Dragan Runjaic, BSc

Extreme Percipitation Events in Austria: An Analysis of Return Level Estimates

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme:

Mathematics

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Mag.rer.nat. Dr.rer.nat.

Siegfried Hörmann

Institute of Statistics

Graz, May 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

This thesis is about analysing return levels of extreme precipitation events in Austria. Previously, such investigations were made on the basis of a network of meteorological measuring stations across Austria, some of them providing data over several decades. However, for the bigger part of these stations the historical time series is limited to daily precipitation amounts. Nowadays, radar based technologies allow for measurements with much higher resolution in time and in space. In turn, however, these augmented data are only available for a comparably short time span and are less exact than the ones obtained with a physical measuring device. The question we are going to pursue in this project, is whether return levels of extreme precipitation events calculated with this new data lead to satisfactory quality. To this end we do a comparison between results based on radar and classical data. We are going to see that there is a quite satisfactory correspondence. In particular, extremal rainfall levels of longer duration obtained from both methods are highly correlated. Along with an empirical analysis we provide the mathematical background from extreme value theory which is used in this context.

Contents

List of Figures

1. Introduction

In this thesis we investigate extreme rainfall events in Austria. The values that we analyse are precipitation amounts for a certain duration and return level at a variety of spatial locations. To be more specific, consider a location s (described by longitude and latitude), a duration of d units (e.g. 3 hours, 1 day, etc.) and a return level of $r > 1$ years. For given s, d, r, the goal is to determine the corresponding extreme rainfall level $\ell = \ell(s, d, r)$. If $\ell = x$, it means that the probability, that the maximal amount of rain in one year within a period of length d at location s (short $R_{\max}(s, d)$) to exceed x is equal to $1/r$:

$$P(R_{\max(s,d)} \geq x) = 1/r. \tag{1.1}$$

In other words, $\ell(s, d, r)$ is the $1/r$-quantile of the distribution of $R_{\max}(s, d)$.

Our work was motivated by the existence of two different data sets which contain estimates for $\ell(s, d, r)$ on a huge variety of arguments s, d, r. The first data set was obtained via the so-called OEKOSTRA (Oesterreichweit koordinierte Starkniederschlagsregionalisierung und - Auswertung) model and the second by the INCA (Integrated Nowcasting Through Comprehensive Analysis) model. Both systems use extreme value theory to estimate extreme rainfall levels throughout Austria. However, they are based on entirely different raw datasets. Roughly speaking, OEKOSTRA is based on precipitation measurements on a network of locations for measuring devices based at fixed spatial locations. Some of the stations collected data over more than a century. INCA, in contrast, is based on radar data which are available at a very high resolution (1×1 km). The raw data, however, only date back to 2004, which means that it contains a relatively short period for estimating the quantiles x in (1.1). The overlap of raw data for INCA and OEKOSTRA is at most 2 years (depending on the location). Hence, the two estimates for $\ell(s, d, r)$ obtained from INCA and OEKOSTRA are expected to be more or less independent. So, not surprisingly, this fact will yield some discrepancy between both models.

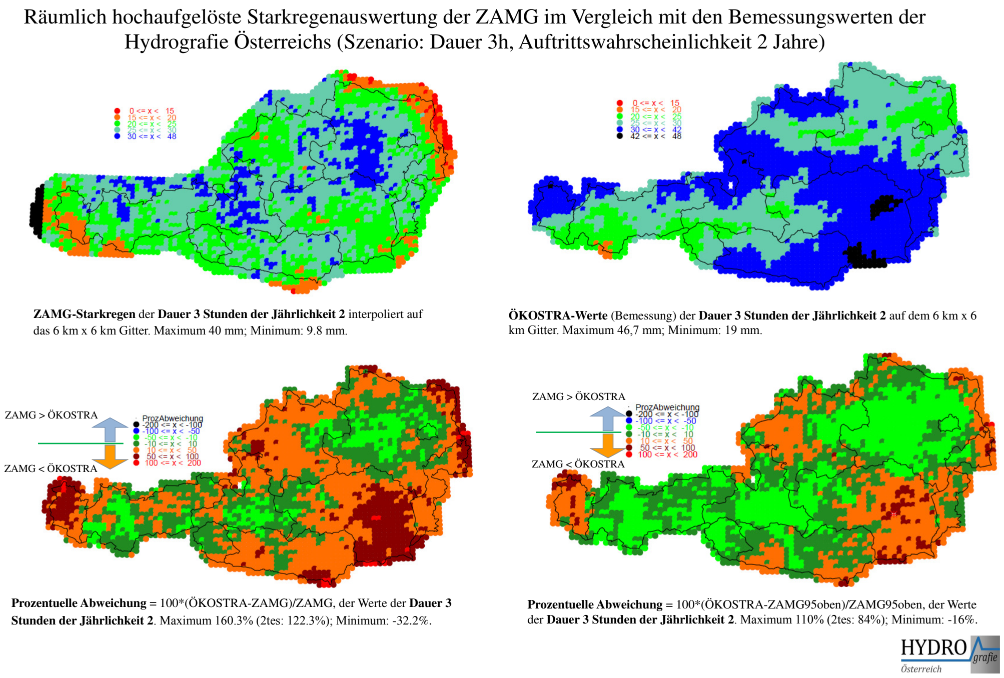
The target of this thesis is to compare these differences. In particular, we would like to answer the question whether they are within a natural range that can be explained by statistical estimation errors.

Before we outline the content of our work, let us briefly describe how this problem originated and how it evolved before we were entering the stage. The starting point is the so-called HORA project, which is an internet platform (`www.hora.gv.at`) providing Austria-wide maps for risks of different natural disasters, including hale, storms, earthquakes, floodings, etc. The platform is implemented by one of Austria's federal ministries (previously called BMLFUW) and the Austrian Insurance Association. In connection with the HORA project, the head of reinsurance of the GRAWE group (Dr. Thomas Hlatky) requested scientific support from Wegener Center Graz for comparison of the data. There was a major concern that the extreme rainfall levels calculated from the two data would be highly inconsistent. This concern was mainly based on results of the BMLFUW. See Figure 1.1. Given these large discrepancies it was unclear whether the high resolution radar-data provided by the ZAMG (Zentralanstalt für Meteorologie und Geodynamik) are realiable enough for this purpose. The reason for the huge differences seen in this evaluation remained unclear. Despite of the presumed inconsistency, the study Beck and Zingerle, 2013 concluded that `INCA` is an accurate tool to predict precipitation events in Austria.

Given that this problem has a significant statistical component, *Prof. Douglas Maraun* from the Wegener Center Graz, decided to pass the request for scientific support further to the Institute of Statistics of TU Graz. *Our goal is now to make a systematic comparison of the two data and to clarify the situation.*

The rest of the thesis is organized as follows:

In Section 2 we give a detailed survey over the datasets which we use in our analysis. In Section 3 we do an empirical comparison of the two data sets. We focus on $d = 24$ hours and $d = 3$ hours and $r = 2$ years. Then, in Section 4 and Section 5 we describe basic methods from extreme value theory, which provides the mathematical framework for this type of problem. In Section 6 we show how extremal quantiles as in (1.1) can be



Räumlich hochaufgelöste Starkregenauswertung der ZAMG im Vergleich mit den Bemessungswerten der Hydrografie Österreichs (Szenario: Dauer 3h, Auftrittswahrscheinlichkeit 2 Jahre)

ZAMG-Starkregen der **Dauer 3 Stunden der Jährlichkeit 2** interpoliert auf das 6 km x 6 km Gitter. Maximum 40 mm; Minimum: 9.8 mm.

ÖKOSTRA-Werte (Bemessung) der **Dauer 3 Stunden der Jährlichkeit 2** auf dem 6 km x 6 km Gitter. Maximum 46,7 mm; Minimum: 19 mm.

Prozentuelle Abweichung = 100*(ÖKOSTRA-ZAMG)/ZAMG, der Werte der **Dauer 3 Stunden der Jährlichkeit 2**. Maximum 160.3% (2tes: 122.3%); Minimum: -32.2%.

