic. Consequently, the stand-alone estimation of λ_1^* can be performed easily. This also facilitates for easy evaluation of any estimator for λ_1^*, as the sample median can be observed and is close to the true median of the distribution. Furthermore, we replace $\lambda_2, \ldots, \lambda_5$ with the parameters $\lambda_2^*, \ldots, \lambda_5^*$. These new parameters can take any values on the real line, whereas λ_2 and λ_3 are bounded to a finite interval. This removes the need for some inequality constraints on the parameters for all optimisation procedures. We get the reparameterisation scheme

$$\begin{aligned}
\lambda_1 &= \lambda_1^* - \frac{\lambda_2}{2}\left\{(1-\lambda_3)\frac{(\frac{1}{2})^{\lambda_4}-1}{\lambda_4} - (1+\lambda_3)\frac{(\frac{1}{2})^{\lambda_5}-1}{\lambda_5}\right\},\\
\lambda_2 &= \log(1+e^{\lambda_2^*}),\\
\lambda_3 &= \frac{1-e^{\lambda_3^*}}{1+e^{\lambda_3^*}},\\
\lambda_4 &= \log(1+e^{\lambda_4^*}),\\
\lambda_5 &= \log(1+e^{\lambda_5^*}) - \frac{1}{2}.
\end{aligned} \tag{3.26}$$

The quantile function of an FPLD can now be rewritten as

$$Q(p;\boldsymbol{\lambda}^*) = \lambda_1^* + \frac{g(\lambda_2^*)}{1+e^{\lambda_3^*}}\left\{e^{\lambda_3^*}\frac{p^{g(\lambda_4^*)} - (\frac{1}{2})^{g(\lambda_4^*)}}{g(\lambda_4^*)} - \frac{(1-p)^{h(\lambda_5^*)} - (\frac{1}{2})^{h(\lambda_5^*)}}{h(\lambda_5^*)}\right\}, \tag{3.27}$$

where $g(\lambda) = \log(1+\exp(\lambda))$ and $h(\cdot) = g(\cdot) - 1/2$. The vector of transformed parameters $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_5^*)^T$ is annotated with a star. This reparameterisation of the FPLD limits the parameter space of λ_4 and λ_5 from the real line to the intervals $(0, \infty)$ and $(-1/2, \infty)$, respectively. For an all-purpose parameterisation of the FPLD, this is obviously a poor choice. However, the motivation behind this parameterisation is to model diurnal temperature range. Negative values of λ_4 lead to a negative support of the FPLD, which is not allowed. Additionally, as is discussed in Section 4.3, our chosen estimator for the parameters of the FPLD is only consistent

when $\lambda_4, \lambda_5 > -1/2$.

In this parameterisation, both λ_2, λ_4 and λ_5 are transformed using the same function, g. Using Taylor expansion, we can show that

$$g(x) = \log(1 + e^x) \approx \begin{cases} e^x, & x \gg 1 \\ x, & x \ll 1 \end{cases} . \tag{3.28}$$

The exponential behaviour of $g(x)$ for negative values of x makes it so that large shifts in $\boldsymbol{\lambda}^*$ results in small changes in $\boldsymbol{\lambda}$. It is therefore harder for the parameters $\boldsymbol{\lambda}$ to approach very small values. This is a desirable property when modelling diurnal temperature range in Norway. As λ_4 and λ_5 grow small and/or negative, a tiny change in the parameters can lead to much heavier tails. Diurnal temperature range in Norway will seldom take values much larger than ~ 30K. Consequently, too heavy tails in the FPLD are not desirable. The purpose of the parameterisation on λ_2 is simply to ensure that the parameter is strictly positive. However, the linear growth of λ_2 for large values of λ_2^* is advantageous, as we do not wish for the distributional range of the FPLD to become too large. If the reparameterisation of λ_2 to λ_2^* had an exponential growth for all values of λ_2^* this would severely complicate any numerical procedures.

In all numerical implementations, the parameters $\boldsymbol{\lambda}^*$ are used. However, in the remainder of the thesis, we will mainly reference the parameters $\boldsymbol{\lambda}$ from (3.6), as their interpretation is more intuitive.

Chapter 4

Inference for the FPLD

In Chapter 3, the FPLD is proposed as a model for diurnal temperature range. However, in order to perform modelling with this distribution, we must first establish some parameter estimation methods. We present different methods for local parameter estimation, and a method for spatial interpolation of the FPLD. Lastly, we provide the necessary conditions for consistency of the presented estimators.

4.1 Local parameter estimation

We present three alternatives for local parameter estimation of the FPLD. These are maximum likelihood estimation, the method of moments and the method of quantiles.

4.1.1 Maximum likelihood estimation

Even though there is no analytical expression for the probability density function of an FPLD, we can estimate it numerically, using the expression in (3.17). The log-likelihood of $\boldsymbol{\lambda}$ for observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ equals

$$\begin{aligned} l(\boldsymbol{\lambda}; \boldsymbol{y}) &= \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\lambda}) \\ &= -n \log \frac{\lambda_2}{2} - \sum_{i=1}^{n} \log \left\{ (1-\lambda_3) p_{(i)}^{\lambda_4 - 1} + (1+\lambda_3)(1-p_{(i)})^{\lambda_5 - 1} \right\}, \end{aligned} \tag{4.1}$$

where $p_{(i)} = p(y_i, \boldsymbol{\lambda})$ is found by solving the equation

$$y_i = Q(p; \boldsymbol{\lambda}), \tag{4.2}$$

for p.

The maximum likelihood estimator for $\boldsymbol{\lambda}$ is found by maximising (4.1). Note that, in the log-likelihood all $p_{(i)}$, $i = 1, \ldots, n$ depend on $\boldsymbol{\lambda}$ as well as y_i. The problem of maximising the likelihood is therefore much more complex than it seems at a first glance. Additionally, since the support is constantly updated along with the parameter values (see (3.11)),

the log-likelihood function takes the value of negative infinity each time $\min\{y_1, \ldots, y_n\}$ or $\max\{y_1, \ldots, y_n\}$ is sent outside of the function support. The difference between a terrible fit and the best possible fit of the data might therefore be quite minuscule.

In practice, the maximum likelihood estimation becomes highly computationally costly. For a set of n observations, the equation $p = Q^{-1}(y; \boldsymbol{\lambda})$ must be solved numerically n times per iteration of the chosen optimisation scheme. This significantly slows down the algorithm for a large number of observations. Due to the non-linearity of the log-likelihood, a large number of iterations are also required in order to locate the maximum likelihood estimator with sufficient precision.

4.1.2 The method of moments

Another popular approach for parameter estimation is the method of moments (e.g. Casella and Berger, 2002). This method attempts to match empirical moments from available data with the theoretical moments of a given distribution. As seen in Section 3.5, the rth moment of an FPLD is an analytical expression of $\boldsymbol{\lambda}$. In order to estimate $\boldsymbol{\lambda}$ using the method of moments, one needs five or more empirical moments of the data. Denote the rth moment of an FPLD as m_r. The method of moments estimator for $\boldsymbol{\lambda}$ is found by

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \sum_{r=1}^{R} \left(\frac{1}{n} \sum_{i=1}^{n} y_i^r - m_r \right)^2, \tag{4.3}$$

with R the number of moments and n the number of data. The common approach for the method of moments is to set R equal to the number of parameters one is estimating. Often, the parameter estimators can then be solved analytically for m_1 to m_R. However, it is not possible to find an analytical solution for the FPLD. Additionally, we find that the first five moments are not enough to obtain estimates with high performance for diurnal temperature range, meaning that we need to compute the first 10-20 moments of the distribution. This is problematic, as the moments of diurnal temperature range grow exponentially with r.

4.1.3 The method of quantiles

We can apply least squares estimation to estimate the parameters of the FPLD. For a given independent random sample $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(n)}$, the ith order statistic can be expressed as

$$y_{(i)} = \mathbb{E}[y_{(i)}] + \varepsilon_i, \;\; i = 1, \ldots, n. \tag{4.4}$$

The error terms have an expected value of zero. However, they are heteroscedastic, meaning that the variance can differ for each ε_i, $i = 1, \ldots, n$.

The error terms are not independent and do not come from a symmetrical distribution. Tarsitano (2010) perform least squares estimation for the FPLD, based on the work of Öztürk and R. F. Dale (1985). He ignores the problems of heteroscedasticity and dependency of the error terms, as his goal is only an approximate estimation of $\boldsymbol{\lambda}$. We continue this tradition in good conscience. The expected value of the ith order statistic is available in closed form and can be analytically expressed for an FPLD as (Tarsitano, 2010)

$$\begin{aligned}\mathbb{E}[y_{(i)}] =& \lambda_1 + \frac{\lambda_2}{2}\left\{\frac{1-\lambda_3}{\lambda_4}\left[\frac{\Gamma(n+1)\Gamma(i+\lambda_4)}{\Gamma(i)\Gamma(n+1+\lambda_4)} - 1\right] + \right. \\ & \left. \frac{1+\lambda_3}{\lambda_5}\left[1 - \frac{\Gamma(n+1)\Gamma(n+1-i+\lambda_5)}{\Gamma(n+1-i)\Gamma(n+1+\lambda_5)}\right]\right\}.\end{aligned} \tag{4.5}$$

One can estimate $\widehat{\boldsymbol{\lambda}}$ by minimising the least squares expression

$$S(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \left(y_{(i)} - \mathbb{E}[y_{(i)}; \boldsymbol{\lambda}]\right)^2, \tag{4.6}$$

i.e.

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} S(\boldsymbol{\lambda}). \tag{4.7}$$

The minimisation (4.7) is non-linear and difficult to perform. However, as pointed out by Tarsitano (2005), for large samples the expected value of an order statistic can be simplified.

$$\mathbb{E}[y_{(i)}] \to Q(p_i) \text{ as } n \to \infty, \tag{4.8}$$

where $p_i = i/n$. The quantile function of an FPLD is much easier to minimise than the expected value of an order statistic. When one is modelling diurnal temperature range with an FPLD, several years of data are available at each weather station, meaning that we have thousands of measurements. We therefore assume that the asymptotic result in (4.8) always holds for our data. The minimisation problem (4.7) can now be simplified. The new minimisation problem becomes

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \left(y_{(i)} - Q\left(\frac{i}{n}; \boldsymbol{\lambda}\right)\right)^2. \tag{4.9}$$

A weakness of the method of quantiles is that it mainly focuses on the bulk of the observations $\boldsymbol{y}$. The method might return a parameter estimator $\widehat{\boldsymbol{\lambda}}$ such that $Q(0; \widehat{\boldsymbol{\lambda}}) \not< y_{(1)}$ and $y_{(n)} \not< Q(1; \widehat{\boldsymbol{\lambda}})$, because the fit of the model is good for the bulk of the data. This problem can be addressed by introducing inequality constraints to (4.9), demanding that $Q(0; \boldsymbol{\lambda}) < y_{(1)}$ and $y_{(n)} < Q(1; \boldsymbol{\lambda})$.

For minimisation of the quantile distance (4.9), we implement an optimisation algorithm that is a combination of the global minimisation algorithm `CRS` (Kaelo and Ali, 2006) and the local minimisation algorithm `COBYLA` (Powell, 1994), implemented in the `R`-package `nloptr` (Johnson, 2011).

The method of parameter estimation where one attempts to minimise the distance between a set of quantiles from the data and the corresponding quantiles from a statistical model, hereby referred to as the method of quantiles, is not that common in applied statistics. The first mentions of the method we have found are by Aitchison and Brown (1957). They apply the method of quantiles for parameter estimation of the three-parameter log-normal distribution. The method of quantiles is also known as percentile matching. Some applications of the method are found within the financial literature (e.g. Bignozzi, Macci, and Petrella, 2018; Sgouropoulos, Yao, and Yastremiz, 2015). Extensive theory for the method is presented by Koenker (2005).

Bignozzi, Macci, and Petrella (2018) state that parameter estimation methods based on quantile matching can be preferable when distributions are heavy-tailed or their support varies with the parameters. The support of an FPLD varies with its parameters and becomes heavy-tailed when λ_4 or λ_5 decreases. Additionally, Tarsitano (2005) performs the method of quantiles for estimating the parameters of an FPLD, using only five quantiles. He concludes that the method has several advantages, however, there is no available theoretical justification for the choice of quantiles. Our method applies n empirical quantiles, where n is the length of $\boldsymbol{y}$. Consequently, the problem of choosing the five best quantiles does not occur for us. Additionally, Bhatti et al. (2018) perform the method of quantiles on a Pareto distribution. They conclude that the method outperforms both the method of moments and the method of maximum likelihood estimation for the given distribution. Given that the FPLD is a combination of two Pareto distributions, this is a promising result.

It is clear from these findings that the method of quantiles might be a preferable choice when estimating the parameters of the FPLD. Additionally, the method of quantiles can easily be combined with the method of quantile regression, which is described in Section 4.2. For the remainder of this thesis, we choose to focus on the method of quantiles for performing local parameter estimation of the FPLD.

4.2 Spatial parameter estimation

The methods presented in Section 4.1 can only be applied at locations where observations of data are already available. In order to model data at locations without any observations, some spatial technique must be applied. We wish to express the parameters of the FPLD as functions of some

explanatory variables. A standard approach for this kind of modelling is to model each parameter as functions of a linear combination of some explanatory variables, $\lambda_i^j = g(\mathbf{x}_i^T \boldsymbol{\beta}^j), i = 1, \ldots, n,\ j = 1, \ldots, 5$. Under such a model, all of the local estimation methods can be applied for estimating the regression coefficients $\boldsymbol{\beta}^j$. However, the resulting loss function is quite difficult to minimise. For k available explanatory variables, the parameter vector takes a length of $5k$. The problem becomes highly computationally costly, and optimisation procedures often gets stuck in local minima. We have not been able to get any acceptable results using this method of parameter estimation. The method of quantiles seems to always end up in local minima with extremely heavy tails of the FPLD, and the method of maximum likelihood estimation is too computationally demanding.

We choose to perform spatial modelling of diurnal temperature range in a regression framework. One possible approach consists of performing regression separately on each of the five parameters of the FPLD. For a set of locations, one can estimate $\boldsymbol{\lambda}$ locally at each location. It is then possible to perform some regression technique, independently for each of the five parameters. However, due to the complex interactions between all of the parameters, we have little confidence in such an approach. Figure 4.1 displays how the square distance, applied in the method of quantiles (4.9), behaves when parameter values are slightly changed. In the left plots, a negative change in both λ_1 and λ_3 of approximately 0.1 cause a change in square distance from ~ 0 to approximately e^8. However, if the change in λ_3 is positive instead of negative, the square distance changes with approximately e^4. Not only is the change in the quantile distance extremely reactive to small changes in $\boldsymbol{\lambda}$, it also depends heavily on the direction of any changes. After having transformed the parameters from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^*$, the changes are not quite as bad. However, small changes in parameter still lead to large changes in our loss function. Consequently, estimating each of the parameters separately might lead to a poor model fit, as five small and independent parameter errors added together might result in large errors when combined.

Consequently, we perform regression on the distribution of diurnal temperature range itself. Standard linear regression models assume that all responses are distributed with identical variance σ^2, and attempts to model the mean value of the response. As seen in the data exploration in Section 2.2, the distribution of diurnal temperature range can take many different shapes, and is highly skewed. The standard linear regression framework cannot model such behaviour and is not an adequate solution. Another option is the method of quantile regression. This method does not place as many assumptions on the distribution of our observations as that of standard linear regression and might, therefore, achieve a better model fit to diurnal temperature range observations.

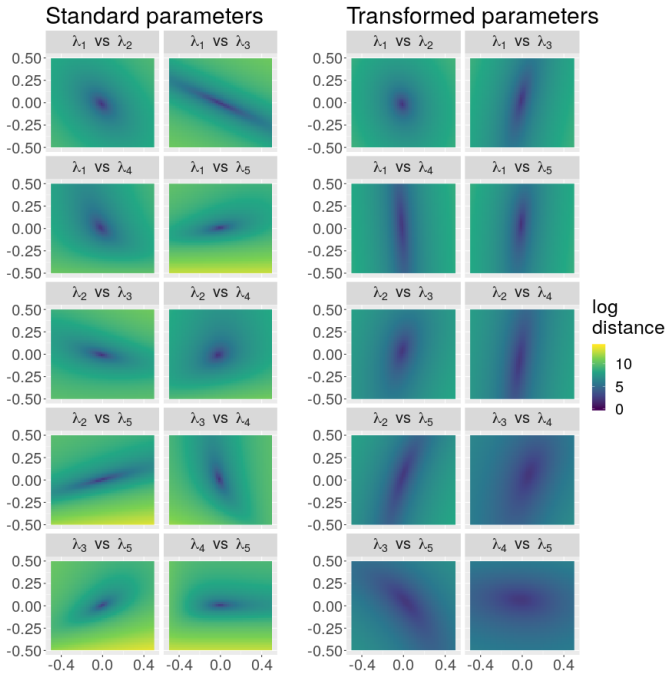

Figure 4.1: For the parameter values $\widehat{\boldsymbol{\lambda}} = (3, 2, 0.4, 0.2, 0.1)^T$, 20000 data samples are simulated. The squared quantile distance (4.9) between the simulated data and an FPLD($\boldsymbol{\lambda}$) is calculated and displayed, for $\boldsymbol{\lambda} \in [\widehat{\boldsymbol{\lambda}} - 0.5, \widehat{\boldsymbol{\lambda}} + 0.5]$. The plots display the logarithm of the squared quantile distance as parameters are changed pairwise. In the plot with label λ_i vs λ_j, the difference $\lambda_i - \widehat{\lambda}_i$ is displayed along the x-axis. The y-axis measures the difference $\lambda_j - \widehat{\lambda}_j$. In the right plots, labelled "Transformed parameters", we start by transforming $\widehat{\boldsymbol{\lambda}}$ to $\widehat{\boldsymbol{\lambda}}^*$, using the transformations in (3.26). The transformed parameters are then changed pairwise in the same fashion.