Prozentuelle Abweichung = 100*(ÖKOSTRA-ZAMG95oben)/ZAMG95oben, der Werte der **Dauer 3 Stunden der Jährlichkeit 2**. Maximum 110% (2tes: 84%); Minimum: -16%.

Figure 1.1.: Discrepancy in extreme rainfall levels are seen in the top two plots.

estimated. A crucial point will be to obtain confidence intervals for $\ell(s, d, r)$ and to subsequently assess whether the two methods are comparable, which will also be done in Section 6. We conclude the main part of the thesis by analyzing how sampling effects and sample size impacts the estimation results. Finally, in Appendix A, we give an overview on how our code and database is structured.

2. Datasets

As we have explained in the Introduction, we are dealing with two datasets for extreme rainfall events. We will from now on refer to these datasets as `OEKOSTRA` and `INCA`. In the following we explain these two data in detail.

2.1. The `OEKOSTRA` data

The `OEKOSTRA` data has been provided by *Dr. Viktor Weilguni* from the *Bundesministerium für Nachhaltigkeit und Tourismus*. We obtained the data in `dbf` format, divided amongst 11 files, 578Kb in size each with 918 rows and 52 columns. The columns describe mostly metadata, like names, altitude, coordinates, etc. and precipitation data for each of the 917 measuring stations in Austria. Each file is related to one return period: $r = 1, 2, 3, 5, 10, 20, 25, 30, 50, 75, 100$ (in years). The return durations are ranging from 5 minutes up to 6 days. We transformed the files into *csv* files and eliminated columns that were of no interest for our analysis.

Besides the corresponding extremal rainfall levels, the variables given in the *csv* files are provided in Table 2.1.

As already mentioned above, the `OEKOSTRA` dataset consists of records from 917 measuring sites over Austria and is based on data which altogether spans from year 1895 to 2006. The range of available data depends on the station. There is no case where data is present for the whole range from 1895 to 2006. Most stations span a length of 45 years. See Figure 2.1.

Durations for extreme rainfall events considered in the OEKOSTRA data are $d =$ 5min, 10min, 15min, 20min, 30min, 45min, 60min, 90min, 2h, 3h, 4h, 5h, 9h, 12h, 18h and 1d, 2d, 3d, 4d, 5d, 6d. Intraday precipitation measurements

`Station`	unique integer number assigned to the measuring station;
`Name`	name of the station (in most cases a nearby city);
`Waters`	nearby important lake or river;
`Land`	province;
`Owner`	owner of the station;
`Height`	meters above sealevel;
`Latitude`	latitudinal GPS coordinate;
`Longitude`	longitudinal GPS coordinate;
`Cone Coord`	cone coordinate (Bessel 1841-Ellipsoid);
`From`	first date at which data was recorded;
`To`	last date at which data was recorded;
`Years`	number of years that have recordings;

Table 2.1.: Variables in the `OEKOSTRA` dataset.

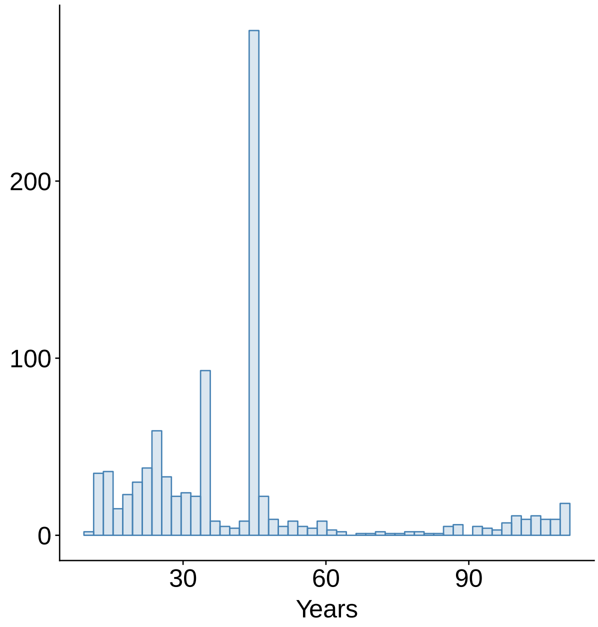


Figure 2.1.: Number of years where data is present for a station.

are only available in 221 measuring stations. Hence, extreme rainfall levels for durations $d < 1$ day are available in these stations. In the other 696 stations only the accumulated precipitation within a day was recorded

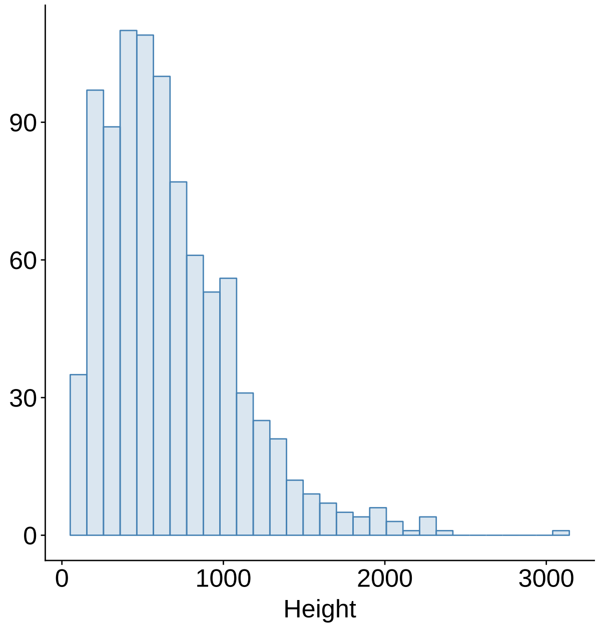


Figure 2.2.: Histogram of altitude-levels of stations.

and hence for these sites only extreme levels for durations $d \geq 1$ day are available.

It should be stressed at this point that the OEKOSTRA data contain only the extremal levels and no raw precipitation data. Extremal levels are quantities that have to be estimated using statistical theory. We will outline the underlying theory in Chapter 4. Since we have no raw data we cannot judge the quality of the estimates. We have no information about the estimation procedure used.

Due to the topological shape of Austria, the altitude of the measuring devices varies in a wide range from 117 up to 3105 meters above sealevel. Most of the stations are located between 200m and 700m. See Figure 2.2.

We observed that a station is more likely to have no intraday data the more years of data it contains. See Figure 2.4. Note that locations with only a few years of observations are not necessarily recently established measuring stations. E.g., the station "Schattendorf" has only 13 years of data ranging

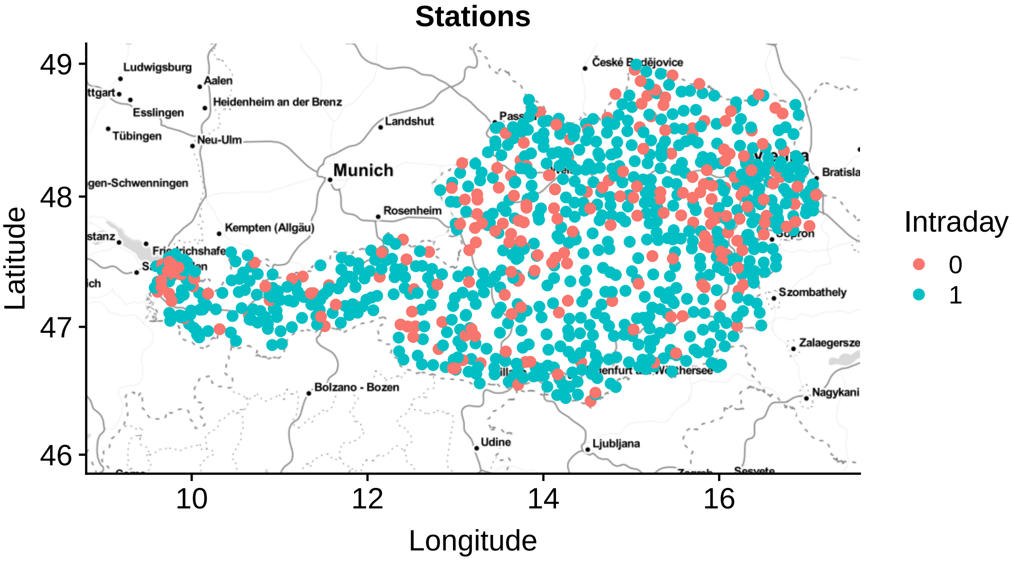


Figure 2.3.: Map of measuring locations for OEKOSTRA. We distinguish between stations where intraday data available (0), and where no intraday data available (1).

from year 1957 to 1972.

2.2. The INCA data

2.2.1. The raw precipitation dataset

The raw INCA data has been provided by *Dr. Heimo Truhetz* from the *Wegener Center* in Graz. [1] We obtained the data as ncdf4 files, which provide a high-level R interface to data files as binary data files that are portable across platforms. They also include metadata information. [2]

The data is 56,9 GB in size, divided among 160 files, each one belonging to one particular month and year starting from January 2004 until April 2017 and has a very high spatial resolution of 1km x 1km over Austria. Those 160

[1] https://wegcenter.uni-graz.at/

[2] https://cran.r-project.org/web/packages/ncdf4/ncdf4.pdf

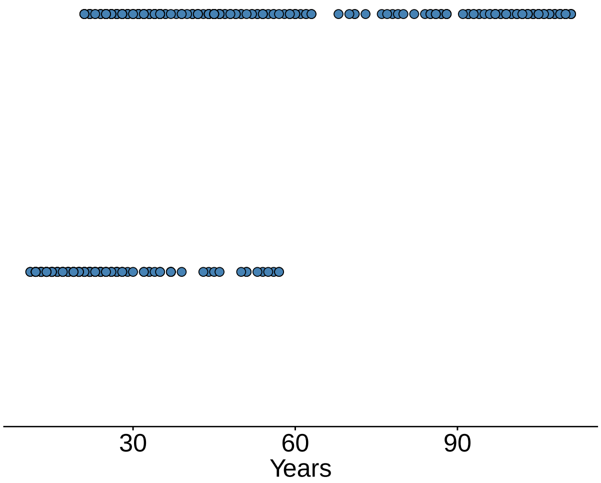


Figure 2.4.: Observation period in years. Top row: no intraday data available, bottom row: intraday data available.

files each contain precipitation values on a 1x1km grid over Austria in 15 minutes intervals. The data is complete, meaning that there are no spatial or temporal missing values on any of the available variables.

The data also includes one file containing the orography of the measuring points. It is based on the INCA system. [3]

2.2.2. The return level estimates

The return level estimates for the INCA data have been provided by *Dr. Alexander Beck* from the ZAMG [4] and were calculated on the basis of the data described in the section above. We obtained this data as a punch of plain *txt* files, each belonging to one of the durations $d = 3h, 6h, 9h, 12h, 18h, 24h, 48h,$

[3] https://www.zobodat.at/pdf/BerichteGeolBundesanstalt_88_0007-0016.pdf

[4] https://www.zamg.ac.at/cms/en

$72h$, $96h$. They are formatted in a tabular way and contain the return level estimates for $r = 1.1, 2, 3, 5, 10$ years, including 0.95% confidence intervals for the precipitation levels.

Notice that the data starts with the return period of 1.1 years. As a matter of fact, since we are considering return times of annual maxima, only return levels greater than 1 year can be considered.

Given the relatively short time period of about 13 years of available data, it was only used to predict events up to a return period of 10 years.

The return levels were estimated on the basis of the R-package *extRemes*. [5] Further information on the calculations can be found in Beck and Zingerle, 2013.

[5] https://cran.r-project.org/web/packages/extRemes/extRemes.pdf

3. An empirical comparison of the data

Our data are available for a variety of different durations (from 5 min up to 6 days) and return periods (from 1 year up to 100 years). Thus there are $21 \times 11 = 231$ possibilities to combine those two values. In this thesis we will limit ourselves to the return period of 2 years and the durations $d = 24\text{h}$ and $d = 3\text{h}$. All our methods are applicable for each other setting as well and the analysis can be extended in a straight forward way.

The aim of this chapter is to give an empirical comparison of the two datasets. In essence, the goal is to compare

$$\hat{\ell}_{\mathtt{IN}}(s, d, r) \quad \text{and} \quad \hat{\ell}_{\mathtt{OE}}(s, d, r)$$

when $r = 2$ years and $d = 3\text{h}$ or $d = 24\text{h}$, over a set of locations $s \in \mathcal{S}$. Here $\hat{\ell}_{\mathtt{IN}}$ and $\hat{\ell}_{\mathtt{OE}}$ are the estimated extreme rainfall levels provided by the `INCA` and `OEKOSTRA` data, respectively.

A first problem is that the spatial locations $\mathcal{S} = \mathcal{S}_{\mathtt{IN}}$ from the `INCA` data and the locations $\mathcal{S} = \mathcal{S}_{\mathtt{OE}}$ from the `OEKOSTRA` data are not identical. `OEKOSTRA` data are provided only at locations where there exist physical measuring devices, while `INCA` has pseudo-stations on a 1x1 km grid across Austria. Our approach here was to calculate the nearest `INCA` pseudo-station for each of the physical measuring sites from the `OEKOSTRA` data. Hence our comparison is basically restricted to $\mathcal{S}_{\mathtt{OE}}$. Due to the high resolution of `INCA` data, this approach seems justified.

This choice leads to two data vectors $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_n)'$, with $n = 917$ or $n = 221$, depending on the duration d. Basic ways to compare the relation between x and y is to look at correlation between these vectors or to regress y on x. This is done in Section 3.1. Alternatively, in order to reduce the impact of noise, we also consider some transformations based on grouping or smoothing in the subsequent sections. More precisely, we do the following:

1. *Smoothing within a given radius* $\delta > 0$: x_i is replaced by
$$\mathrm{med}\{x_j \colon |s_j - s_i| \leq \delta\}.$$
Same with y_i. Here s_i is the spatial coordinate related to x_i.
2. *Grouping the data within a range of sea-levels*: For a given $k \leq n$, the stations x_i are divided into k groups according to their sea-level $h(x_i)$. First, we re-index x_i with respect to h such that
$$h(x_1) \leq \cdots \leq h(x_n)$$
and then we calculate $m := [n/k]$ as the size for each group. Next $x = (x_1, \ldots, x_n)'$ is split into
$$x_1, \ldots, x_m, x_{m+1}, \ldots, x_{2m}, \ldots$$
and we obtain k groups $g_1, \ldots, g_k$ by setting
$$g_i := \{x_{(i-1)m+1}, \ldots, x_{im}\}$$
We associate to each group $\mathrm{med}(g_i)$ and $\min_{x_j \in g_i}\{h(x_j)\}$.
Note that for some choices of k, namely when k is not a factor of n, the last group might be left with fewer or more stations then the other groups since $n/k \notin \mathbb{N}$.
3. *Grouping within a regular grid:* For a given $k \in \mathbb{N}$ a regular $k \times k$ grid over Austria is calculated by looking at the longitude $\mathrm{lon}(x_i)$ and latitude $\mathrm{lat}(x_i)$ of each station. Let $L_{\mathrm{lon}} := \min_x \mathrm{lon}(x_i)$ and $U_{\mathrm{lon}} := \max_x \mathrm{lon}(x_i)$ and $L_{\mathrm{lat}} := \min_x \mathrm{lat}(x_i)$ and $U_{\mathrm{lat}} := \max_x \mathrm{lat}(x_i)$.
This means each grid segment g_{ij} spans a longitude of $S_{\mathrm{lon}} := (U_{\mathrm{lon}} - L_{\mathrm{lon}})/k$ and a latitude of $S_{\mathrm{lat}} := (U_{\mathrm{lat}} - L_{\mathrm{lat}})/k$. Hence one station x_k belongs to a grid segment g_{ij} if
$$iS_{\mathrm{lon}} < \mathrm{lon}(x_k) \leq (i+1)S_{\mathrm{lon}} \text{ and } jS_{\mathrm{lat}} < \mathrm{lon}(x_k) \leq (j+1)S_{\mathrm{lat}}.$$
Each grid segment g_{ij} is then associated with the $\mathrm{med}\{x_k : x_k \in g_{ij}\}$.
Note that for large values of k many of those grid segments may be empty.