4.2.1 Quantile regression

Regression can be considered one of the great pillars of modern statistics. In a classical regression setting, an n-dimensional vector of observations, or responses, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ are made available. For each observation, k different numerical or categorical explanatory variables are provided. These provide important information concerning each of the observations and are collected in a design matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$, with dimension $n \times k$.

The response vector is modelled as a linear combination of the explanatory variables, plus an error term. The importance of each explanatory variable is decided by a regression coefficient vector, $\boldsymbol{\beta}$. The classical linear regression model can be written

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.10}$$

where all error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are independent and identically distributed with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$. The motivation behind any regression model is to gain information about data that have not yet been observed, or possibly to examine the effects of given explanatory variables on the available response. Gaining information about unobserved data is made possible by the model assumptions, that all observations depend on the same explanatory variables with the same regression coefficients. Already observed data are used to estimate the regression coefficient vector $\widehat{\boldsymbol{\beta}}$, which can be used for modelling the behaviour of new observations. In a classical linear regression model, the estimated conditional mean of new observations equals $\widehat{\mathbb{E}}[y_{\text{new}}] = \mathbf{x}_{\text{new}}^T \widehat{\boldsymbol{\beta}}$.

Sometimes, it is not enough to be able to predict the conditional mean of new observations. In a quantile regression setting, specific quantiles in the distribution of y is modelled, instead of the mean value. For a given p-quantile, $0 \leqslant p \leqslant 1$, the response is modelled as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_p. \tag{4.11}$$

However, the mean values of the error terms are not necessarily equal to zero. The errors must not even be identically distributed. Instead it is demanded that all error terms are independent, and that $P(\varepsilon_{ip} \leqslant 0) = p$ for all error terms in $\boldsymbol{\varepsilon}_p = (\varepsilon_{1p}, \ldots, \varepsilon_{np})^T$. This implies

$$p = P(\varepsilon_{ip} \leqslant 0) = P(\mathbf{x}_i^T \boldsymbol{\beta}_p + \varepsilon_{ip} \leqslant \mathbf{x}_i^T \boldsymbol{\beta}_p) = P(y_i \leqslant \mathbf{x}_i^T \boldsymbol{\beta}_p), \tag{4.12}$$

meaning that

$$Q_{y_i}(p|\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}_p \tag{4.13}$$

for all $i = 1, \ldots, n$. The estimator of $\boldsymbol{\beta}_p$ is found from the minimisation problem

$$\widehat{\boldsymbol{\beta}}_p = \arg\min_{\boldsymbol{\beta}_p} \sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_p), \tag{4.14}$$

where $\rho_p(\cdot)$ is the loss function

$$\rho_p(\varepsilon) = \varepsilon \cdot \left(p - I_{(-\infty,0)}(\varepsilon)\right), \tag{4.15}$$

and $I_A(\cdot)$ is the indicator function

$$I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases} \tag{4.16}$$

This stems from the fact that for any random variable y, its theoretical p-quantile, q_p, is defined as

$$q_p = \arg\min_q \mathbb{E}[\rho_p(y - p)]. \tag{4.17}$$

The estimation procedure for $\boldsymbol{\beta}$ (4.14) can be considered as the empirical equivalent to this definition of a quantile (Fahrmeir et al., 2013). Quantile regression has been applied multiple times within the field of climate science due to its absence of assumptions compared to the standard linear regression model, and its ability to model any quantile of a distribution (e.g. Tareghian and Rasmussen, 2013; Cannon, 2011).

The quantile function of the FPLD is a non-linear function of the parameters $\boldsymbol{\lambda}$. One might therefore assume that the quantile regression model, which assumes that the quantiles of the response is a linear function, is a bad fit for modelling data that is distributed as the FPLD. However, the quantile regression model does not assume that the entire quantile function of the response is a linear function. It merely assumes that a given p-quantile can be modelled as a linear combination of the available explanatory variables. Given the correct choice of explanatory variables, the quantile regression therefore might be a good fit for diurnal temperature range data, even though we assume its distribution to be that of the FPLD.

Quantile regression can also be performed using a Bayesian formulation. Assume that all error terms $\varepsilon_{ip} = y_i - \mathbf{x}_i^T \boldsymbol{\beta}_p$ are independently distributed and follows the asymmetric Laplace distribution with density

$$f_p(\varepsilon) = p(1-p)e^{-\rho_p(\varepsilon)}, \tag{4.18}$$

where $\rho_p(\cdot)$ is the loss function (4.15) from the standard quantile regression procedure (Yu and Moyeed, 2001). The likelihood function of $\boldsymbol{y}$ equals

$$L(\boldsymbol{y}|\boldsymbol{\beta}_p) = p^n(1-p)^n \exp\left\{ -\sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_p) \right\}, \tag{4.19}$$

and we see that the value of $\boldsymbol{\beta}_p$ that minimises (4.14) also maximises the likelihood (4.19). The posterior distribution of $\boldsymbol{\beta}_p$ equals

$$\pi(\boldsymbol{\beta}_p|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\beta}_p)p(\boldsymbol{\beta}_p), \tag{4.20}$$

for some appropriate prior distribution $p(\boldsymbol{\beta}_p)$. The estimator $\widehat{\boldsymbol{\beta}}_p$ is found

using the expected value of the posterior distribution of $\boldsymbol{\beta}_p$. Calculating the expected value of the posterior distribution (4.20) is incredibly costly. However, this value can be estimated, using a Markov Chain Monte Carlo method.

Methods for Bayesian statistics are often more computationally intensive. However, they come with the property that one can estimate the entire posterior distribution of some parameter, not only the mean value or some other given statistic of the distribution. We choose to focus on the Bayesian implementation of quantile regression.

4.2.2 Sampling from the posterior of $\boldsymbol{\beta}_p$

Kozumi and Kobayashi (2011) demonstrate that it is possible to simulate from the posterior distribution of $\boldsymbol{\beta}_p$ (4.20) using a Gibbs sampler, which often is preferred due to its high rate of convergence. This is performed by utilising a mixture representation of the asymmetric Laplace distribution.

Proposition (Kozumi and Kobayashi, 2011)**.** Let z be a standard exponential variable and u a standard normal variable. The random variable

$$\varepsilon = \theta_p z + \gamma_p \sqrt{z} u, \tag{4.21}$$

with

$$\theta_p = \frac{1-2p}{p(1-p)}, \text{ and } \gamma_p^2 = \frac{2}{p(1-p)}, \tag{4.22}$$

follows the asymmetric Laplace distribution with parameter p (4.18).

This can be proved by comparing the characteristic function of ε with that of a random variable distributed according to the asymmetric Laplace distribution (4.18). First, the characteristic function of ε is found:

$$\begin{aligned}
\phi_\varepsilon &= \mathbb{E}\left[e^{it(\theta_p z + \gamma_p \sqrt{z} u)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{it(\theta_p z + \gamma_p \sqrt{z} u)} \mid z\right]\right] \\
&= \mathbb{E}\left[e^{z(it\theta_p - \frac{1}{2}t^2\gamma_p^2)}\right] \\
&= \left(1 + \frac{1}{2}t^2\gamma_p^2 - it\theta_p\right)^{-1}.
\end{aligned} \tag{4.23}$$

Then, the characteristic function of a random variable v following the asym-

metric Laplace distribution is calculated:

$$\begin{aligned}
\phi_v &= \mathbb{E}\left[e^{itv}\right] \\
&= \int_{-\infty}^{\infty} p(1-p)\exp\left\{v\left(it - \left(p - I_{(-\infty,0)}(v)\right)\right)\right\}\mathrm{d}v \\
&= p(1-p)\left(\int_{-\infty}^{0}\exp\left\{v\left(it-(1-p)\right)\right\}\mathrm{d}v + \int_{0}^{\infty}\exp\left\{v\left(it-p\right)\right\}\mathrm{d}v\right) \\
&= p(1-p)\left(\frac{1}{(1-p)+it} + \frac{1}{p-it}\right) \\
&= \left(1 + t^2\frac{1}{p(1-p)} - it\frac{1-2p}{p(1-p)}\right)^{-1} \\
&= \left(1 + \frac{1}{2}t^2\gamma_p^2 - it\theta_p\right)^{-1}.
\end{aligned} \tag{4.24}$$

The two characteristic functions are identical, meaning that the distributions of v and ε are identical.

One can now rewrite the quantile regression response $\boldsymbol{y}$ as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_p = \mathbf{X}\boldsymbol{\beta}_p + \theta_p\boldsymbol{z} + \gamma_p\sqrt{\boldsymbol{z}}\boldsymbol{u}, \tag{4.25}$$

where all z_i and u_i are mutually independent, $i = 1, 2, \ldots, n$. Following Kozumi and Kobayashi (2011), we assume a multivariate normal prior for $\boldsymbol{\beta}_p$ on the form

$$\boldsymbol{\beta}_p \sim N(\boldsymbol{\beta}_{p0}, \boldsymbol{B}_{p0}), \tag{4.26}$$

with hyperparameters $\boldsymbol{\beta}_{p0}$ and $\boldsymbol{B}_{p0}$. Yu and Moyeed (2001) show that all posterior moments of $\boldsymbol{\beta}_p$ exist under a normal prior. The full joint probability density function of $\boldsymbol{y}$, $\boldsymbol{\beta}_p$, and $\boldsymbol{z}$ can be expressed as

$$\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\beta}_p, \boldsymbol{z}) =& f(\boldsymbol{y}|\boldsymbol{\beta}_p, \boldsymbol{z})f(\boldsymbol{\beta}_p)f(\boldsymbol{z}) \\
&\propto \left(\prod_{i=1}^{n} z_i^{1/2}\right)\exp\left\{-\sum_{i=1}^{n}\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_p - \theta_p z_i)^2}{2\gamma_p^2 z_i}\right\}\times \\
&\quad \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_{p0})^T\boldsymbol{B}_{p0}^{-1}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_{p0})\right\}\times\exp\left\{-\sum_{i=1}^{n} z_i\right\}.
\end{aligned} \tag{4.27}$$

The kernels of the full conditional distributions of $\boldsymbol{\beta}_p$ and $\boldsymbol{z}$ can be extracted from (4.27). One can find that the full conditional distributions are proportional to known distributions. The posterior distribution of $\boldsymbol{\beta}_p$ follows a Gaussian multivariate distribution,

$$[\boldsymbol{\beta_p}|\boldsymbol{y}, \boldsymbol{z}] \sim N(\widehat{\boldsymbol{\beta}}_p, \widehat{\boldsymbol{B}}_p), \tag{4.28}$$

with

$$\widehat{\boldsymbol{B}}_p^{-1} = \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^T}{\gamma_p^2 z_i} + \boldsymbol{B}_{p0}^{-1} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_p = \widehat{\boldsymbol{B}}_p \left\{ \sum_{i=1}^n \frac{\mathbf{x}_i (y_i \theta_p z_i)}{\gamma_p^2 z_i} + \boldsymbol{B}_{p0}^{-1} \boldsymbol{\beta}_{p0} \right\}. \tag{4.29}$$

The full conditional distribution of z_i is proportional to a generalised inverse Gaussian distribution,

$$[z_i | \boldsymbol{y}, \boldsymbol{\beta}_p, \boldsymbol{z}_{-i}] \sim GIG\left(\frac{1}{2}, a_{ip}, b_p\right), \ i = 1, \ldots, n, \tag{4.30}$$

where $a_{ip} = (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_p)^2 / \gamma_p^2$, $b_p = 2 + \theta_p^2 / \gamma_p^2$. The notation $\boldsymbol{z}_{-i}$ is used for the vector $\boldsymbol{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)^T$. The probability density function of a generalised inverse Gaussian variable is given by

$$f(z; \nu, a, b) = \frac{(b/a)^{\nu/2}}{2K_\nu(\sqrt{ab})} z^{\nu-1} \exp\left\{ -\frac{1}{2}\left(\frac{a}{z} + bz\right) \right\}, \tag{4.31}$$

where $K_\nu(\cdot)$ is a modified Bessel function of the third kind (Kozumi and Kobayashi, 2011).

The full conditional distributions of both $\boldsymbol{\beta}$ and $\boldsymbol{z}$ are known parametric distributions from which we can sample directly. It is therefore possible to create a Gibbs sampler for estimating the posterior distributions of these parameters. Our implemented Gibbs sampler is described in Algorithm 1. We choose the hyperparameters $\boldsymbol{\beta}_{p0} = \mathbf{0}$, and $\boldsymbol{B}_{p0} = 10 \cdot \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. No testing is performed for finding the optimal hyperparameters.

One can generate samples from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ using the Cholesky composition of $\boldsymbol{\Sigma}$ (e.g. Gentle, 2009). First, generate the samples x_i, $i = 1, \ldots, n$, from the standard Gaussian distribution $N(0, 1)$. The number n is the length of $\boldsymbol{\mu}$. The random vector $\boldsymbol{x}^* = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{x}$, with $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, is then multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Generating data from a generalised inverse Gaussian is not as easy as generating data from a multivariate Gaussian distribution. Several algorithms are available, the most popular being that of Dagpunar (1989). However, these methods can be quite time-consuming. We implement a generalised inverse Gaussian sampler proposed by Hörmann and Leydold (2014) that attempts to tackle some of the inefficiency problems of previously proposed algorithms. The sampling algorithm divides the parameter space into three separate domains and performs the ratio-of-uniforms method (e.g Gamerman and Lopes (2006)), using different upper and lower bounds in each domain. The ratio-of-uniforms method only depends on the algebraic kernel of a distribution, i.e. the computationally intensive calculations of the Bessel function need not be performed. The division into three separate subdomains guarantees a low rejection constant for the entire

Algorithm 1 Gibbs sampler

Input:

$\boldsymbol{y},\ p,\ \mathbf{X},\ \boldsymbol{\beta}_{p0},\ \boldsymbol{B}_{p0},\ \boldsymbol{\beta}_p^{(0)},\ \boldsymbol{z}^{(0)},\ \text{max\_iter}$

Initialise:

$n \leftarrow \text{length}(\boldsymbol{y})$
$\theta_p \leftarrow \frac{1-2p}{p(1-p)}$
$\gamma_p^2 \leftarrow \frac{2}{p(1-p)}$
$b_p \leftarrow 2 + \theta_p^2/\gamma_p^2$

Execute:

for $i = 1, 2, \ldots, \text{max\_iter}$ **do**
  $\widehat{\boldsymbol{B}}_p^{(i)} \leftarrow \left\{ \sum_{j=1}^n \frac{\mathbf{x}_j \mathbf{x}_j^T}{\gamma_p^2 z_j^{(i-1)}} + \boldsymbol{B}_{p0}^{-1} \right\}^{-1}$
  $\widehat{\boldsymbol{\beta}}_p^{(i)} \leftarrow \widehat{\boldsymbol{B}}_p^{(i)} \left\{ \sum_{j=1}^n \frac{\mathbf{x}_j (y_j - \theta_p z_j^{(i-1)})}{\gamma_p^2 z_j^{(i-1)}} + \boldsymbol{B}_{p0}^{-1} \boldsymbol{\beta}_{p0} \right\}$
  $\boldsymbol{\beta}_p^{(i)} \sim N(\widehat{\boldsymbol{\beta}}_p^{(i)}, \widehat{\boldsymbol{B}}_p^{(i)})$
  for $j = 1, 2, \ldots, n$ **do**
    $a_{jp}^{(i)} \leftarrow (y_j - \mathbf{x}_j^T \boldsymbol{\beta}_p^{(i)})^2 / \gamma_p^2$
    $z_j^{(i)} \sim GIG(\frac{1}{2}, a_{jp}^{(i)}, b_p)$
  end for
end for

Return:

$\boldsymbol{\beta}_p^{(1)}, \ldots, \boldsymbol{\beta}_p^{(\text{max\_iter})}$

parameter domain. Dagpunar (1989) also performs the ratio-of-uniforms method. However, by not dividing the parameter space into smaller subdomains, his method is not able to guarantee a uniformly bounded rejection constant (Hörmann and Leydold, 2014).

After running the Gibbs sampler for some time, it will converge. Convergence implies that all values of $\boldsymbol{\beta}_p^{(i)}$ follow the posterior distribution of $\boldsymbol{\beta}_p$, for $i > k$. k is the iteration where the Gibbs sampler has reached convergence (e.g. Gamerman and Lopes, 2006). In order to sample from the posterior distribution of $\boldsymbol{\beta}_p$, one must therefore run the Gibbs sampler for a large number of iterations. Having done so, one must create plots, similar to those in Figure 5.5, of $\boldsymbol{\beta}_p^{(i)}$ for $i = 1, 2, \ldots, N$. We can examine the plots for a visual inspection of the convergence properties of the Gibbs sampler. A cutoff-value k should be chosen, so that we believe the Gibbs sampler has reached convergence before iteration number k. An estimator for $\boldsymbol{\beta}_p$ is the expected mean of the posterior distribution. This is estimated, using the Gibbs sampler, as

$$\widehat{\boldsymbol{\beta}}_p = \frac{1}{N-k} \sum_{i=k+1}^{N} \boldsymbol{\beta}_p^{(i)}, \tag{4.32}$$

where N is the number of iterations of the Gibbs sampler, ans k is the cutoff-value.