Details are provided in Sections 3.2–3.4.

3.1. Comparing the raw data when $d = 24$h

We now consider the return period of $r = 2$ years and the duration $d = 24$h. With the above described choice of spatial locations s we plotted maps of Austria with the corresponding extreme rainfall levels, based on the `INCA` and `OEKOSTRA` data. See Figure 3.1. There are obvious similarities in both plots, though the `OEKOSTRA` precipitation levels are in tendency smaller. A linear relationship between x and y is suggested by a scatter plot (Figure 3.2). In this figure we have marked the corresponding sea-levels in order to explore whether they suggest an impact on the regression line. The linear relationship is furthermore suggested by the regression analysis (Equation 3.1). Figure 3.3 provides an overview on the distribution of the measuring stations with respect to the sea-level.

We regressed `OEKOSTRA` onto `INCA` and obtained the following model

$$\widehat{\texttt{INCA}} = 0.599 \cdot \texttt{OEKOSTRA} + 33.524. \tag{3.1}$$

Here is a summary of the regression analysis.

```
              Estimate   Std. Error t value Pr(>|t|)
(Intercept) 33.52376   0.99985    33.53   <2e-16
Oekostra     0.59862   0.01582    37.83   <2e-16

Residual standard error: 6.899 on 915 degrees of freedom
Multiple R-squared:   0.61,Adjusted R-squared:  0.6096
F-statistic:  1431 on 1 and 915 DF,  p-value: < 2.2e-16
```

Additionally we considered a linear model with the altitude of the measuring station as extra explanatory variable. It yields the model

$$\widehat{\texttt{INCA}} = 0.001h + 0.585 \cdot \texttt{OEKOSTRA} + 33.44, \tag{3.2}$$

and the following summary of the regression analysis:

```
              Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   3.344e+01   9.982e-01     33.500     <2e-16
```

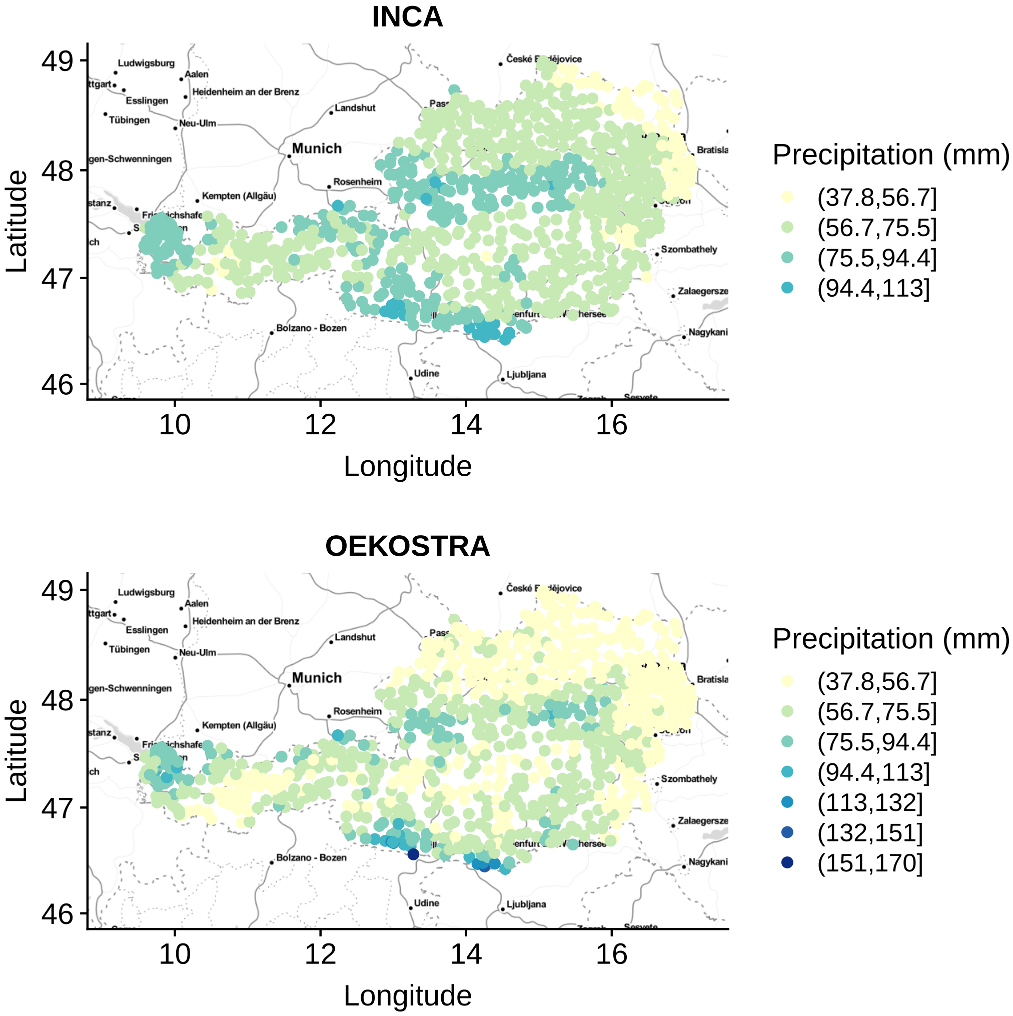


Figure 3.1.: OEKOSTRA and INCA maps.

```
Oekostra      5.855e-01  1.680e-02    34.851    <2e-16
Height        1.331e-03  5.819e-04    2.288     0.0224

Residual standard error: 6.883 on 914 degrees of freedom
Multiple R-squared:  0.6122,Adjusted R-squared:  0.6114
F-statistic: 721.5 on 2 and 914 DF,  p-value: < 2.2e-16
```

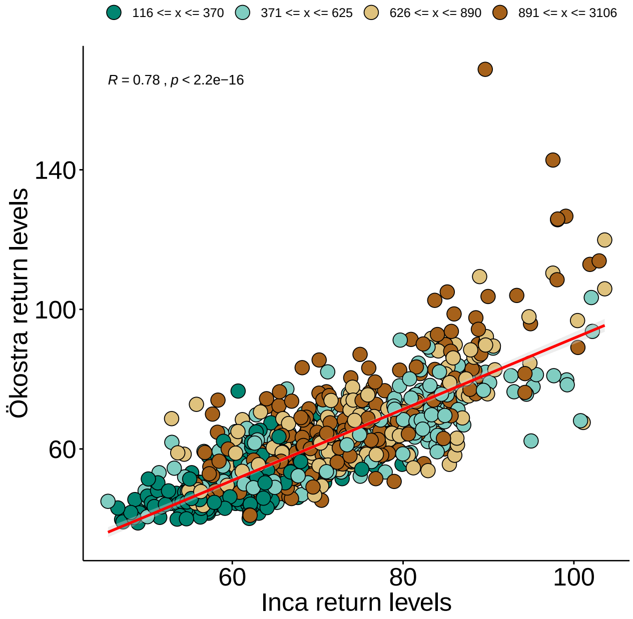


Figure 3.2.: Correlation (R) between `OEKOSTRA` and `INCA` extremal levels.

which does not suggest a very high impact of the height variable on our model, but it is still significant at the usual 5% confidence level.

Finally, we considered if there might be an interaction between the two regressor variables. This gives rise to the model

$$\widehat{\texttt{INCA}} = -0.0003h \cdot \texttt{OEKOSTRA} + 0.02h + 0.836 \cdot \texttt{OEKOSTRA} + 19.0722, \qquad (3.3)$$

and yields the corresponding regression table:

```
                    Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)         1.907e+01  1.913e+00    9.968      <2e-16
Oekostra            8.364e-01  3.311e-02    25.258     <2e-16
Height              2.012e-02  2.236e-03    9.000      <2e-16
Oekostra*Height    -3.157e-04  3.637e-05   -8.681      <2e-16

Residual standard error: 6.619 on 913 degrees of freedom
Multiple R-squared:  0.6418,Adjusted R-squared:  0.6406
F-statistic: 545.2 on 3 and 913 DF,  p-value: < 2.2e-16
```

The outcome of the regression analysis suggests clear evidence for an interaction effect. The R^2 increased quite notably, suggesting that this model does a significantly better job as compared to the model which doesn't use the variable `Height` as covariate.

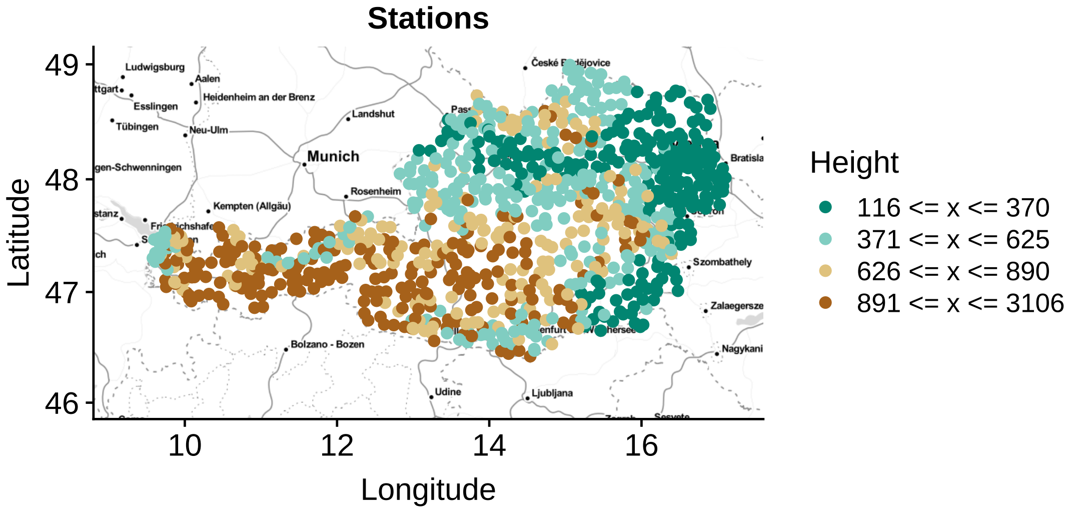


Figure 3.3.: Sea-levels of measuring stations.

We noticed among the `OEKOSTRA` data two outliers (station 114637 and 113811) with very hight precipitation levels (Figure 3.1) and hence also considered to remove those. Their overall leverage on the regression, however, is rather small.

3.2. A smoothing approach

In this and in the subsequent sections we are trying to remove the effect of noisy data, by pooling and averaging data with respect to geographical or topological features.

In the first approach we calculate the median precipitation value over nearby stations which are in the range of some given radius δ. The goal is to reduce noise by profiting from spatial correlation. Since Austria only spans a few degrees in longitude and even fewer degrees in latitude, said radius had to be quite small for a meaningful calculation. A station x_i is considered to be within a radius δ of a location s if

$$\sqrt{(\mathrm{lat}(x_i) - \mathrm{lat}(s))^2 + (\mathrm{lon}(x_i) - \mathrm{lon}(s))^2} \leq \delta.$$

As expected, the correlation-coefficient between the pooled `INCA` and `OEKOSTRA` data rises. The maximal correlation is achieved at $\delta = 0.25$. See Figure 3.4.

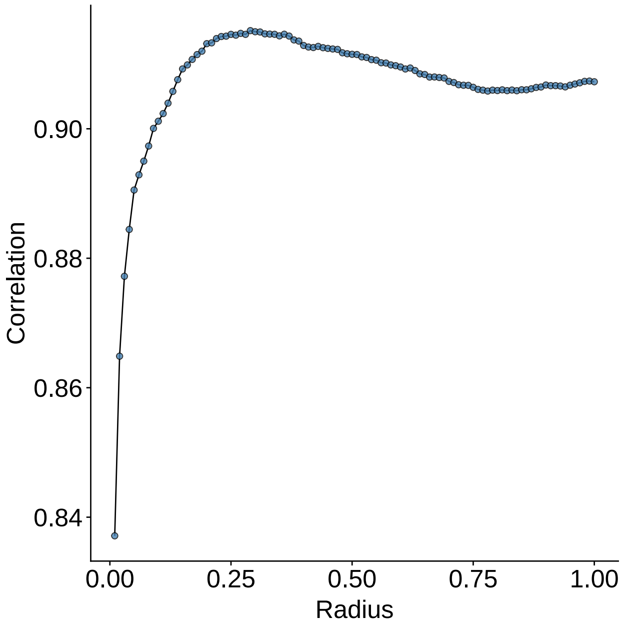


Figure 3.4.: Correlation (R) between `OEKOSTRA` and `INCA` with data pooling by radius.

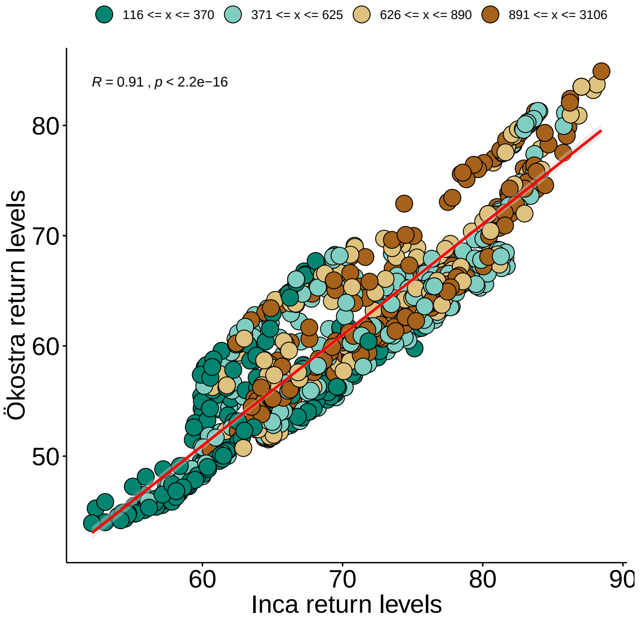


Figure 3.5.: Correlation (R) between `OEKOSTRA` and `INCA` smoothed with radius 0.25.

When fitting a linear model we obtain

$$\widehat{\mathtt{INCA}} = 0.833 \cdot \mathtt{OEKOSTRA} + 19.276. \tag{3.4}$$

The regression table for $\delta = 0.25$ indicates a very good fit:

```
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   19.27575   0.75494      25.53     <2e-16
Oekostra      0.83257    0.01217      68.41     <2e-16

Residual standard error: 3.277 on 915 degrees of freedom
Multiple R-squared:  0.8365,Adjusted R-squared:  0.8363
F-statistic:  4680 on 1 and 915 DF,  p-value: < 2.2e-16
```