4.2.3 Interpolation of the parameters of the FPLD

Having developed regression models for different quantiles in the distribution of diurnal temperature range, we wish to further apply these in the

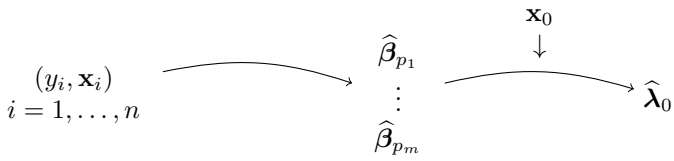

Figure 4.2: Diagram of our two-step scheme for spatial interpolation of the parameters of the FPLD.

estimation of the parameters of the FPLD. Advantageously, quantile regression pairs well with the method of quantiles. These two methods can be combined into a two-step procedure for interpolation of the parameters of the FPLD. The first step of our interpolation procedure performs quantile regression on observed data in order to estimate a set of regression

coefficients

$$\widehat{\boldsymbol{\beta}}_p = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_p \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right), \ p \in \{p_1, \ldots, p_m\}. \tag{4.33}$$

A set of explanatory variables are necessary in order to perform the regression with any success. Having performed the quantile regression, we now choose any location of interest, with the explanatory variables $\mathbf{x}_0$. For the second step of our interpolation scheme, the quantiles $\widehat{q}_{p_j} = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}_{p_j}$ are estimated, and then applied in estimating the FPLD parameters

$$\widehat{\boldsymbol{\lambda}}_0 = \arg\min_{\lambda} \sum_{j=1}^{m} \left(\widehat{q}_{p_j} - Q(p_j; \boldsymbol{\lambda}) \right)^2. \tag{4.34}$$

A diagram summarising the method is displayed in Figure 4.2.

4.3 Consistency

Neither quantile regression, nor the method of quantiles are excessively applied in the statistical literature. We therefore present some theoretical properties of the two methods. More exactly, we present the necessary conditions for the consistency of the resulting parameter estimators. Consistency of an estimator is an asymptotic requirement on the distribution of the estimator. As the sample size grows, a consistent estimator $\widehat{\theta}$ converges in probability to the true parameter θ, i.e. $||\widehat{\theta} - \theta|| \to 0$ as $n \to \infty$. This should be a minimal requirement for any statistical estimator. If the estimator is not consistent it cannot be trusted and other estimators should be applied instead. After the conditions for consistent estimators have been presented, the rate of convergence fir the consistent estimators are also examined.

4.3.1 The method of quantiles

The method of quantiles proceeds to find the FPLD quantile function $Q(p; \widehat{\boldsymbol{\lambda}})$ that is most similar in L^2 to some empirical quantile function (see Section 4.1.3). Assume $\boldsymbol{y}$ are identically and independently distributed as an FPLD with unknown parameters $\boldsymbol{\lambda}$, i.e. $y_i \sim \text{FPLD}(\boldsymbol{\lambda})$ for $i = 1, \ldots, n$. From $\boldsymbol{y}$, a set of sample p-quantiles, $\widehat{q}_p, \ p \in \{p_1, \ldots, p_m\}$, are constructed. The method of quantiles estimator for $\boldsymbol{\lambda}$ is

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \sum_{j=1}^{m} \left(\widehat{q}_{p_j} - Q(p_j; \boldsymbol{\lambda}) \right)^2. \tag{4.35}$$

Before applying this methodology, we must first establish that the parameter estimator $\widehat{\boldsymbol{\lambda}}$ converges in probability to the true parameter $\boldsymbol{\lambda}$. As stated by Sgouropoulos, Yao, and Yastremiz (2015), the estimator (4.35) is the empirical counterpart to the minimiser of

$$\int_0^1 \left\{\widehat{Q}(p) - Q(p;\boldsymbol{\lambda})\right\}^2 \mathrm{d}p, \tag{4.36}$$

where $\widehat{Q}(p)$ is the empirical quantile function of $\boldsymbol{y}$. That is to say, if m approaches infinity, then the sum in (4.35) will approach the integral in (4.36). If we can show that (4.36) converges to zero as $n \to \infty$, then so will (4.35) converge to zero as $n, m \to \infty$. The following theorem is stated:

Theorem 1 (Shorack and Wellner, 2009)**.** Assume that the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are i.i.d. according to some distribution F. The empirical distribution function of $\boldsymbol{y}$ is $F_n(y) = \sum_{i=1}^n I_{[y_i,\infty)}(y)$. Denote

$$d(F_n, F) = \left[\int_0^1 \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 \mathrm{d}t\right]^{1/2}. \tag{4.37}$$

Then, $d(F_n, F) \to 0$ as $n \to \infty$ if and only if both

$$F_n \xrightarrow{d} F \quad \text{and} \quad \int_0^1 \left[F_n^{-1}(t)\right]^2 \mathrm{d}t \to \int_0^1 \left[F^{-1}(t)\right]^2 \mathrm{d}t. \tag{4.38}$$

These conditions hold for i.i.d. random variables from an FPLD. The first condition, that $F_n \xrightarrow{d} F$, holds directly from the Glivenko-Cantelli theorem (Shorack and Wellner, 2009). The integral $\int_0^1 \left[F^{-1}(t)\right]^2 \mathrm{d}t$ is, by definition, equal to the second moment of y, $\mathbb{E}[y^2]$. As previously demonstrated in Section 3.5, the second moment of an FPLD with parameters $\boldsymbol{\lambda}$ exists if $\lambda_4, \lambda_5 > -1/2$. As long as the second moment of y exists, $\int_0^1 \left[F_n^{-1}(t)\right]^2 \mathrm{d}t \to \int_0^1 \left[F^{-1}(t)\right]^2 \mathrm{d}t$ is a direct consequence of the fact that $F_n \xrightarrow{d} F$. Thus, for $\lambda_4, \lambda_5 > -1/2$ the sum in (4.35) converges to zero for the true value of $\boldsymbol{\lambda}$.

In some situations, direct observations might not be available, but a finite set of quantiles is. The question arises then, whether the estimator from the method of moments still consistent as only $n \to \infty$, while m is constant. Koenker (2005) claims that this is true, provided that all estimated quantiles are consistent. Given m quantiles $\boldsymbol{Q}(\boldsymbol{p};\boldsymbol{\lambda})$, $\boldsymbol{p} = (p_1, \ldots, p_m)^T$ and their estimators $\widehat{\boldsymbol{q}}_{\boldsymbol{p}}$, assume that the estimated quantiles converges in distribution to the true quantiles, $\sqrt{n}(\widehat{\boldsymbol{q}}_{\boldsymbol{p}} - \boldsymbol{Q}(\boldsymbol{p};\boldsymbol{\lambda})) \rightsquigarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}(\boldsymbol{\lambda}))$ and the functions $Q(\cdot;\boldsymbol{\lambda})$ and $\boldsymbol{\Omega}(\boldsymbol{\lambda})$ satisfy some natural continuity and rank conditions. Under these assumptions, Koenker (2005) claims that

$$\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \rightsquigarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{4.39}$$

with

$$\boldsymbol{\Sigma} = (\boldsymbol{J}^T\boldsymbol{J})^{-1}\boldsymbol{J}^T\boldsymbol{\Omega}\boldsymbol{J}(\boldsymbol{J}^T\boldsymbol{J})^{-1}, \tag{4.40}$$

and

$$\boldsymbol{J} = \nabla_{\boldsymbol{\lambda}}\boldsymbol{Q}(\boldsymbol{p};\boldsymbol{\lambda}) = \begin{pmatrix} \frac{\partial Q}{\partial \lambda_1}(p_1) & \cdots & \frac{\partial Q}{\partial \lambda_5}(p_1) \\ \vdots & \ddots & \vdots \\ \frac{\partial Q}{\partial \lambda_1}(p_n) & \cdots & \frac{\partial Q}{\partial \lambda_5}(p_n) \end{pmatrix}. \tag{4.41}$$

The gradient of the quantile function of the FPLD, for a single value of p, equals

$$\nabla_{\boldsymbol{\lambda}} Q(p;\boldsymbol{\lambda}) = \begin{pmatrix} 1 \\ \frac{1}{2}\left\{(1-\lambda_3)\frac{p^{\lambda_4}-1}{\lambda_4} - (1+\lambda_3)\frac{(1-p)^{\lambda_5}-1}{\lambda_5}\right\} \\ -\frac{\lambda_2}{2}\left\{\frac{p^{\lambda_4}-1}{\lambda_4} + \frac{(1-p)^{\lambda_5}-1}{\lambda_5}\right\} \\ \frac{\lambda_2}{2}(1-\lambda_3)\frac{p^{\lambda_4}(\lambda_4 \log p - 1)+1}{\lambda_4^2} \\ -\frac{\lambda_2}{2}(1+\lambda_3)\frac{(1-p)^{\lambda_5}(\lambda_5 \log(1-p)-1)+1}{\lambda_5^2} \end{pmatrix}. \tag{4.42}$$

From (4.41) and (4.42) it is clear that the matrix $\boldsymbol{J}$ has full rank if five or more unique p-quantiles are provided. Consequently, $\boldsymbol{J}^T\boldsymbol{J}$ is non-singular. As λ_4 and λ_5 converges to zero, the gradient converges to

$$\lim_{\lambda_4,\lambda_5\to 0} \nabla_{\boldsymbol{\lambda}} Q(p;\boldsymbol{\lambda}) = \begin{pmatrix} 1 \\ \frac{1}{2}\left\{(1-\lambda_3)\log p - (1+\lambda_3)\log(1-p)\right\} \\ -\frac{\lambda_2}{2}\left\{\log p + \log(1-p)\right\} \\ \frac{3\lambda_2}{4}(1-\lambda_3)(\log p)^2 \\ -\frac{3\lambda_2}{4}(1+\lambda_3)(\log(1-p))^2 \end{pmatrix}, \tag{4.43}$$

i.e. the determinant of $\boldsymbol{J}$ is bounded away from infinity for all values of $\boldsymbol{\lambda}$ and $p \notin \{0,1\}$. Consequently, the determinant $|\boldsymbol{\Sigma}|$ is also bounded away from infinity. Thus, as the estimated quantiles converge to the true quantile values, so does $\widehat{\boldsymbol{\lambda}}$ converge to $\boldsymbol{\lambda}$.

4.3.2 Quantile regression

Assume that the vector $\boldsymbol{y}$ takes the form

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}_p + \varepsilon_{ip},\ i = 1,\ldots,n, \tag{4.44}$$

with independent, but not necessarily identically distributed, errors ε_{ip} such that $P(\varepsilon_{ip} \leqslant 0) = p$. The pth conditional quantile of y_i then equals

$$Q_{y_i}(p) = \mathbf{x}_i^T\boldsymbol{\beta}_p. \tag{4.45}$$

We want to find the conditions so that the quantile regression estimator $\widehat{\boldsymbol{\beta}}_p$ from (4.14) converges in probability to the true value, $\boldsymbol{\beta}_p$, i.e. $||\widehat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p|| \to 0$ as $n \to \infty$. Koenker (2005) performs an extensive analysis of quantile regression, and presents three conditions that together are necessary and sufficient for consistency of $\widehat{\boldsymbol{\beta}}_p$.

Condition 1 (Koenker, 2005)**.** Given that the pth quantile function of $\boldsymbol{y}|\mathbf{X}$ takes the form (4.45) with the conditional distribution functions F_i of y_i, $i = 1, \ldots, n$, the conditional distribution functions satisfy

$$\sqrt{n}(a_n(\delta) - p) \to \infty \quad \text{and} \quad \sqrt{n}(p - b_n(\delta)) \to \infty,$$

for all $\delta > 0$ as $n \to \infty$, with

$$a_n(\delta) = n^{-1} \sum_{i=1}^{n} F_i(\mathbf{x}_i^T \boldsymbol{\beta}_p - \delta)$$
$$b_n(\delta) = n^{-1} \sum_{i=1}^{n} F_i(\mathbf{x}_i^T \boldsymbol{\beta}_p + \delta).$$

Condition 2 (Koenker, 2005)**.** There exists $d > 0$ such that

$$\liminf_{n\to\infty} \inf_{\|u\|=1} n^{-1} \sum_{i=1}^{n} I_{[0,d)} \left(|\mathbf{x}_i^T u| \right) = 0.$$

Condition 3 (Koenker, 2005)**.** There exists $D > 0$ such that

$$\limsup_{n\to\infty} \sup_{\|u\|=1} n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i^T u)^2 \leqslant D.$$

Under Conditions 1, 2 and 3, the estimator $\widehat{\boldsymbol{\beta}}_p$ converges in probability to $\boldsymbol{\beta}_p$. Condition 1 is the only condition on the distribution of $\boldsymbol{y}$. It holds for any random vector $\boldsymbol{y}$ such that the cumulative distribution functions are strictly increasing functions. As seen in Section 3.4, the cumulative distribution function of an FPLD is strictly increasing if we put constrictions on λ_2 and λ_3, meaning that Condition 1 holds for the FPLD. Conditions 2 and 3 are discussed further in Section 4.3.3, as these depend on the specific choice of the design matrix.

The rate of convergence of $\widehat{\boldsymbol{\beta}}_p$ is also examined by Koenker (2005). In the setting of linear quantile regression with i.i.d. sampling, he proposes two conditions for establishing the rate of convergence.

Condition 4 (Koenker, 2005)**.** The distribution functions F_i of y_i, $i = 1, \ldots, n$, are absolutely continuous, with continuous densities f_i uniformly bounded away from 0 and ∞ at the points $f_i(Q_i(p))$.

Condition 5 (Koenker, 2005)**.** There exist positive definite matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_{1p}$ such that

(i) $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \boldsymbol{D}_0$,

(ii) $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_i(Q_i(p)) \mathbf{x}_i \mathbf{x}_i^T = \boldsymbol{D}_{1p}$,

(iii) $\max_{i=1,\ldots,n} \|\mathbf{x}_i\| / \sqrt{n} \to 0$.

Under Conditions 4 and 5, the estimator of $\boldsymbol{\beta}$ converges with the following rate:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, p(1-p)\boldsymbol{D}_{1p}^{-1}\boldsymbol{D}_0\boldsymbol{D}_{1p}^{-1}\right). \tag{4.46}$$

In Section 3.4, we find that the probability density function of the FPLD is bounded away from infinity. Additionally, the probability density function can only equal zero when $p \in \{0, 1\}$. Consequently, Condition 4 holds for data that are distributed according to the FPLD, when neither the minimum-quantile nor the maximum-quantile are included in the quantile regression. Condition 5 is mostly a reformulation of Conditions 2 and 4. If a matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ is positive definite, then $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y} > 0 \; \forall \; \boldsymbol{y} \in \mathbb{R}^k$. We get

$$\boldsymbol{y}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \boldsymbol{y} = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{y}^T \mathbf{x}_i \right)^2 \geqslant 0, \tag{4.47}$$

and

$$\boldsymbol{y}^T \left(\frac{1}{n} \sum_{i=1}^{n} f_i(Q_i(p)) \mathbf{x}_i \mathbf{x}_i^T \right) = \frac{1}{n} \sum_{i=1}^{n} f_i(Q_i(p)) \left(\boldsymbol{y}^T \mathbf{x}_i \right)^2 \geqslant 0. \tag{4.48}$$

Strictness of the inequalities in (4.47) and (4.48) is guaranteed if Conditions 2 and 4 holds. Condition 2 ensures that $\sum_{i=1}^{n} (\boldsymbol{y}^T \mathbf{x}_i)^2 > 0$, while Condition 4 ensures that $f_i(Q_i(p)) > 0 \; \forall \; i = 1, \ldots, n$. Part (iii) of Condition 5 holds whenever the maximum row of the design matrix does not converge to infinity.

Having established the rate of convergence for $\widehat{\boldsymbol{\beta}}_p$, and using the fact that $\widehat{q}_p = \mathbf{x}^T \widehat{\boldsymbol{\beta}}_p$, we are able to identify the matrix $\boldsymbol{\Omega}$ in the rate of convergence of $\widehat{\boldsymbol{\lambda}}$ in (4.40) as

$$\boldsymbol{\Omega} = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_m \end{pmatrix}, \tag{4.49}$$

where

$$\Omega_i = p_i(1-p_i) \mathbf{x}^T \boldsymbol{D}_{1p_i}^{-1} \boldsymbol{D}_0 \boldsymbol{D}_{1p_i}^{-1} \mathbf{x}. \tag{4.50}$$

Consequently, the rates of convergence of $\widehat{\boldsymbol{\beta}}_p$ and $\widehat{\boldsymbol{\lambda}}$ change with the values of $\boldsymbol{p}$ and $\mathbf{x}$ (and $\boldsymbol{\lambda}$, in the case of $\widehat{\boldsymbol{\lambda}}$), but both schemes are consistent given that the conditions on the design matrix and data observations hold.

4.3.3 Discussion of assumptions

Several of the conditions necessary for consistency of the estimators from the method of quantiles and the quantile regression put constraints on the choice of design matrix and responses. These constraints are not unreasonable and will hold for most standard applications. However, it is important to justify whether the conditions hold in the case of modelling diurnal temperature range using the FPLD.

In order to guarantee consistency of the quantile regression estimators, we assume that all observations can be written on the form $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_p + \varepsilon_{ip}$, from (4.45). This is the standard assumption in a quantile regression framework. We feel safe in assuming that the data takes this form, for some choice of explanatory variables $\mathbf{X}$. We cannot, however, claim that all observations at a given location are i.i.d. Climate is usually correlated in time. It is therefore likely that observations that are close in time will display some clear dependencies. Since we have chosen not to focus on modelling diurnal temperature range in time, we have to accept some inaccuracies. It is our belief that the temporal dependencies within each season are unable to considerably affect the general spatial trends of diurnal temperature range. We therefore choose to ignore the possible dependencies between the error terms. Conditions 1, 2 and 3 must hold in order to guarantee consistency. As discussed in Section 4.3.2, Condition 1 holds for data that is distributed following the FPLD. Condition 2 holds whenever the rows of the design matrix $\mathbf{X}$ consist of a finite set of vectors that are equal under scaling, i.e. $\mathbf{x}_i = b\mathbf{x}_j,\ b \in \mathbb{R}$. We interpret $n \to \infty$ so that not only the number of observations per location goes to infinity, but the number of locations in itself also goes to infinity. In this situation, Condition 2 holds. Condition 3 ensures that the growth of the design matrix is controlled as $n \to \infty$. This holds for a design matrix consisting of real-world location and climate data. The same can be said for Condition 5.

We have chosen to implement a Bayesian quantile regression framework based on the work of Yu and Moyeed (2001) and Kozumi and Kobayashi (2011). The formulation in (4.18) places some additional assumptions on ε_{ip}, namely that all errors are i.i.d. as an asymmetric Laplace distribution. This is not true for our data, which we assume to be distributed as the FPLD. Furthermore, we assume that the distribution of diurnal temperature range differs in space, meaning all our data are not identically distributed. Yu and Moyeed (2001) claim, based on empirical findings, that one can apply this framework no matter the original distribution of the data. Since then, this framework has been applied extensively and with

much success for data with a variety of underlying distributions (e.g. Lum, Gelfand, et al., 2012; Rodrigues and Fan, 2017). The posterior consistency of a Bayesian quantile regression under a misspecified asymmetric Laplace distribution is examined by Sriram, Ramamoorthi, Ghosh, et al. (2013). They find that consistency of the regression estimator holds under some quite mild conditions on the design matrix, for a large group of true distributions of the response, including location- and scale-models. We therefore assume that the use of an asymmetric Laplace distribution as a prior on the error terms is unable to substantially influence the results of our quantile regression.

Chapter 5

Simulation studies

In Section 4.3, we find that our chosen methods obtain consistent estimators for a reasonable choice of the design matrix and data. Additionally, the rate of convergence for the methods are found. Unfortunately, we are not able to find the matrices D_0 or D_{1p} in practice. In order to test the performance of our proposed methods we perform some simulation studies. For several different scenarios, we simulate data from distributions with known parameters. The already known parameters are then estimated using simulated data and the method of quantiles or quantile regression. The correctness and rate of convergence of the estimators are evaluated.

5.1 The method of quantiles

Different values of the parameter vector $\boldsymbol{\lambda}$ are chosen. For each vector, we simulate 100 different realisations of $\boldsymbol{y}_{n_i}^{(j)} = (y_1^{(j)}, \ldots, y_{n_i}^{(j)})^T$, $j = 1, \ldots, 100$. This is performed for a set of different lengths n_i, $i = 1, \ldots, m$. The estimators $\widehat{\boldsymbol{\lambda}}_{n_i}^{(j)}$ are found using the method of quantiles, from Section 4.1.3, for each realisation $\boldsymbol{y}_{n_i}^{(j)}$, $i = 1, \ldots, m$, $j = 1, \ldots, 100$. Credible intervals for $\widehat{\boldsymbol{\lambda}}$ are then created and displayed for each of the realisation lengths, n_i. Figure 5.1 displays the results of the simulation study for the parameter value $\boldsymbol{\lambda} = (6, 3, 0.5, 0.2, 0.4)^T$. It is clear that the estimated parameters converge to the true parameters as $n_i \to \infty$. The rate of convergence for the estimator is quite low. This is due to the fact that the approximation $F(y_{(i)}) = i/n$ converges slowly. If we replace $y_{(i)}$ with the quantile function for the true parameter, $Q(\frac{i}{n}; \boldsymbol{\lambda})$, the method of quantiles returns the exact true parameter $\boldsymbol{\lambda}$ almost all of the time and for all values of $n \geqslant 5$.

Results from the simulation study do not always look as good as those in Figure 5.1. An example can be seen in Figure 5.2. Further testing finds that the applied optimisation algorithm sometimes gets stuck in local minima, which can cause sudden spikes in the size of the credible intervals, like those seen in Figure 5.2. In Figure 5.3, all 100 estimated parameters from Figure 5.2 are plotted for each of the realisation lengths. We see that as n grows, estimated values are not spread out evenly but clearly focused in separate areas. For λ_3, e.g., there are no estimated parameters between $\lambda_3 = 0$ and $\lambda_3 = -0.5$, but at $\lambda_3 = -0.5$, several estimated values are

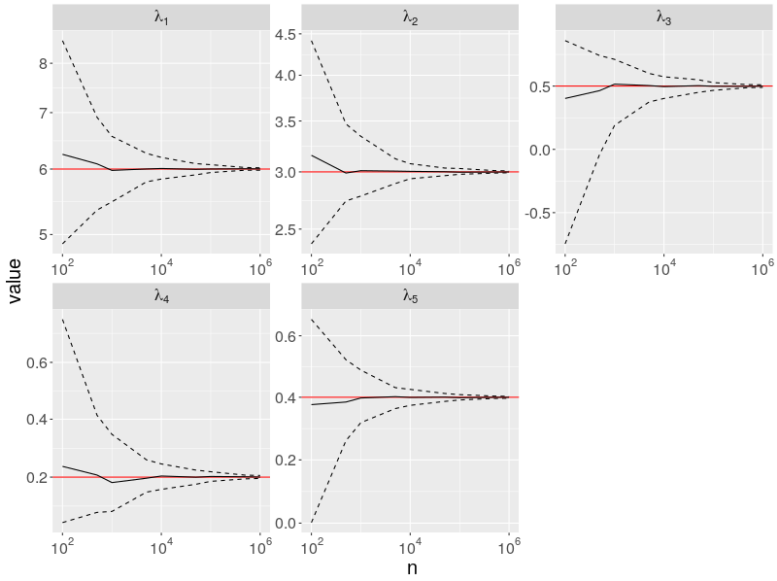

Figure 5.1: Convergence plots for $\widehat{\boldsymbol{\lambda}}$, found using the method of quantiles. For a given sample size, 100 realisations of data are simulated from an FPLD with parameters $\boldsymbol{\lambda} = (6, 3, 0.5, 0.2, 0.4)^T$. The parameters are estimated for each data sample. The medians of all estimated parameters are plotted with solid black lines. The true values of the parameters are displayed using red lines. The 0.05- and 0.95-quantiles of the estimated parameters are displayed using dashed lines.

found. This implies that there exist local minima where the optimisation algorithm gets stuck. We find that, when one of the estimated parameters have a large error, the other four are likely to have large errors as well. The red dots in Figure 5.3 display some examples of erroneous parameters that have been estimated together. Even though some of the red dots are close to their true parameter values, most of them are far away from their corresponding correct values. This implies that the problem might lie in our chosen optimisation algorithm, described in Section 4.1.3, and not in the estimation procedure in itself. The problem therefore might be fixed using some other optimisation tools. However, it is difficult to say for sure what may cause these problems. Estimating a credible interval using only 100 simulations is not optimal either. However, computations grow time-consuming as n increases. It is possible that the credible intervals would behave better if more simulations were performed.

The method of quantiles is also tested when only a small set of quantiles

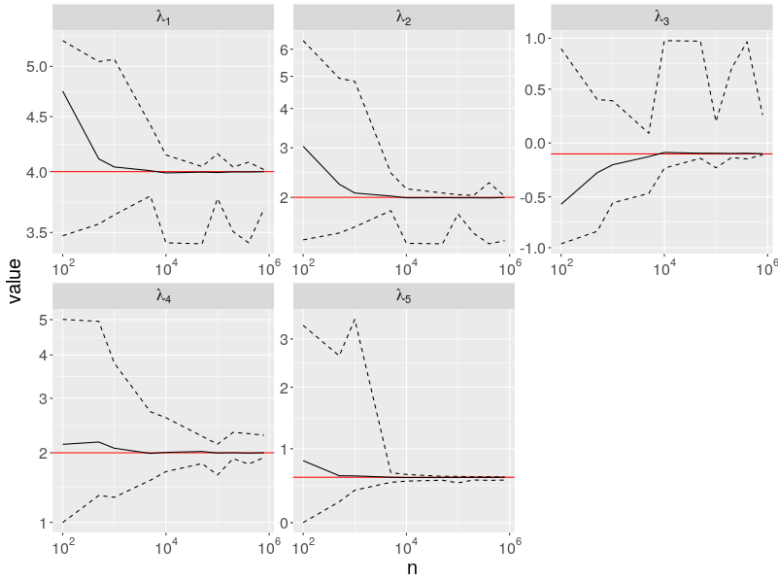

Figure 5.2: Convergence plots for $\widehat{\boldsymbol{\lambda}}$, found using the method of quantiles. For a given sample size, 100 realisations of data are simulated from an FPLD with parameters $\boldsymbol{\lambda} = (4, 2, -0.1, 2, 0.6)^T$. The parameters are estimated for each data sample. The medians of all estimated parameters are plotted with solid black lines. The true values of the parameters are displayed using red lines. The 0.05- and 0.95-quantiles of the estimated parameters are displayed using dashed lines.

are available. We simulate n_{sample} samples from the FPLD with parameters $\boldsymbol{\lambda} = (4, 3, 0.1, 0.7, 0.2)^T$. A set of n_q evenly spaced sample quantiles q_{p_i} are calculated, for $p_i = i/(n_q + 1)$, $i = 1, \ldots, n_q$. The method of quantiles is applied, using the n_q sample quantiles, in order to estimate $\boldsymbol{\lambda}$. This procedure is repeated 100 times, and for several combinations of n_{sample} and n_q. For each value of n_{sample} and n_q, 95%-credible intervals are estimated for $\widehat{\boldsymbol{\lambda}}$. Figure 5.4 displays the width of these intervals. The true parameter value is contained within all of the displayed credible intervals. We interpret n_{sample} as a measure for the accuracy of the sample quantiles. It is clear from all figures that the performance of the method of quantiles increases with the number of sample quantiles, and the correctness of these. However, the contours in the plots do not take the shape of ellipses. The shapes are more those of rounded rectangles. This indicates that, if n_{sample} is small, i.e. the errors of the estimated quantiles are large, the performance of the method of quantiles cannot be improved much by simply increasing n_q, and

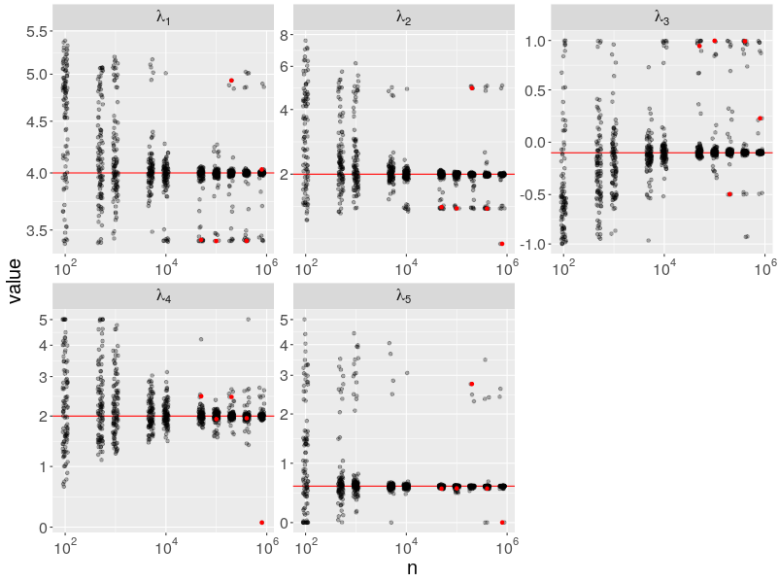

Figure 5.3: Different values of $\widehat{\boldsymbol{\lambda}}$, found using the method of quantiles. For a given sample size, 100 realisations of data are simulated from an FPLD with parameters $\boldsymbol{\lambda} = (4, 2, -0.1, 2, 0.6)^T$. The parameters are estimated for each data sample. The true values of the parameters are displayed using red lines. Red dots display parameter values that have been estimated together for different sample sizes.

vice versa.

5.2 Quantile regression

In order to test the performance of our quantile regression on data that is distributed according to an FPLD, we make use of the compact form FPLD from (3.20). If we fixate the values of λ_4 and λ_5, the FPLD can be written on the linear form

$$Q(p) = \beta_{1p} \cdot a + \beta_{2p} \cdot b + \beta_{3p} \cdot c, \tag{5.1}$$

with

$$\beta_{1p} = 1, \beta_{2p} = p^{\lambda_4}, \beta_{3p} = -(1-p)^{\lambda_5}, \tag{5.2}$$

and

$$b = \frac{\lambda_2(1-\lambda_3)}{2\lambda_4}, \quad c = \frac{\lambda_2(1+\lambda_3)}{2\lambda_5}, \quad a = \lambda_1 - b + c. \tag{5.3}$$

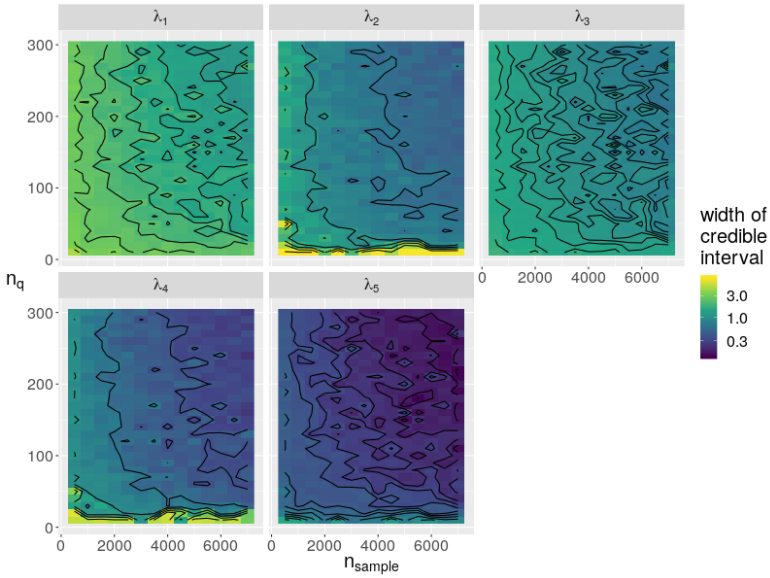

Figure 5.4: Convergence plots for $\widehat{\boldsymbol{\lambda}}$, found using the method of quantiles. For a given sample size, and a given number of quantiles, 100 realisations of data are simulated from an FPLD with parameters $\boldsymbol{\lambda} = (4, 3, 0.1, 0.7, 0.2)^T$. Credibility intervals for $\widehat{\boldsymbol{\lambda}}$ are constructed for each value of n_q and n_{sample}. The width of each credible interval is displayed in a raster plot for each of the parameters $\lambda_1, \ldots, \lambda_5$. Contours are added on top of each plot. The true parameter value $\boldsymbol{\lambda}$ is found inside all the credibility intervals for $\widehat{\boldsymbol{\lambda}}$.

We can now simulate data y_i from the quantile functions $Q(p) = a_i\beta_{1p} + b_i\beta_{2p} - c_i\beta_{3p}$, for $i = 1, \ldots, n$, where a_i, b_i and c_i can take any arbitrary values. We perform quantile regression on simulated data for varying sample sizes, and test whether the estimated regression coefficients $\widehat{\boldsymbol{\beta}}_p$ are equal to the true parameters in (5.2). Figure 5.5 displays credible intervals for all three regression coefficients for different values of p. As the sample size increases, the credible intervals quickly become very small. It is clear that this result does not mean that a quantile regression is perfect for modelling the FPLD, as any non-linear effects are removed by fixating λ_4 and λ_5. This is, however, a clear indication that the method is able to model changes in different quantiles with high performance.

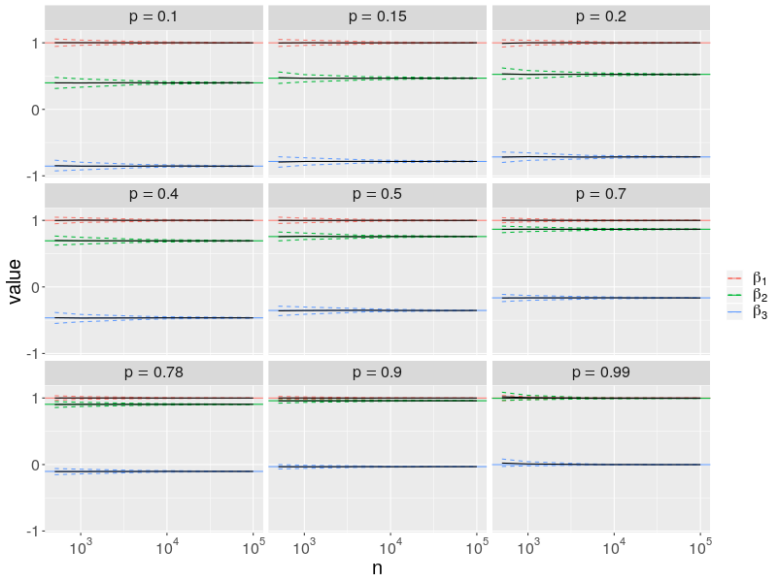

Figure 5.5: For a set of p-quantiles, the values a_i, b_i and c_i, $i = 1, \ldots, n$, are drawn uniformly from the intervals $[-10, 10]$, $[0.1, 12]$ and $[0.1, 13]$, respectively. Data y_i are then simulated using (5.1). A Bayesian quantile regression is performed on the data in order to estimate β_1, β_2 and β_3. The Gibbs sampler, described in Section 4.2.2, is run for 600 iterations, and the first 200 iterations are removed. This entire procedure is repeated 100 times. From the resulting $100 \cdot 400$ samples of each β, credible intervals are created and displayed for each value of n and p. The mean of the β-samples are plotted using black solid lines. The true values of the βs are plotted using coloured solid lines. λ_4 and λ_5 are fixated to the values 0.4 and 1.5, respectively.