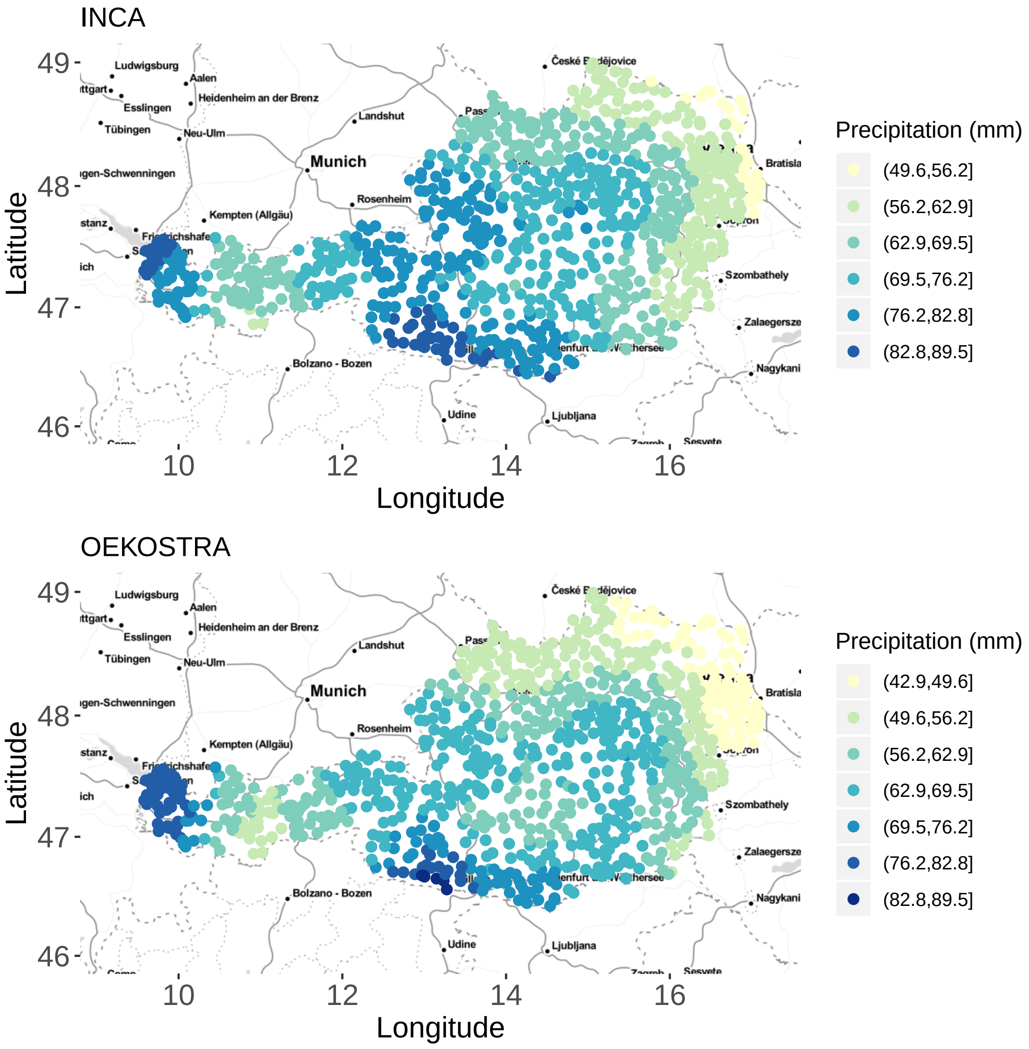


Figure 3.6.: Scatterplot between OEKOSTRA and INCA maps smoothed with radius $\delta = 0.25$.

3.3. Grouping within a range of sea-levels

Our next approach is based on grouping different data according to a range of sea-levels and then calculate the mean of the extreme precipitation levels within those groups. Grouping the stations into similar height levels might improve the correlation since sea-levels have an impact on precipitation

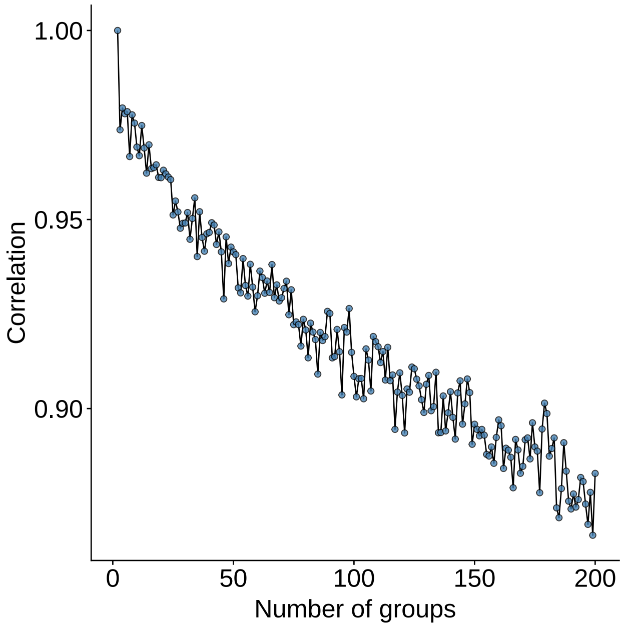


Figure 3.7.: Correlation (R) between OEKOSTRA and INCA extreme precipitation levels with the height method.

amounts.

To this end we first sorted the data with respect to ascending sea-levels and then divided them into $k > 0$ groups. The first group contains the $m := \lfloor n/k \rfloor$ stations with the lowest sea-level, the second group also contains $\lfloor n/k \rfloor$ stations with the lowest sea-level among the remaining stations, etc. We do this for both, OEKOSTRA and INCA data. For both data, we then compute m corresponding means and compare them as in the previous sections.

Looking at groups where many stations are combined into one group indeed yields quite strong correlation. And, not surprisingly, this correlation declines as the number of groups grows, i.e. when we tend towards the raw data, where each station forms one group. (See Figure 3.7). In Figure 3.8 we grouped data into 50 height levels. Notice that the group of biggest altitudes are all found below the regression line which suggests that INCA might systematically overestimate precipitation on higher altitude.

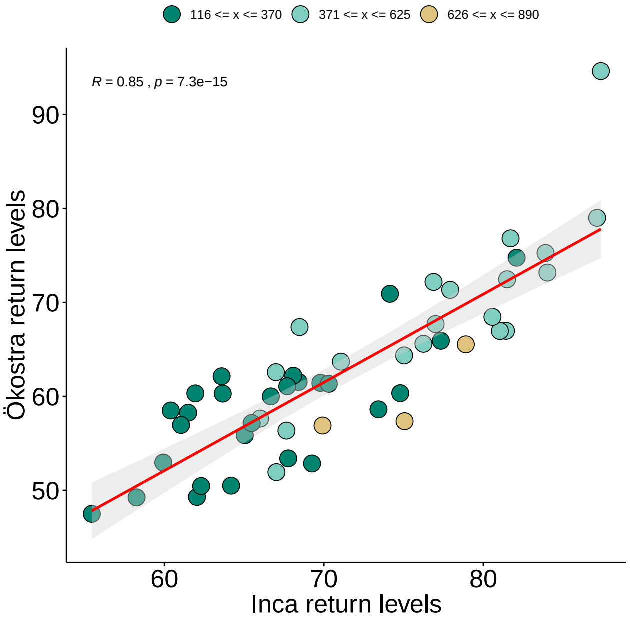


Figure 3.8.: Correlation (R) between pooled OEKOSTRA and INCA with 50 sea-levels. Note that the height levels indicating the color code refer to the minimal height of stations with a group.

3.4. Grouping with a rectangular grid

In this approach we divide Austria into a regular $k \times k$ grid and calculate the median of precipitations within the grid segments, in order to capture local effects and even out the effect of statistical outliers. Note that in some segments we may have no stations and therefore the assigned value is 0 or NA, respectively. We selected from 1x1 to 20x20 and 50x50 and 100x100 segments. It is obvious that the greater our grid size, the more empty segments we obtain. In the 100x100 case the map is already filled with around 90% of empty segments.

We had a closer look at the gridsize 19x19 since the correlation locally spiked there. Presumably, this is the ideal size to capture local climate conditions in specific areas.