Chapter 6

Case study: Diurnal temperature range in Norway

Having established methods for parameter estimation of the FPLD, we apply these to the data described in Chapter 2. The modelling of diurnal temperature range is divided into three parts. In Section 6.2, the method of quantiles is applied for local parameter estimation of the FPLD at all 55 weather stations, separately. In Section 6.3, we perform quantile regression of diurnal temperature range, without any parameter estimation of the FPLD. Two quantile regression models with different explanatory variables are fitted to the observations of diurnal temperature range, and the results are evaluated. In Section 6.4, quantile regression and the method of quantiles are combined for spatial interpolation of the parameters of the FPLD at all weather stations. The quantile regression is performed for locations with and without available observations of diurnal temperature. Eventually, all our different estimation methods are compared in Section 6.5. Evaluation of the developed models are performed using techniques described in Section 6.1. Modelling of diurnal temperature range is always performed separately for each of the four seasons of the year.

6.1 Evaluation

In order to model diurnal temperature range we must first define some methods for evaluating our results. For a first evaluation we examine different quantiles of the model fit. The spatial models are also evaluated using cross-validation, in order to test their properties for parameter estimation at locations without available temperature observations. Finally, a competing model for the median of diurnal temperature range, with spatial random effects, is created and compared with our chosen models.

6.1.1 Evaluation of quantiles

As both the method of quantiles and the quantile regression are performed by describing key quantiles of a statistical distribution, quantile-quantile plots appear as an obvious approach for qualitative evaluation. Quantile-quantile plots are created by pairwise plotting of estimated quantiles against

the quantiles of the true distribution. In this case, the true quantiles are represented by sample quantiles from the available observations of diurnal temperature range. If the quantiles are estimated correctly, they will obtain the same values as those of the available observations. The quantile-quantile plot then simply displays several dots that follow a straight line with slope 1. If, however, the estimation procedure is not correct, points in the quantile-quantile plot will not coincide with the straight line.

We examine the estimated median of diurnal temperature range, and find the discrepancies between the estimated medians and the sample medians at all weather stations,

$$d_i = \widehat{m}_i - m_i, \tag{6.1}$$

where m_i is the median at weather station number i for a given season, and $\widehat{m_i}$ is its estimator. The sample median is a highly robust statistic, meaning that the value is close to the true median of the distribution of diurnal temperature range. As discussed in Section 3.6, the parameter λ_1^* is identical to the median of the FPLD. Consequently, we are able to partially evaluate the estimator $\widehat{\boldsymbol{\lambda}}$ by comparing the estimated value of the median with the sample median of the available data. Obviously, an evaluation of the median exclusively, is not able to provide us with a complete evaluation of our estimation procedures. In general, we are interested in estimating quantiles for all possible quantiles and not only the median. However, from the data exploration in Section 2.2, we have found the spatial patterns of all quantiles of diurnal temperature range to be quite similar. Additionally, errors in the tails of the distribution of diurnal temperature range might be affected by a lack of data or other extreme events that are not accounted for. The median is not affected by such problems and is a better choice for an early examination of modelling errors. The estimator is therefore perfect for comparisons of different estimation methods, and for easy detection of flaws in our models.

6.1.2 Cross-validation

The aim of the quantile regression is to model diurnal temperature range in locations where no temperature observations are available. It is therefore of great importance to test the method using out-of-sample estimation. out-of-sample estimation means that we estimate $\widehat{\boldsymbol{\lambda}}$ for a given location without including any data from that location in the estimation procedure. The opposite of out-of-sample estimation is in-sample estimation, meaning that data from the given location is included in the estimation procedure. Cross-validation is performed for evaluation of the out-of-sample estimation properties of our chosen methods.

Cross-validation is a commonly used method for evaluating the performance of statistical learning methods (e.g. James et al., 2013). In order

to perform cross-validation, a small number of weather stations are removed from the available data. Data from the remaining stations are used to estimate all regression coefficients $\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$. Due to the model assumptions, the estimated regression coefficients should fit the data from the removed weather stations as well. Consequently, quantiles are estimated at all left-out weather stations, and we estimate $\widehat{\boldsymbol{\lambda}}$ for each of these. We can now compare the fit of the quantile regression model and the FPLD at both the left-out stations and at those that were included in the estimation procedure for $\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$. This is repeated multiple times, for ensuring that the model fit are evaluated both in-sample and out-of-sample at all weather stations. We remove five randomly selected weather stations each time.

6.1.3 Comparison of competing models

Several methods have been developed. The method of quantiles can be applied for local parameter estimation of the FPLD, while a combination of quantile regression and the method of quantiles can be applied for spatial parameter estimation of the FPLD. The spatial parameter estimation can be performed both in-sample and out-of-sample. Weaknesses in these methods can be evaluated by comparing the results from the different methods against each other.

Another potential weakness in our proposed spatial estimation scheme is that no spatial random effect is included for modelling diurnal temperature range. It is an obvious fact that observations of diurnal temperature range that are close in space are stronger correlated than observations that are far away from each other. Incorporation of such features of spatially correlated data when modelling climate data is important and can improve the performance of our estimation techniques. Consequently, we wish to evaluate how much our estimation procedure might gain from incorporating a spatial random effect, in addition to the spatial fixed effects already included in the quantile regression model. A competing model for estimation of the median of diurnal temperature range is constructed. Denote the median of diurnal temperature range as m. We wish to model the median with a spatial random field, at the geographical coordinates $\boldsymbol{s} = (\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_n^T)^T$. The spatial random field model for $\boldsymbol{m} = (m_1, \ldots, m_n)^T$ is similar to that of a standard regression model. The expected value of m_i is equal to $\mu(\boldsymbol{s}_i, \boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$ are the explanatory variables from the location $\boldsymbol{s}_i$ and $\boldsymbol{\beta}$ is a regression coefficient vector. However, in a spatial random field model, there are two different error terms. The median values equal

$$\boldsymbol{m}(\boldsymbol{s}; \boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{s}), \sigma^2) = \boldsymbol{\mu}(\boldsymbol{s}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_s(\boldsymbol{s}; \boldsymbol{\Sigma}(\boldsymbol{s})) + \boldsymbol{\varepsilon}_0(\sigma^2). \tag{6.2}$$

The error term $\boldsymbol{\varepsilon}_s$ is a multivariately distributed Gaussian random variable, with mean zero and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{s})$, representing the spatial random component of the model. The standard error terms $\boldsymbol{\varepsilon}_0$ are also

included. ε_{0i} is a univariately distributed Gaussian random variable with mean zero and variance σ^2 for $i = 1, \ldots, n$. The covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{s})$ ensures that observations which are close in space are more correlated than those far away from each other (Omre, 2018). We model the median of diurnal range with a Gaussian random field with a Matern covariance matrix. All model parameters are estimated using maximum likelihood estimation. The optimisation is performed using the `R`-package `geoR` (Ribeiro Jr and Diggle, 2018; R Core Team, 2018).

Five different methods, with different strengths and weaknesses, are now available for estimating the median of diurnal temperature range at a given weather station. These are described in Table 6.1. We can compare the results from all models to evaluate how different model deficiencies might effect our results.

6.2 The method of quantiles

The parameters of the FPLD are estimated locally at all weather stations, for all seasons, using the method of quantiles from in Section 4.1.3.

In Figure 6.1, the fitted FPLD is plotted with observations of diurnal temperature range for different locations and seasons. The observations are the same as those found in Figure 2.4. It is clear from these plots that the FPLD can take many shapes. For almost all weather stations, the fit of the distribution to observed data is quite satisfactory. For a few weather stations and some seasons, the fitted FPLD does not agree well with the observed temperature range. One example of this is the model fit at Hovden for winter (lower right plot in Figure 6.1). This is also an example of a case where the support of the fitted FPLD is unable to cover all observations of diurnal temperature range, as discussed in Section 4.1.3. Inequality constraints demanding that $Q(0; \boldsymbol{\lambda}) < y_{(1)}$ and $y_{(n)} < Q(1; \boldsymbol{\lambda})$ have been added in the optimisation scheme for the method of quantiles. However, our chosen optimisation algorithm only considers these constraints as soft constraints, i.e. failures to upheld the constraints are justified if the corresponding loss in the quantile distance is large enough. Another choice of optimisation algorithm might lead to other results where the constraints on the support hold, but the overall fit to data is worse. Still, these constraints hold for most locations and seasons.

For each weather station and season, quantile-quantile plots are made for the fitted FPLDs against observations of diurnal temperature range. Figure 6.2 displays the corresponding plots for the estimated distributions from Figure 6.1. Even for the challenging shape of diurnal temperature range seen at Hovden, most of the quantiles are fitted well to the data and problems only occur in the tails. This seems to hold for all available weather stations.

Method	Description
Spatial interpolation, in-sample (SI)	Spatial interpolation of $\boldsymbol{\lambda}$, and thus λ_1^*, using quantile regression and the method of moments. Estimation is performed in-sample.
Spatial interpolation, out-of-sample (SO)	Spatial interpolation of $\boldsymbol{\lambda}$, and thus λ_1^*, using quantile regression and the method of moments. Parameters are estimated out-of-sample. Of the 55 weather stations, 5 stations are left out in training the method. Parameters are then estimated for the five left-out stations only.
Median-field (MF)	A Gaussian spatial random field with a Matern correlation matrix is created for the median. The median is estimated out-of-sample, in the same way as the spatial interpolation method (SO).
Quantile regression (QR)	Median regression, i.e. quantile regression with $p = 0.5$, is performed. All explanatory variables from Table 6.2 are applied for the modelling. The median is estimated out-of-sample, in the same way as the spatial interpolation method (SO).
Local parameter estimation (L)	Local estimation of $\boldsymbol{\lambda}$ is performed at each weather station, separately. Estimation is performed using the method of quantiles. Estimation is performed in-sample.

Table 6.1: Five different methods for estimation of the median of diurnal temperature range at a given weather station, for a given season.

6.3 Quantile regression

Bayesian quantile regression, as described in Section 4.2, is performed on all observations of diurnal temperature range data. Explanatory variables are chosen such that they are available from geographical information systems. In addition, information concerning daily mean temperature is included, as this is available at all weather stations where the diurnal temperature range has been observed. All of the explanatory variables from Figure 2.7 are chosen for modelling of diurnal temperature range, as there are clear signs of dependencies between these variables and the quantiles of our temperature range observations. These are described in Table 6.2. Two different quantile regressions are fitted to the diurnal temperature range data. We

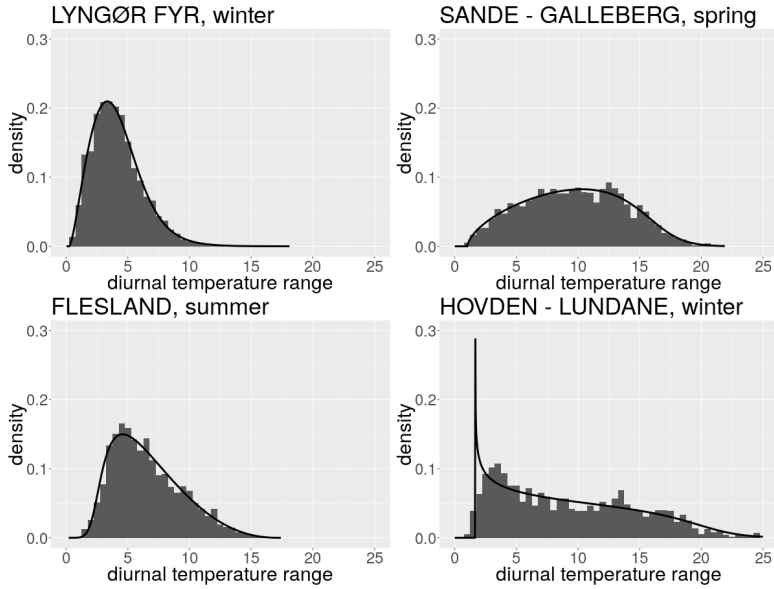

Figure 6.1: Histograms displaying observed diurnal temperature range are plotted along with the probability density functions of FPLDs with parameters estimated using the method of quantiles. All data are collected from the period 01/01/1989 - 31/12/2018. Selected seasons and locations are indicated above each plot.

first develop a purely geographical model, using $\beta_0, \ldots, \beta_4$ only. Then, the temperature information is included for a second regression model. Thus, for any $p \in (0, 1)$, we develop the following quantile regression models:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_p + \varepsilon_{ip}, \ \mathbf{x}_i = (1, x_1, x_2, \ldots, x_k)^T, \ \boldsymbol{\beta}_p = (\beta_{0p}, \beta_{1p}, \ldots, \beta_{kp})^T, \quad (6.3)$$

where y_i, $i = 1, \ldots, n$, is an observation of diurnal temperature range, and $k \in \{4, 6\}$. $\boldsymbol{\beta}_p$ and $\mathbf{x}_i$ are the corresponding regression coefficients and explanatory variables, described in Table 6.2 and ε_{ip} is an error term, described in (4.11). Geographical information like longitude and the distance to the sea can easily be found for any location of interest, using map-data available online. One might argue that observations of daily mean temperature are unavailable at most locations in general. However, while the statistical literature on modelling diurnal temperature range is lacking, a lot of effort and success has been put into the modelling of mean temperature (e.g. Maraun and Widmann, 2018; Haylock et al., 2008). We assume that there already exists satisfactory spatial and temporal models for Norway, which are able to describe the historical mean and variance of daily mean

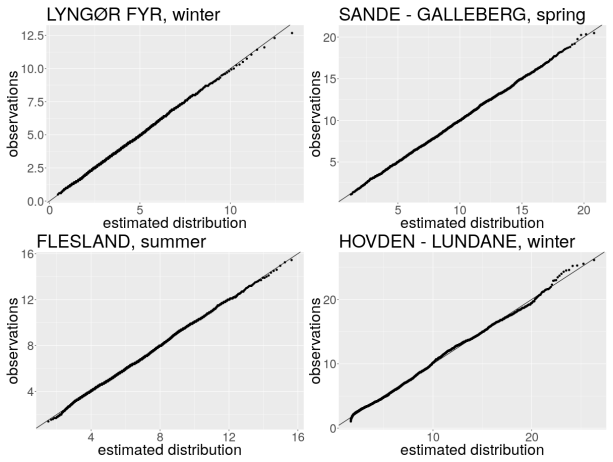

Figure 6.2: Quantile-quantile plots displaying observed diurnal temperature range against FPLDs with parameters estimated using the method of quantiles. Quantiles are plotted for $p \in \{0.001, 0.002, \ldots, 0.999\}$. All data are collected from the period 01/01/1989 - 31/12/2018. Selected seasons and locations are indicated above each plot.

temperature between 1989 and 2018 with a high performance. The Nordic Gridded Climate Data Set version 2 (Lussana, Tveito, and Uboldi, 2018), e.g., is able to model mean temperature quite successfully, even though its modelling of daily minimum and maximum temperature suffers from several inconsistencies. Accordingly, all explanatory variables can easily be provided at any location in Norway.

In order to perform quantile regression, we must first test the Gibbs sampler for convergence, as described in Section 4.2.2. For several different quantiles, output of $\boldsymbol{\beta}_p$ from each iteration is examined. We find that the Gibbs sampler has reached convergence within approximately 100 iterations for all seasons. This holds for both the purely geographical model and the model with added temperature information. Output from the Gibbs sampler is displayed in Figure 6.3. Based on these findings, we choose to run the algorithm for 600 iterations, and remove the first 150 iterations each time, when modelling diurnal temperature range. From the plots we find that the most important explanatory variables for winter seem to be the information concerning the mean daily temperature, and the intercept.

We apply the purely geographical quantile regression model to diurnal temperature range data. The model is fitted to data from all weather stations, for each season. The estimation of $\widehat{\boldsymbol{\beta}}_p$ is performed in-sample, meaning that no data is left out during the estimation procedure for $\widehat{\boldsymbol{\beta}}_p$.

Explanatory variable	Corresponding regression coefficient
Intercept	β_0
Longitude	β_1
Latitude	β_2
Altitude	β_3
Distance to the sea	β_4
Historical mean of daily mean temperature	β_5
Historical variance of daily mean temperature	β_6

Table 6.2: The explanatory variables applied in our quantile regression model, and their corresponding regression coefficients.

Results of the median regression model, i.e. quantile regression with $p = 0.5$, are displayed in Figure 6.4. It is clear that the model is able to capture many of the spatial patterns of diurnal temperature range. However, the model fit is not perfect. The model seems unable to capture large differences in temperature range over small distances. Observe, e.g., that there is a considerable jump in the median of diurnal temperature range as we move from weather stations along the coast to those further inland. This jump is not found in the fitted model.

Relative discrepancies are calculated as the difference between the estimated medians and the sample medians of diurnal temperature range, divided by the sample medians, i.e.

$$d_{\text{rel},i} = \frac{\widehat{m}_i - m_i}{m_i} = \frac{d_i}{m_i}. \tag{6.4}$$

The absolute values of the relative discrepancies between the estimated medians and sample medians from Figure 6.4 are displayed in Figure 6.5. We find that most relative median discrepancies for the winter model take values of approximately $10 - 15\%$, with some values reaching as high as 40%. For the summer model, the results are worse. Along the southern coast of Norway, most relative median discrepancies take values of $25 - 30\%$ or more, with some reaching values of more than 50%. The densities of all median discrepancies from all seasons are displayed in Figure 6.6. Both relative and actual discrepancies are displayed.

The model performance is lower for the spring and summer months. It is not immediately evident what cause these differences in performance. In Section 2.2 we find that diurnal temperature range often take higher values during spring and summer, as demonstrated for four different stations in Figure 2.2. This is also clear from Figure 6.6, where the distributions of relative median discrepancies are much more similar between all seasons

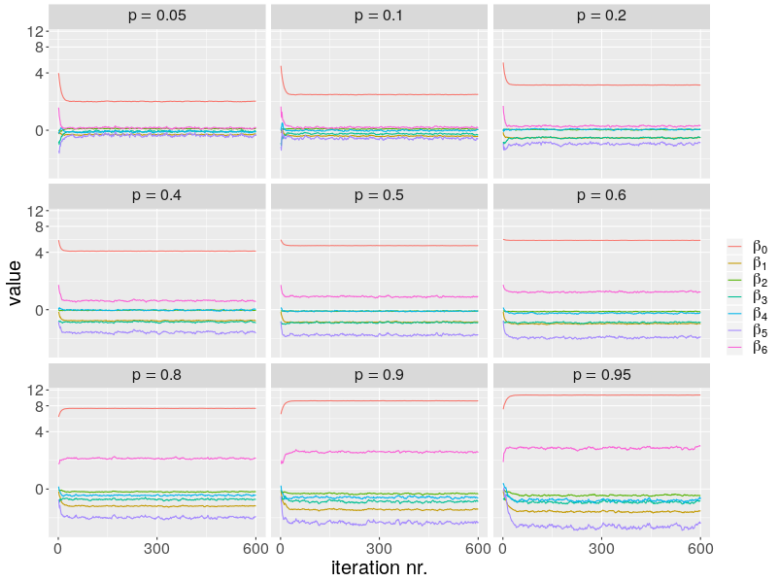

Figure 6.3: Output from the Gibbs sampler of our quantile regression model with all the explanatory variables from Table 6.2. The model is fitted to data from all 55 weather stations for nine different p-quantiles. All data are collected from the winter months each year between 1989 and 2018. All coefficients are standardised, i.e. each row of the design matrix are subtracted by their mean and divided by their sample standard deviation.

than those of the standard median discrepancies. Thus, larger values of diurnal temperature range seems to lead to larger errors. The variability in diurnal temperature range is also higher during summer and spring. This is clearly seen in Figure 2.3. During autumn and winter, the modes are more concentrated, and there is less variability in the skewness of diurnal temperature range than during spring and summer. This can make it more difficult to correctly model the quantiles of diurnal temperature range at all weather stations. Studying the median of diurnal temperature range in Figure 2.5, we also see that the variability in space is higher during summer and spring than for the other two seasons.

We calculate the mean of the medians of diurnal temperature range over all weather stations. The mean is then subtracted from all locations, and new median values are calculated. We denote these as standardised medians. The standardised medians are displayed in Figure 6.7. A blue dot represents a negative value, while red dots represent positive values.

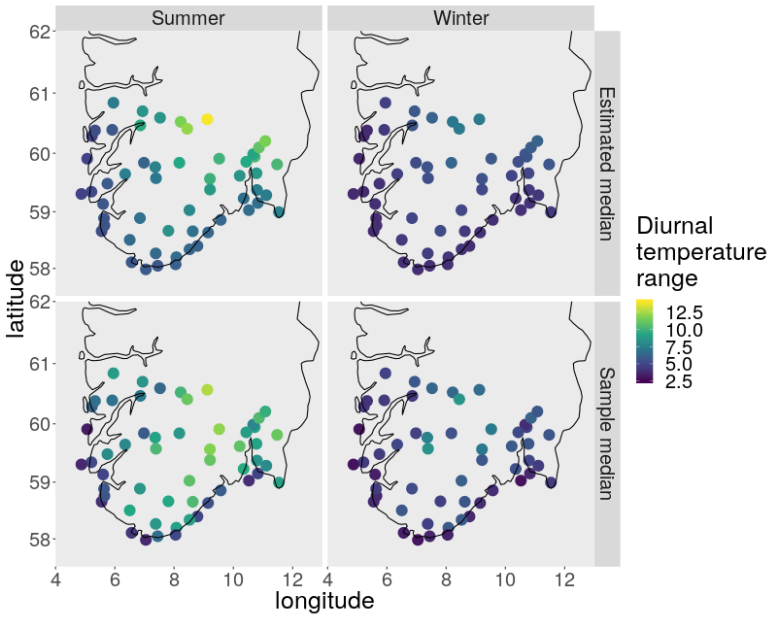

Figure 6.4: Median regression is performed in-sample for diurnal temperature range, using the data from all 55 weather stations. The data consist of diurnal temperature range for the years 1989 - 2018. The left plots consist of summer data, and the right plots consist of winter data. Lower plots display sample medians of diurnal temperature range. The upper plots display estimated medians from the purely geographical quantile regression model with regression coefficients $\beta_0, \ldots, \beta_4$ from Table 6.2.

The absolute values of the standardised median is represented by the radius of each dot. Below the standardised medians, we have displayed the corresponding median discrepancies from our purely geographical median regression model. The same colour codes apply, i.e. red colouring means that the estimated median is larger than the sample median, and blue colouring means that the estimated median is smaller than the sample median. The absolute value of each discrepancy is represented by the radius of each dot. During all seasons, the change in the median of diurnal tem-

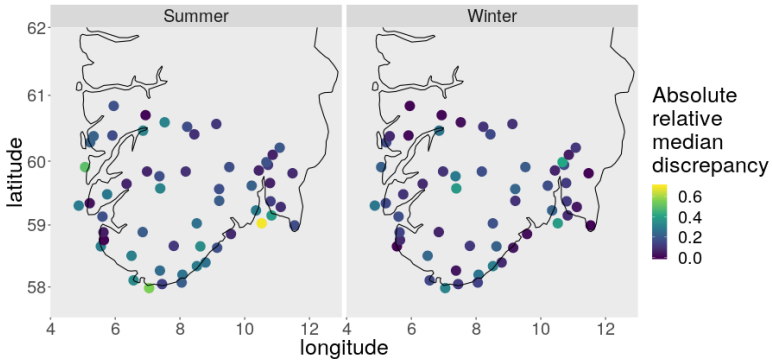

Figure 6.5: Absolute values of the relative discrepancies between estimated medians and sample medians are displayed for summer and winter, for the median regression procedure from Figure 6.4.

perature range is very abrupt as the distance from the sea is increased. The same pattern is found in all the median discrepancies. Dots along the coast are almost solely red, while those a bit further inland are blue. We find that the negative standardised medians often correspond to positive median discrepancies, and vice versa. That is to say, for low median values, the estimated median is too high, and for high median values, the estimated median is too low. This implies that the quantile regression model is unable to model the abrupt changes in the data, and instead takes values somewhere in the middle of the two extremes. A similar pattern is found for other quantiles than the median, as well.

Having examined the fit of the purely geographical quantile regression model, we now add information concerning daily mean temperature. This time all the explanatory variables in Table 6.2 are applied for modelling the quantiles of diurnal temperature range. Some of the results of the median regression model are displayed in Figure 6.8. The difference for the winter model after adding temperature is substantial. Almost all of the spatial patterns in the median of diurnal temperature range seems to have been captured, and it is hard to find any substantial differences between the estimated medians and the sample medians of diurnal temperature range. A similar result does not emerge for the summer model. The new regression model, with all explanatory variables from Table 6.2, is able to add more variability to the estimated medians of diurnal temperature range than the purely geographical model. However, clear differences can still be observed between the sample medians and the estimated medians.

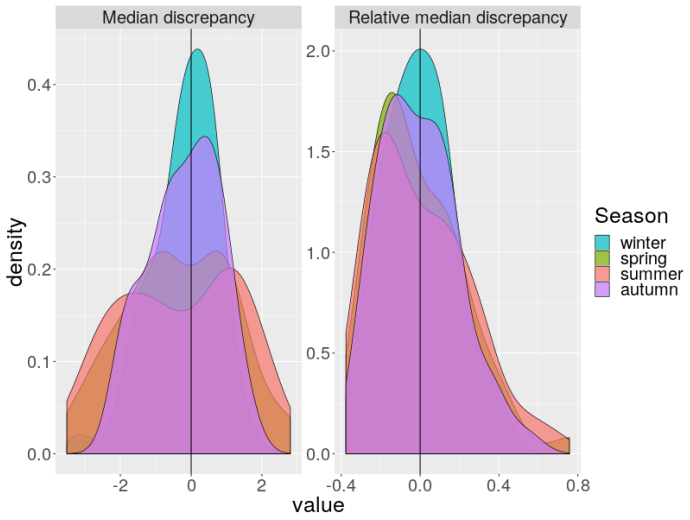

Figure 6.6: Densities for the discrepancies between sample medians and estimated medians from our purely geographical quantile regression models are displayed in the left plot. Relative discrepancies, found with (6.4), are displayed in the right plot.

The relative median discrepancies, defined in (6.4), are displayed for the quantile regression model with added temperature information. The results are found in Figure 6.9. We find that most of the relative discrepancies for the winter model have been reduced to less than 5%. The differences between estimated and sample medians have been reduced for summer, but not with nearly as much as in the winter mode. We display the densities of all median discrepancies for all seasons. These are found in Figure 6.10. This time, there is a clear difference between summer and all other seasons. The inclusion of daily mean temperature into our quantile regression model has substantially improved the performance of both winter, spring and autumn.

Similar plots to those in Figure 6.7 are created for the new median regression model. The plots are seen in Figure 6.11. The abrupt change from red to blue has disappeared along the coast for the median discrepancies for winter, spring and autumn. However, during summer, the patterns are mainly unchanged from Figure 6.5. The addition of temperature information has considerably reduced the size of most discrepancy dots from winter, spring and autumn. The lack in improvement for the summer model might be explained from Figure 2.7, where trends in the median of diurnal temperature range are plotted as functions of all explanatory variables in Table 6.2. One can find that there are strong correlations between the me-

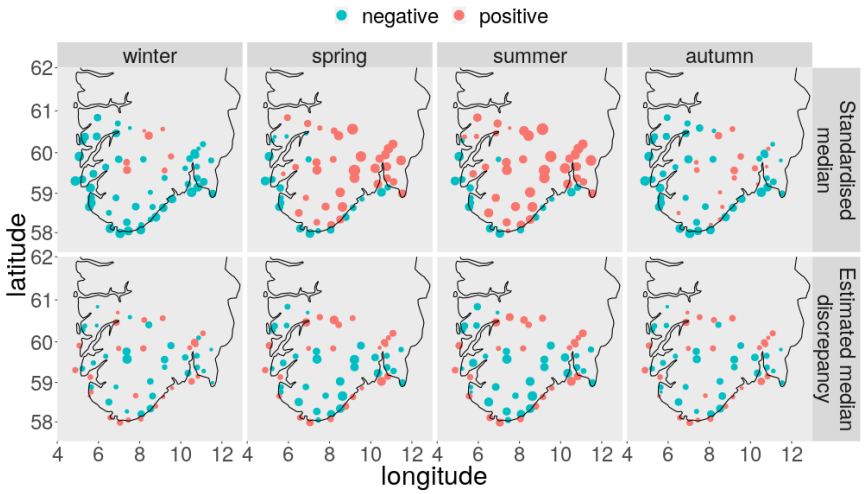

Figure 6.7: Standardised median values are found at all weather stations and displayed in the upper plots. The lower plots display the differences between estimated medians from the purely geographical quantile regression model, and the sample medians. Absolute values are represented by the radius of each dot. Negative and positive data are represented using blue and red colours.

dian of diurnal temperature range and the mean and variance of daily mean temperature for winter, spring and autumn. During summer, however, the correlation is almost completely gone and the slope between the median of diurnal temperature range and the daily mean temperature variables are close to zero. Consequently, it is not possible to gain much improvement from adding these explanatory variables for the summer model.

The quantile regression model with added temperature information obtains a good model fit for most seasons. For further evaluation of the quantile regression performance we examine the model residuals for different quantiles. Residuals are the estimated error terms in a regression model, i.e.

$$\widehat{\varepsilon}_{ip} = y_{ip} - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_p. \tag{6.5}$$

The model assumption from Section 4.2.1, is that $P(\varepsilon_{ip} \leqslant 0) = p$, meaning that we expect a fraction of p of the residuals to be negative. For each weather station we count the fraction of negative residuals. A summary of

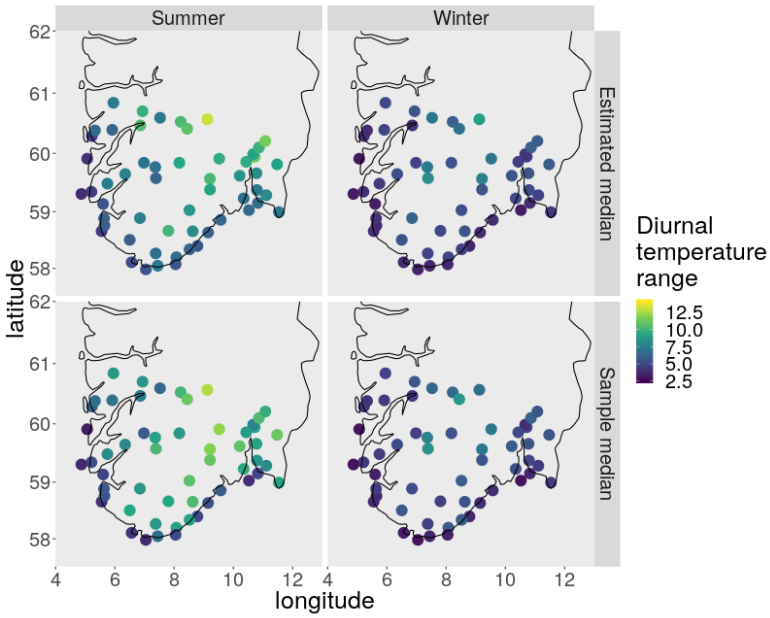

Figure 6.8: Median regression is performed in-sample for diurnal temperature range, using the data from all 55 weather stations. The data consists of diurnal temperature range for the years 1989 - 2018. The left plots consist of summer data, and the right plots consist of winter data. Lower plots display sample medians of diurnal temperature range. The upper plots display estimated medians from the quantile regression model with all regression coefficients from Table 6.2.

all fractions can be seen in Figure 6.12. The fractions of negative residuals seem to be approximately symmetrical around the value of p. However, the spread of the fraction values, especially for summer, is quite large. Consequently, the regression model seems to be well calibrated, albeit with high levels of uncertainty.

When we apply quantile regression for predicting the distribution of unobserved data, we are clearly not able to train our methods on the exact data we are trying to predict, as is done when we perform in-sample estimation. To test the performance when predictions are out-of-sample,

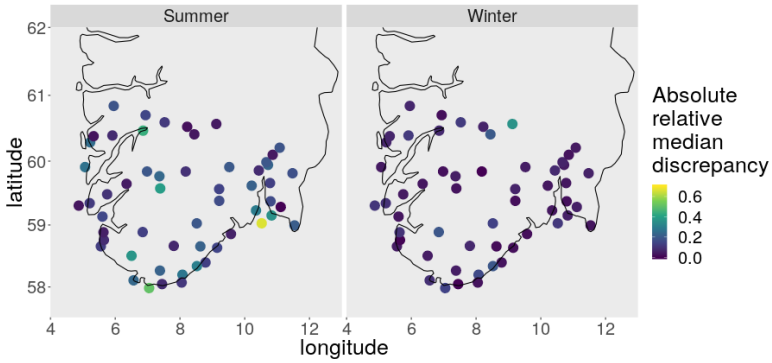

Figure 6.9: Absolute values of the relative discrepancies between estimated medians and sample medians are displayed for summer and winter, for the median regression procedure from Figure 6.8.

we perform cross-validation, as described in Section 6.1.2. The regression model with all explanatory variables from Table 6.2 is fitted to diurnal temperature range observations several times, but each time five random weather stations are removed before the training. The performance of the quantile regression is then tested on the five left-out stations. Figure 6.13 displays the distribution of the resulting discrepancies for three different quantile values. Both out-of-sample discrepancies for the 5 removed stations and in-sample discrepancies for the 50 remaining stations are presented. We find that the performance of our quantile regression scheme is slightly better in-sample than out-of-sample. Compared to the general error of the quantile regression, the additional quantile differences that stem from out-of-sample estimation are quite negligible.