A regression here yields similar results as in previous approaches. We obtain

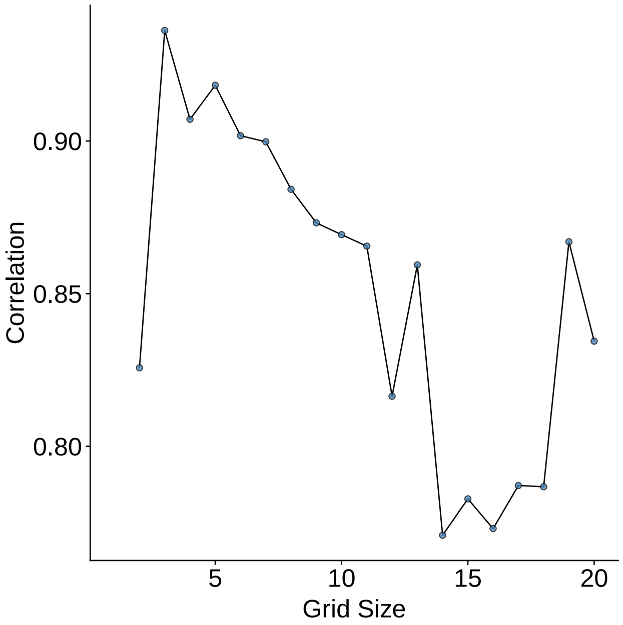


Figure 3.9.: Correlation (R) between `OEKOSTRA` and `INCA` with the grid method.

the linear model

$$\widehat{\texttt{INCA}} = 0.732 \cdot \texttt{OEKOSTRA} + 25.283, \tag{3.5}$$

with

```
              Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   25.28299    1.81030       13.97      <2e-16
Oekostra      0.73152     0.02894       25.27      <2e-16

Residual standard error: 3.277 on 915 degrees of freedom
Multiple R-squared:  0.8365,Adjusted R-squared:  0.8363
F-statistic:  4680 on 1 and 915 DF,  p-value: < 2.2e-16
```

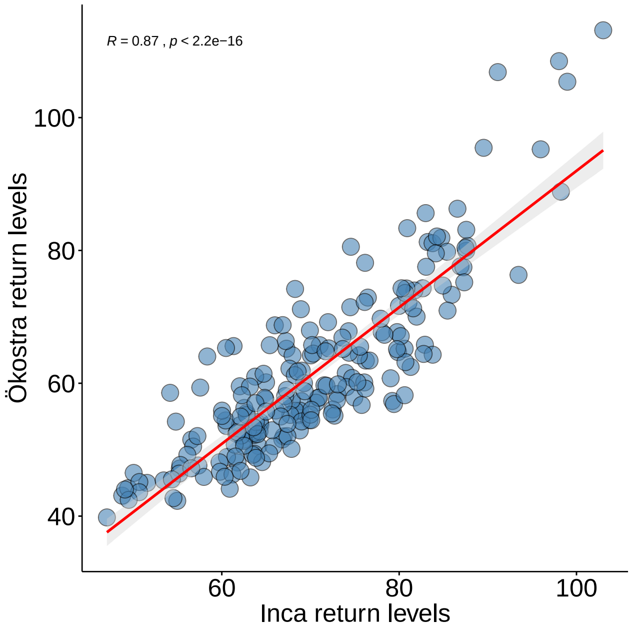


Figure 3.10.: Correlation with gridsize 19.

3.5. Measuring stations with intraday data

Previously we were considering a return period $r = 2$ years and a duration $d = 24$h. For this setup we had extreme precipitation levels available for each of the 917 measuring stations in Austria. In the next section we want to compare our current findings to shorter durations like $d = 3$h. Then, when intraday data are needed, we can base our analysis only 221 stations.

Before we do a similar analysis as above with the shorter durations we first want to see if the measuring stations which provide intraday data, behave similar as those with only daily data. To this end we provide a scatterplot of the precipitation levels where we mark the 221 stations providing intraday-data. See Figure 3.12.

After repeating our calculations with a return period of 2 years and a duration of 24h for only the 221 stations with complete datasets we obtained pretty similar results with all methods. The correlations were in tendency slightly higher. A possible reason for this might be that a station with

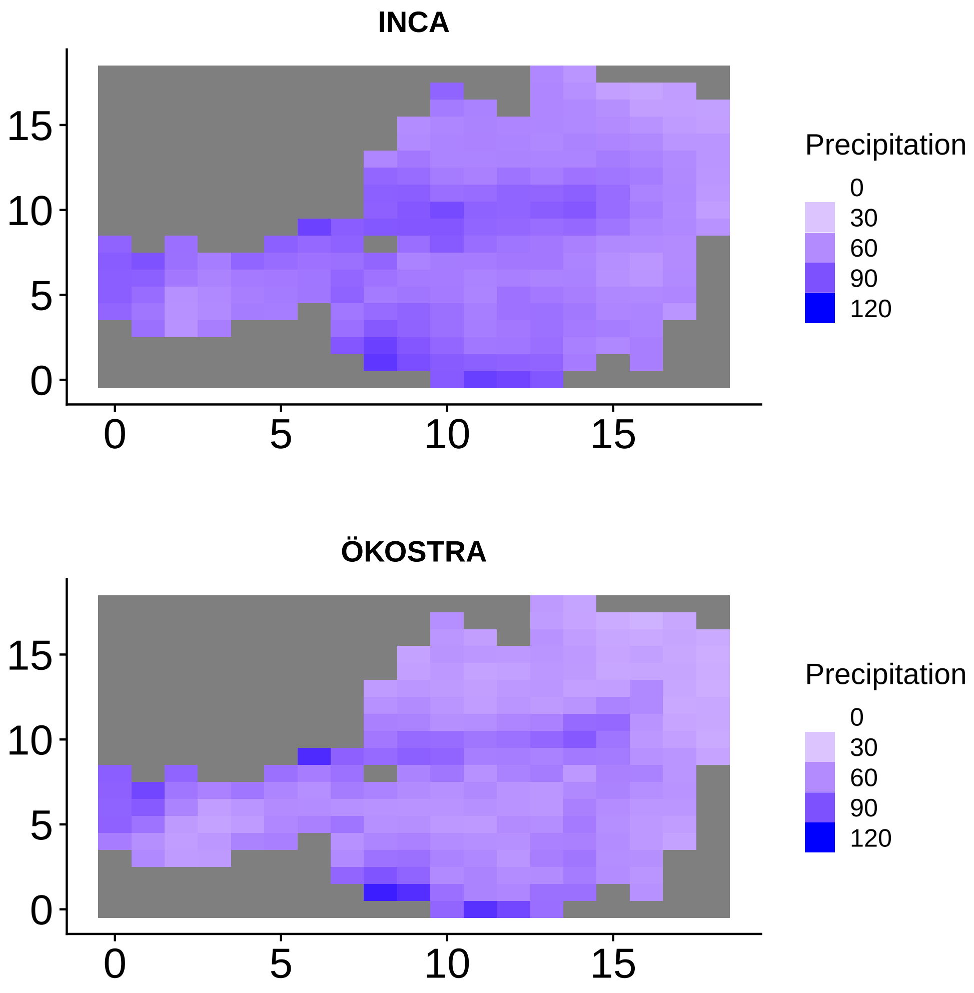


Figure 3.11.: INCA plot with gridsize 19.

complete datasets has to be monitored more carefully than a station which only collects data once a day and it might also be better equipped.

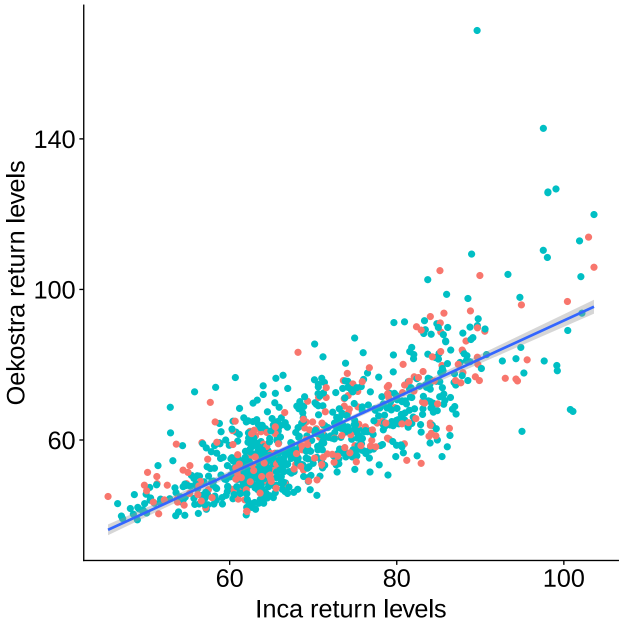


Figure 3.12.: Distinguishing return levels (2 years, 24 hours duration), where red indicates intraday data is available and blue indicates no intraday data is available.