6.4 Interpolation of the parameters of the FPLD

Having estimated a sufficiently large set of the quantiles of diurnal temperature range using Bayesian quantile regression, we now estimate the FPLD parameters using the method of quantiles, as described in Section 4.2.3. Quantile regression is performed for 100 equally spaced p_i-quantiles, $p_i = i/101$, $i = 1, \ldots, 100$, with all the explanatory variables from Table 6.2. The method of quantiles is applied for interpolation of the parameters of the FPLD, using the 100 estimated quantile values. This is performed out-of-sample at all weather station locations. 5 of the 55 weather stations are removed each time. Quantile regression coefficients

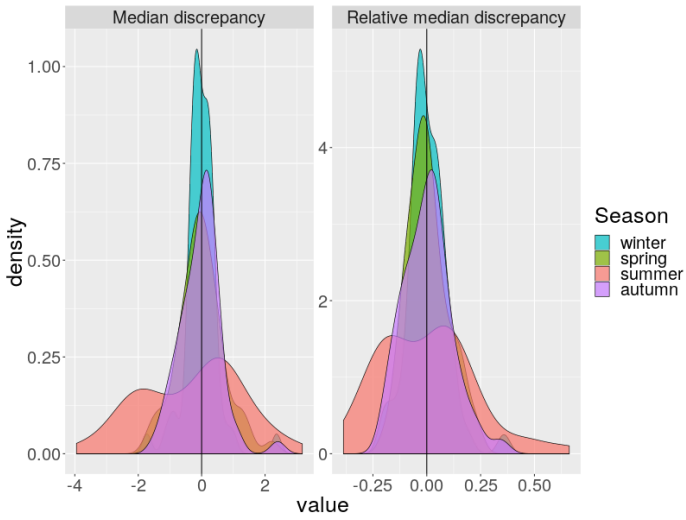

Figure 6.10: Densities for the discrepancies between sample medians and estimated medians from our quantile regression models with added temperature information are displayed in the left plot. Relative discrepancies, found with (6.4), are displayed in the right plot.

are estimated using the remaining 50 stations, and parameters of the FPLD are estimated at the 5 left-out stations. Figure 6.14 displays the estimated distributions at the same locations as those examined in Figures 2.4 and 6.1. The fit of the estimated FPLD is clearly not as good as the fit from the local parameter estimation in Figure 6.1. However, especially for winter, the results show promise. In Figure 6.15, we examine the quantile-quantile plots of the distributions from Figure 6.14. It is clear that the method is struggling with modelling of diurnal temperature range. Even for winter, the right tail of the distribution of the estimated FPLD is quite bad.

6.5 Model comparisons

We compare our different methods for modelling of diurnal temperature range, in order to evaluate their performance. Errors can be introduced in our estimation procedures in many ways, and we wish to test the main source of these. Errors are introduced both in the quantile regression model and through the method of quantiles. Errors also arise from the fact that modelling is performed out-of-sample, and possibly because a spatial random effect has not been included in our models. The question now arises, which of these are the most impactful. We perform estimation of the median

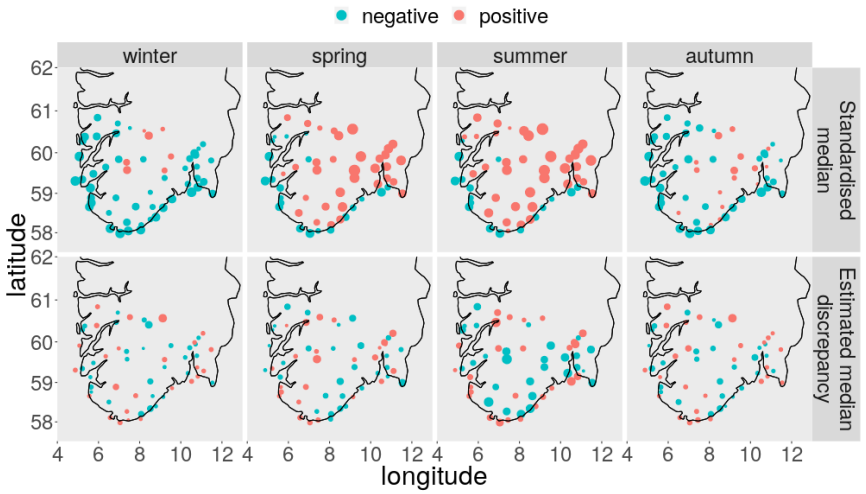

Figure 6.11: Standardised median values are found at all 55 weather stations, and displayed in the upper plots. The lower plots display the differences between the sample medians and the estimated medians from the quantile regression with added temperature information. Absolute values are represented by the radius of each dot. Negative and positive data are represented using blue and red colours.

of diurnal temperature range, for all methods described in Table 6.1, including the Gaussian random spatial field for the median, which we have not yet evaluated. For all methods, the median of diurnal temperature range is estimated at all weather stations, and the resulting median discrepancies are calculated. Densities for all the median discrepancies are displayed in Figure 6.16. It is clear that the local estimation, using the method of quantiles, performs much better than all other methods. Somewhat more surprising, we find that the differences between all other methods are close to negligible. All other methods perform spatial modelling with the same explanatory variables. The Gaussian random field for the median of diurnal temperature range applies a spatial random effect, which is lacking in the other models. This indicates that the most important shortcomings of the interpolation scheme does not stem from the lack of a spatial random effect. Errors that stem from the median regression are very similar to those from the combination of quantile regression and the method of quantiles. This implies that the method of quantiles succeeds in modelling the data

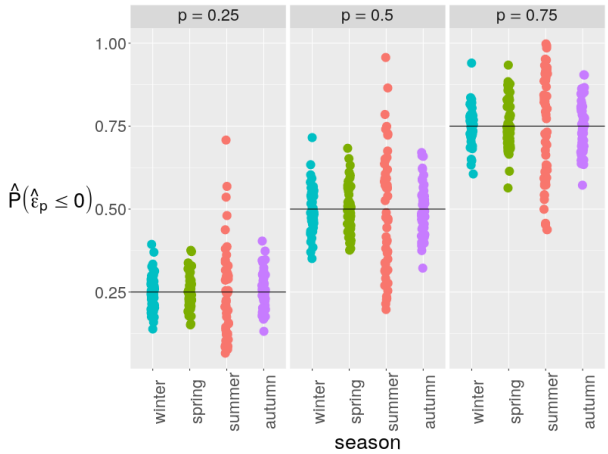

Figure 6.12: Fractions of negative quantile regression residuals, from the model with added temperature information, are calculated and plotted for each weather station and season. A horizontal line is added at the value $\widehat{P}(\widehat{\varepsilon}_p \leqslant 0) = p$.

it is given from the quantile regression. The difference between in-sample estimation and out-of-sample estimation seems close to negligible as well. Consequently, the plots indicate that the main shortcoming of our interpolation scheme stems from the errors introduced in the quantile regression model.

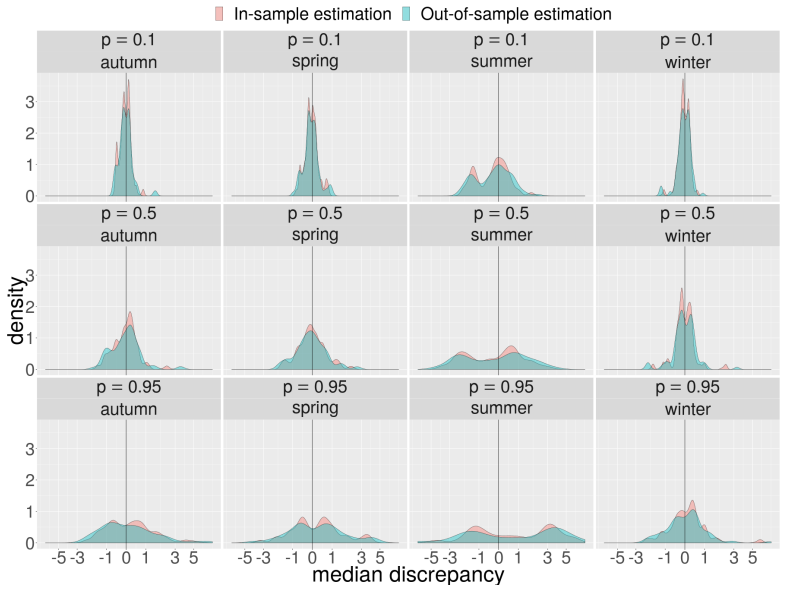

Figure 6.13: Densities of the quantile discrepancies of our quantile regression models with added temperature information are displayed for three different quantiles ($p \in \{0.1, 0.5, 0.95\}$) and all seasons. 5 stations are randomly removed, and the regression coefficients are estimated using data from the remaining 50 stations. Quantiles are estimated and discrepancies are found for the 50 stations (in-sample) and the 5 removed stations (out-of-sample). All the data are taken from the winter months for the years 1989 to 2018. A vertical line is added at $x = 0$.

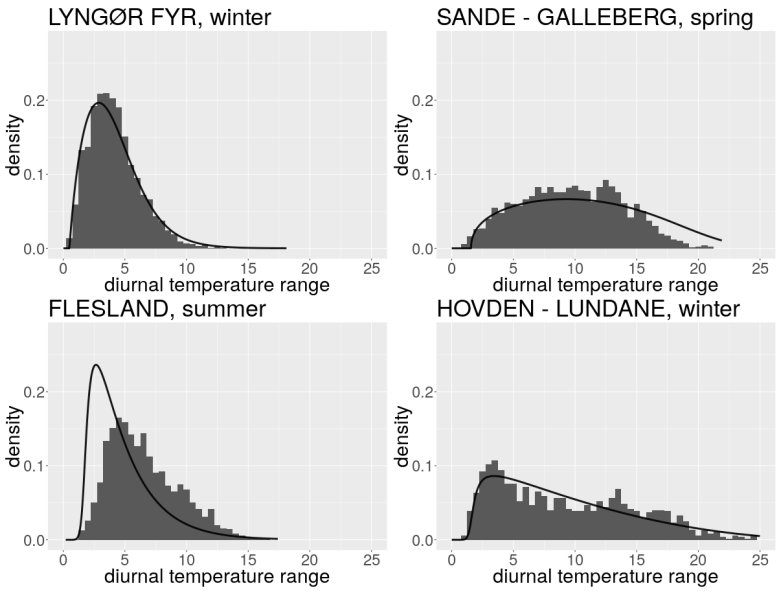

Figure 6.14: Histograms displaying observed diurnal temperature range are plotted with the probability density functions of FPLDs with parameters estimated using a combination of quantile regression and the method of quantiles. The parameters are estimated out-of-sample. All data are collected from the period 01/01/1989 - 31/12/2018. Selected seasons and locations are indicated above each plot.

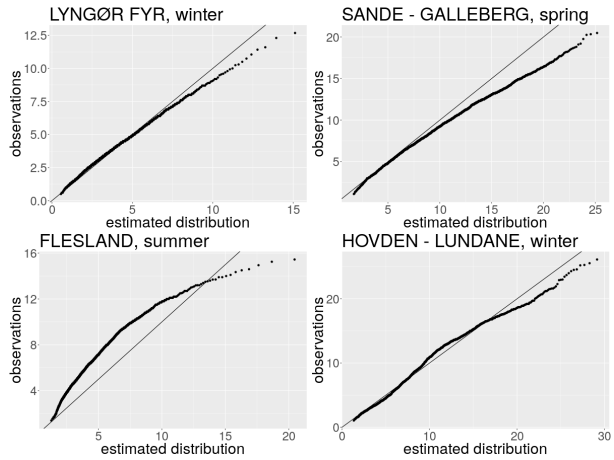

Figure 6.15: Quantile-quantile plots displaying observed diurnal temperature range against FPLDs with parameters estimated using a combination of quantile regression and the method of quantiles. The parameters are estimated out-of-sample. Quantiles are plotted for $p \in \{0.001, 0.002, \ldots, 0.999\}$. All data are collected from the period 01/01/1989 - 31/12/2018. Selected seasons and locations are indicated above each plot.

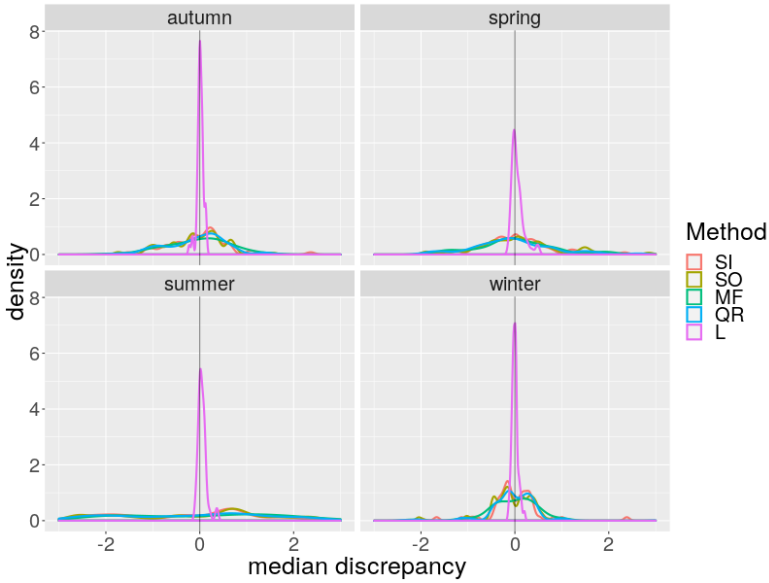

Figure 6.16: Empirical densities of the difference between the sample median of diurnal temperature range and the corresponding estimated median from different estimation procedures, for each season. The different estimation methods are described in Table 6.1. All data are taken from the years 1989 to 2018. Selected seasons are indicated above each plot. A vertical line is added at $x = 0$.

Chapter 7

Discussion

We present the Five-Parameter Lambda Distribution (FPLD) as a model for diurnal temperature range. Different techniques for local and spatial parameter estimation for the FPLD are presented. For local parameter estimation, we choose to focus on the method of quantiles, as we find it less computationally demanding than maximum likelihood estimation and better performing than the method of moments. A method for spatial interpolation of the parameters of the FPLD is developed, which combines quantile regression and the method of quantiles. We provide asymptotic results for the method of quantiles and the quantile regression, thus also providing asymptotic results for the combined interpolation method. The methods are evaluated via a simulation study. Finally, we apply our methods for modelling of diurnal temperature range observations from the southern parts of Norway.

The method of quantiles successfully estimates the correct parameters of the FPLD for all simulation studies in Section 5.1. However, convergence occurs quite slowly. This is likely due to the estimation of the sample quantiles, $F(y_{(i)}) = i/n$ in (4.9), and due to the fact that our optimisation algorithm, described in Section 4.1.3, gets stuck in local minima of the loss function. After performing local parameter estimation of the FPLD for diurnal temperature range, we find that the distribution has a good fit for the majority of the weather stations, with most quantile-quantile plots looking very good. Consequently, the FPLD appears to be a capable choice for modelling diurnal temperature range at any single location. The method of quantiles is able to estimate parameters of the FPLD with reasonably high performance when modelling diurnal temperature range. However, the implemented optimisation algorithm is unable to guarantee that the support of the FPLD is wider than the observed range of diurnal temperature range. These problems mostly occur in the left tail of the FPLD. This might be problematic when the daily mean temperature is close to zero and a small change in diurnal temperature range results in the freezing or melting of water. However, most of the time the support of the FPLD is wider than that of the available observations. Additional optimisation algorithms should be tested, with implemented hard constraints on the support of the fitted FPLD. We have only provided guarantees for a consistent parameter estimator from the method of quantiles when $\lambda_4, \lambda_5 > -1/2$. However, this

does not seem to be a problem in Norway. The right tail in the distribution of diurnal temperature range does not appear too heavy for our estimation procedure, and the maximum values of diurnal temperature range seldom reaches values of more than 30K.

The simulation study for our quantile regression model, in Section 5.2, is unable to evaluate the performance of the regression model when the response is non-linearly distributed. However, we find that the performance of the model is high when modelling data from a linear model. Since this method performs regression for all weather stations simultaneously, the number of available observations is quite sizeable. Accordingly, we expect the method to perform well if the quantiles of diurnal temperature data can truly be represented as linear combinations of the available explanatory variables. Our quantile regression model is not able to fully replicate the spatial patterns in diurnal temperature range when being given geographical explanatory variables only. However, after adding information concerning daily mean temperature, the performance is considerably better for winter, spring and autumn. For the summer model, substantial model errors are still present after the inclusion of information concerning daily mean temperature. We find that the errors most likely stem from two main factors. Information concerning daily mean temperature seems to be among the most important explanatory variables of our quantile regression models. However, during summer, the empirical correlation between daily mean temperature and the quantiles of diurnal temperature range are close to zero, as seen in Figure 2.7. This is not the case for the other three seasons. Additionally, the distribution of diurnal temperature range is has a high variability in space during summer, while there is less variability for other seasons.

All in all, we find the quantile regression model for diurnal temperature range to be very promising. It is our belief that one can achieve better results than what is found in Section 6.3. New explanatory variables can be constructed to include variable interactions, by multiplying two or more of the already available explanatory variables. The performance of our model might also increase by adding other explanatory variables. As an example, the weather station with the largest median discrepancy in the summer model is located at a lighthouse that is far away from the mainland (see Figure 6.9). Consequently, it might be reasonable to distinguish between observations from the mainland, and those out in the sea, through the introduction of a binary explanatory variable. A transformation of variables could also improve our models. The patterns between the median of diurnal temperature range and the available explanatory variables in Figure 2.7 do not seem linear for the geographical information. One might find possible transformations of the explanatory variables which are able to improve the linear trends between the quantiles of diurnal temperature range and the available explanatory variables. Additionally, we might find important de-

pendencies between diurnal temperature range and other climate variables, such as daily precipitation, wind speed and the degree of cloud cover. Especially precipitation and cloud cover have been found to be highly negatively correlated with diurnal temperature range (Zhou et al., 2009; Waqas and Athar, 2018). It is not obvious how such explanatory variables should be incorporated in our model, though. Spatial interpolation of precipitation is much more difficult than interpolation of mean temperature. Daily precipitation is also often equal to zero. For modelling the distribution of diurnal temperature range over thirty years of data, the historical mean of precipitation might therefore be misleading and should perhaps be combined with some other measure, e.g. the frequency of precipitation. It is also difficult to find representative statistics for the degree of cloud cover at a location over the last thirty years.

Examination of the differences between out-of-sample estimation and in-sample estimation, in Figures 6.13 and 6.16, find that the differences are close to negligible. This might not hold if we are able to improve the performance of the quantile regression models. However, it is a promising sign, meaning that our model might be able to generalise the spatial distribution of diurnal temperature range well. The possible errors that stem from not including a spatial random effect in the quantile regression also seems to be close to negligible. Should the performance of our quantile regression model increase, it might be necessary to include such spatial features for further improvement of the model (Lum, Gelfand, et al., 2012). However, at the time of writing, this does not seem to be the most pressing concern.

Having performed quantile regression and the method of quantiles for modelling of diurnal temperature range data, we combine these methods in order to perform spatial interpolation on the parameters of the FPLD. As the interpolation method is based on our quantile regression model, the results for summer are mediocre. However, for the other seasons, the interpolation method is repeatedly able to model diurnal temperature range with a good fit to observed data, even though the method often struggles in the tails of the distribution. We find that the method of quantiles is able to closely reproduce the patterns of the quantile regression. This leads us to the conclusion that, under a functioning quantile regression, our spatial interpolation method for the parameters of the FPLD seems able to model diurnal temperature range in Norway quite well. The interpolation method should be tested further on diurnal temperature range after an improved quantile regression has been developed.

Bibliography

Ahmadabadi, M. N., Farjami, Y., & Moghadam, M. B. (2012). A process control method based on five-parameter generalized lambda distribution. *Quality & Quantity*, *46*(4), 1097–1111.

Aitchison, J. & Brown, J. A. (1957). *The lognormal distribution with special reference to its uses in economics*. Cambridge Univ. Press.

Bhatti, S. H., Hussain, S., Ahmad, T., Aslam, M., Aftab, M., & Raza, M. A. (2018). Efficient estimation of Pareto model: Some modified percentile estimators. *PloS one*, *13*(5), e0196456.

Bignozzi, V., Macci, C., & Petrella, L. (2018). Large deviations for method-of-quantiles estimators of one-dimensional parameters. *Communications in Statistics-Theory and Methods*, 1–26.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & geosciences*, *37*(9), 1277–1284.

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, *50*(1-2), 31–49.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*. BROOKS/COLE.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer, London.

Dagpunar, J. (1989). An easily implemented generalised inverse Gaussian generator. *Communications in Statistics-Simulation and Computation*, *18*(2), 703–710.

Delaney, H. D. & Vargha, A. (2000). The Effect of Nonnormality on Student's Two-Sample T Test. Retrieved June 11, 2019, from %7Bhttps://eric.ed.gov/?id=ED443850%7D

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., & Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, *26*(2), 89–106.

Ejima, K., Pavela, G., Li, P., & Allison, D. B. (2018). Generalized lambda distribution for flexibly testing differences beyond the mean in the distribution of a dependent variable such as body mass index. *International Journal of Obesity*, *42*(4), 930.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression*. Springer.

Gallo, K. P., Easterling, D. R., & Peterson, T. C. (1996). The influence of land use/land cover on climatological values of the diurnal temperature range. *Journal of climate*, *9*(11), 2941–2944.

Gamerman, D. & Lopes, H. F. (2006). *Markov chain monte carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall.

Gentle, J. E. (2009). *Computational statistics*. Springer.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. Chapman and Hall/CRC.

Hanssen-Bauer, I., Drange, H., Førland, E., Roald, L., Børsheim, K., Hisdal, H., ... Sorteberg, A. et al. (2009). Klima i Norge 2100. *Bakgrunnsmateriale til NOU Klimatilpassing., Norsk klimasenter, Oslo, Norway*.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*, *113*(D20).

Hörmann, W. & Leydold, J. (2014). Generating generalized inverse Gaussian random variates. *Statistics and Computing*, *24*(4), 547–557. doi:10.1007/s11222-013-9387-3

Hosking, J. R. & Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, *29*(3), 339–349.

IPCC. (2014). Climate change 2014: Synthesis report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Retrieved December 1, 2018, from https://www.ipcc.ch/site/assets/uploads/2018/05/SYR_AR5_FINAL_full_wcover.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Johnson, S. G. (2011). The NLopt nonlinear-optimization package. Retrieved from http://ab-initio.mit.edu/nlopt

Kaelo, P. & Ali, M. (2006). Some variants of the controlled random search algorithm for global optimization. *Journal of optimization theory and applications*, *130*(2), 253–264.

Koenker, R. (2005). *Quantile regression*. Econometric Society monographs. Cambridge.

Kozumi, H. & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, *81*(11), 1565–1578. doi:10.1080/00949655.2010.496117

Li, H., Reitan, T., & Stenius, S. M. (2017). Regional flomfrekvensanalyse: Utvikling av nye ligninger for flomberegninger i utmålte felt i Norge. *NVEs hustrykkeri*.

Lindgren, F. (2016). Multiscale spatio-temporal modelling and large scale computation. Retrieved June 11, 2019, from http://www.maths.ed.ac.uk/~flindgre/talks/Lindgren_Smogen2016.pdf

Lum, K., Gelfand, A. E. et al. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, *7*(2), 235–258.

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., & Andersen, J. (2018). seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. *Earth System Science Data*, *10*(1), 235.

Lussana, C., Tveito, O., & Uboldi, F. (2018). Three-dimensional spatial interpolation of 2 m temperature over Norway. *Quarterly Journal of the Royal Meteorological Society*, *144*(711), 344–364.

Makowski, K., Wild, M., & Ohmura, A. (2008). Diurnal temperature range over Europe between 1950 and 2005. *Atmospheric Chemistry and Physics*, *8*(21), 6483–6498.

Maraun, D. & Widmann, M. (2018). *Statistical downscaling and bias correction for climate research*. Cambridge University Press.

Marcondes, D., Peixoto, C., & Maia, A. C. (2018). A survey of a hurdle model for heavy-tailed data based on the generalized lambda distribution. *Communications in Statistics-Theory and Methods*, 1–28.

Movahedi, M. M., Khounsiavash, M., Otadi, M., & Mosleh, M. (2017). A new statistical method for design and analyses of component tolerance. *Journal of Industrial Engineering International*, *13*(1), 59–66.

Nair, N. U., Sankaran, P., & Balakrishnan, N. (2013). *Quantile-based reliability analysis*. Springer.

Noorian, S. & Ahmadabadi, M. N. (2018). The Use of the Extended Generalized Lambda Distribution for Controlling the Statistical Process in Individual Measurements. *Statistics, Optimization & Information Computing*, *6*(4), 536–546.

Norwegian Meteorological Institute. (2019). eKlima: Free access to weather- and climate data from Norwegian Meteorological Institute from historical data to real time observations. Retrieved from http://sharki.oslo.dnmi.no/portal/page?_pageid=73,39035,73_39049&_dad=portal&_schema=PORTAL

Omre, H. (2018). *Work Title: Bayesian Spatial Inversion with Conjugate Prior Models*. unpublished.

Öztürk, A. & Dale, R. (1982). A study of fitting the generalized lambda distribution to solar radiation data. *Journal of Applied Meteorology*, *21*(7), 995–1004.

Öztürk, A. & Dale, R. F. (1985). Least squares estimation of the parameters of the generalized lambda distribution. *Technometrics*, *27*(1), 81–84.

Powell, M. J. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis* (pp. 51–67). Springer.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Ramberg, J. S. & Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, *17*(2), 78–82.

Ribeiro Jr, P. J. & Diggle, P. J. (2018). *Geor: Analysis of geostatistical data*. R package version 1.7-5.2.1. Retrieved from https://CRAN.R-project.org/package=geoR

Rodrigues, T. & Fan, Y. (2017). Regression Adjustment for Noncrossing Bayesian Quantile Regression. *Journal of Computational and Graphical Statistics*, *26*(2), 275–284. doi:10.1080/10618600.2016.1172016

Rummukainen, M. (2010). State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(1), 82–96.

Sgouropoulos, N., Yao, Q., & Yastremiz, C. (2015). Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, *110*(510), 742–759.

Shorack, G. R. & Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Sriram, K., Ramamoorthi, R., Ghosh, P. et al. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis*, *8*(2), 479–504.

Tareghian, R. & Rasmussen, P. F. (2013). Statistical downscaling of precipitation using quantile regression. *Journal of hydrology*, *487*, 122–135.

Tarsitano, A. (2004). Fitting the generalized lambda distribution to income data. In *COMPSTAT'2004 Symposium* (pp. 1861–1867). Physica-Verlag/Springer.

Tarsitano, A. (2005). Estimation of the generalized lambda distribution parameters for grouped data. *Communications in Statistics - Theory and Methods*, *34*(8), 1689–1709. doi:10.1081/STA-200066334

Tarsitano, A. (2010). Comparing estimation methods for the FPLD. *Journal of Probability and Statistics*, *2010*.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1–67.

Upadhyay, R. R. & Ezekoye, O. A. (2008). Treatment of design fire uncertainty using Quadrature Method of Moments. *Fire Safety Journal*, *43*(2), 127–139.

Vandeskog, S. M., Haugen, M., & Thorarinsdottir, T. L. (2018). Evaluation of bias corrected precipitation output from the EURO-CORDEX climate ensemble. *NR-notat SAMBA/21/2018*, pp. 20.

Vandeskog, S. M., Thorarinsdottir, T. L., & Steinsland, I. (2019). *Post-proseccing daily minimum and maximum temperature using temperature range*. unpublished.

Waqas, A. & Athar, H. (2018). Observed diurnal temperature range variations and its association with observed cloud cover in northern Pakistan. *International Journal of Climatology*, *38*(8), 3323–3336.

Yu, K. & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437–447.

Zhou, L., Dai, A., Dai, Y., Vose, R. S., Zou, C.-Z., Tian, Y., & Chen, H. (2009). Spatial dependence of diurnal temperature range trends on precipitation from 1950 to 2004. *Climate Dynamics*, *32*(2-3), 429–440.

Appendix A

Shape of the FPLD

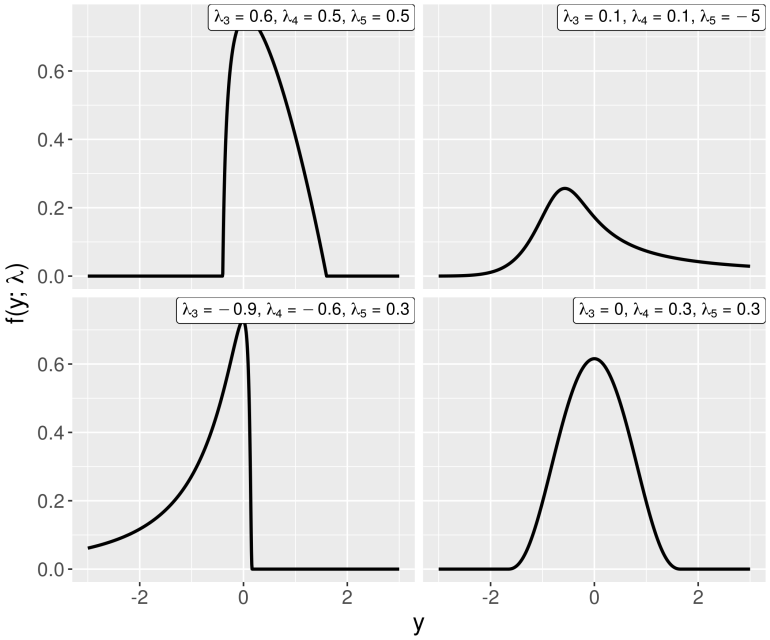

Figure A.1: Probability density function of the FPLD with different values of $\boldsymbol{\lambda}$. $\lambda_1 = 0,\ \lambda_2 = 1$ in all plots.

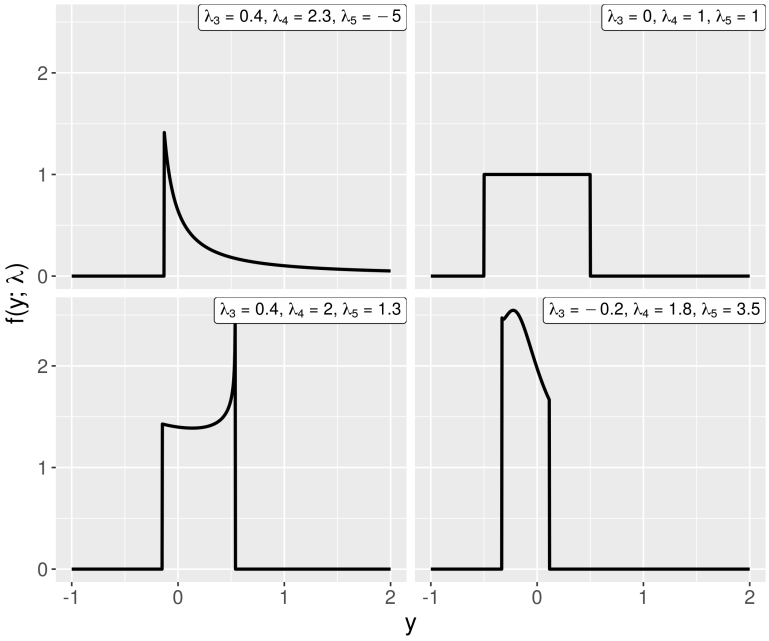

Figure A.2: Probability density function of the FPLD with different values of $\boldsymbol{\lambda}$. $\lambda_1 = 0,\ \lambda_2 = 1$ in all plots.

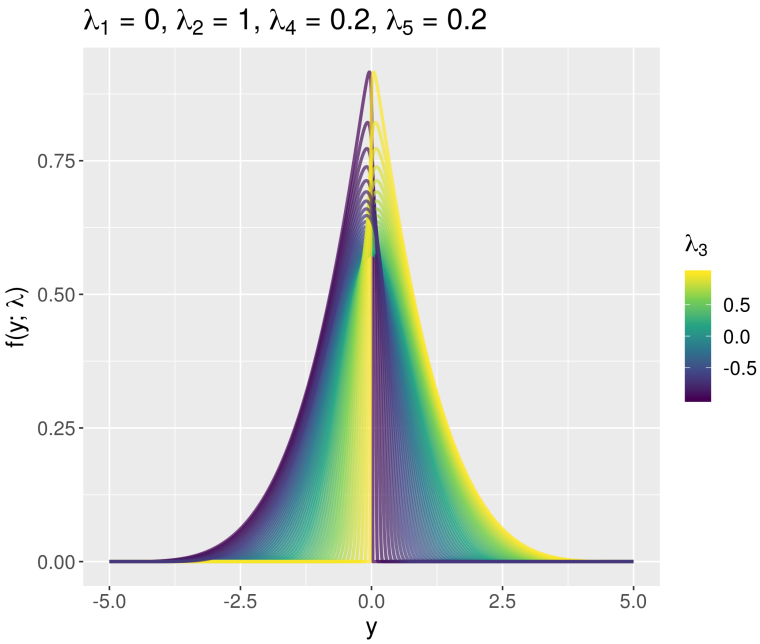

Figure A.3: Probability density function of the FPLD with different values of λ_3.

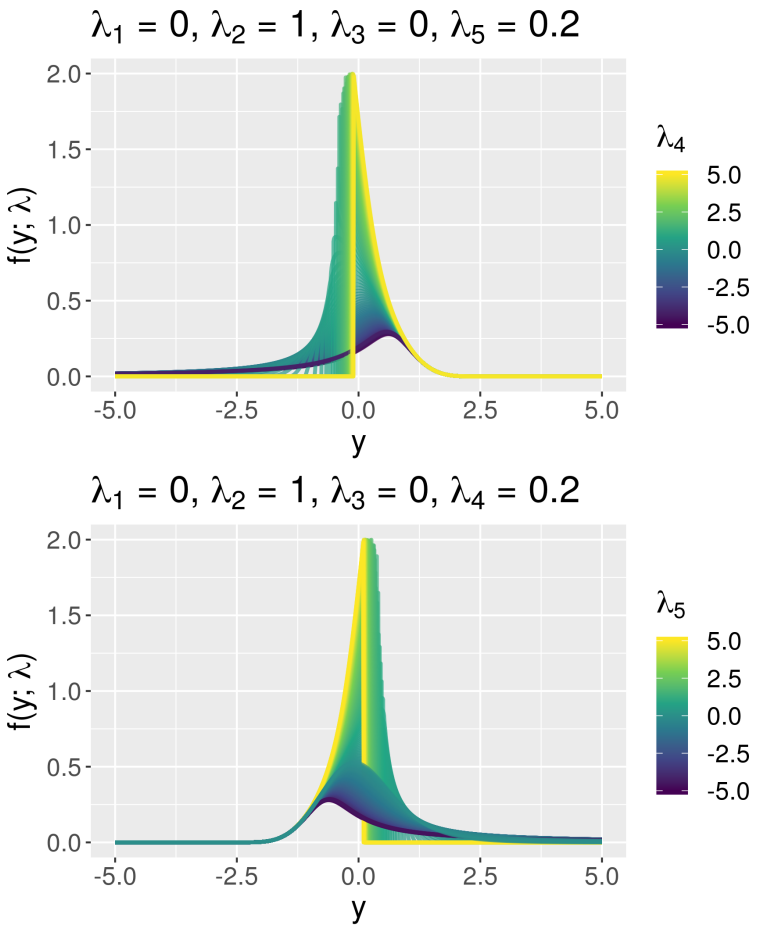

Figure A.4: probability density function of two FPLDs with different values of λ_4 and λ_5 respectively.