3.6. Analysis of shorter durations

Now we repeat our calculations for shorter durations. We kept the return period of 2 years. The initial problem originated from the comparison of the two maps for a return period of 2 years and a duration of 3h (Figure 1.1) and that is why we choose to analyse the 3h duration.

There are again similarities in both plots but not as pronounced as in the $d = 24$h case. It seems that different hight levels are correlated differently, meaning that the regression line would have slopes depending on the sealevel. A weaker linear relationship is suggested by a scatterplot (Figure 3.15). The estimated linear model is given by the equation

$$\widehat{\texttt{INCA}} = 0.403 \cdot \texttt{OEKOSTRA} + 19.579. \tag{3.6}$$

The regression table looks as follows:

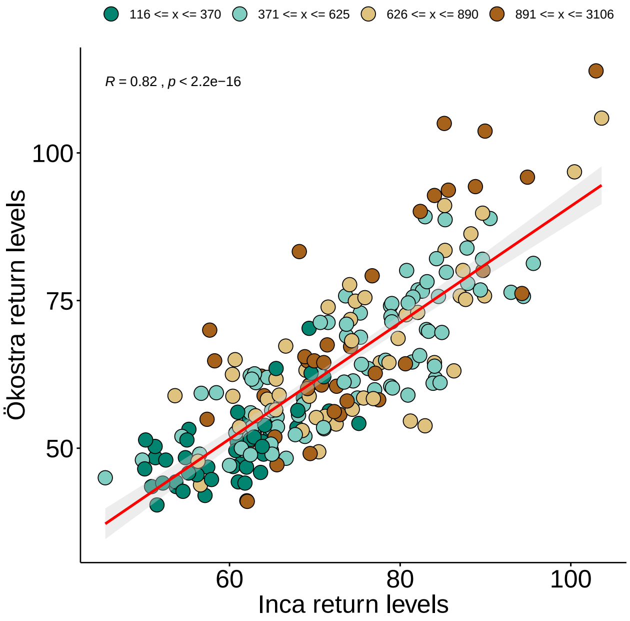


Figure 3.13.: Correlation (R) for the stations with intraday return levels.

```
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  19.57879   1.52994      12.797    <2e-16
Oekostra     0.40281    0.04893      8.232     1.65e-14

Residual standard error: 3.7 on 219 degrees of freedom
Multiple R-squared:  0.2363,Adjusted R-squared:  0.2328
F-statistic: 67.77 on 1 and 219 DF,  p-value: 1.646e-14.
```

The linear relationship is weaker than before and the correlation coefficient unsurprisingly dropped quite a bit (Figure 3.14). We have repeated our other approaches in this case and obtained very similar results. Let us point out that the sea-level method provided most convincing results. The difference in the baseline-correlation without grouping ($R = 0.49$) and the correlations from the sea-level method is quite significant (Figure 3.16).

The regression with height as extra explanatory variable, however, showed no significant impact and from the analysis without intercept we can confirm that `OEKOSTRA` precipitation levels are in tendency again slightly smaller.

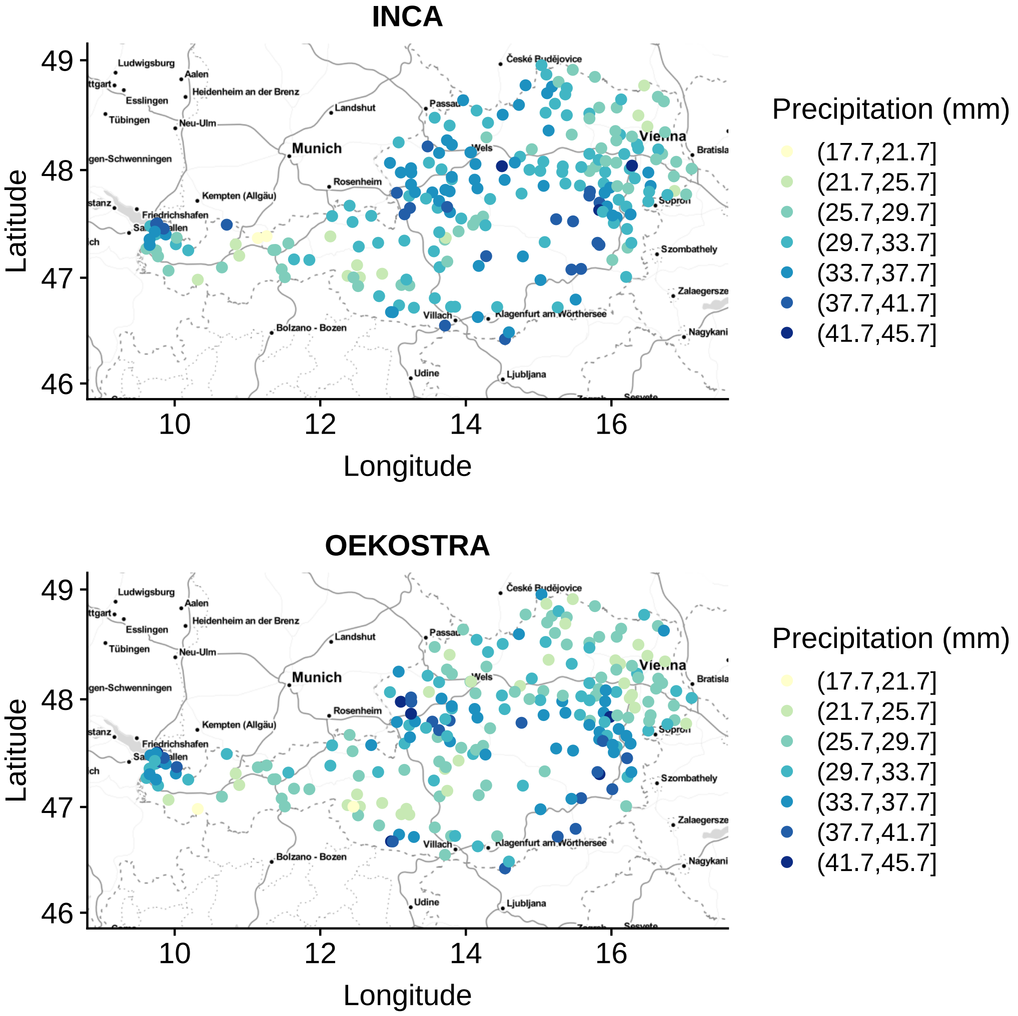


Figure 3.14.: Return levels for 3 hour duration.

$$\widehat{\texttt{INCA}} = -0.002h + 0.398 \cdot \texttt{OEKOSTRA} + 20.702, \tag{3.7}$$

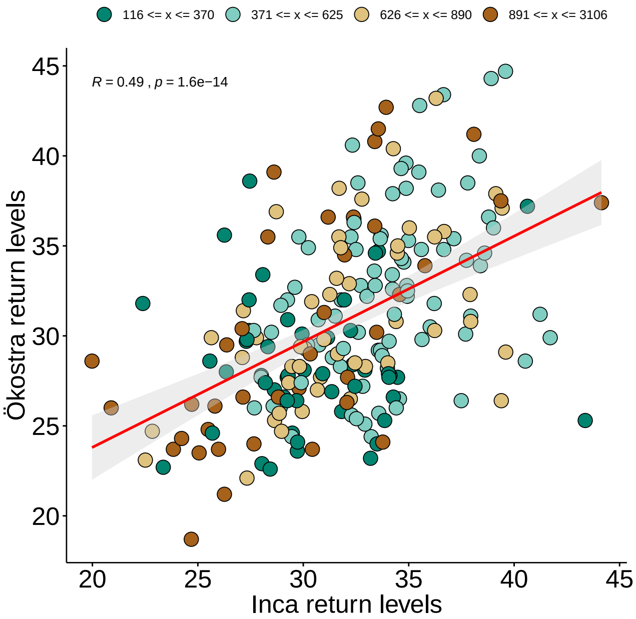


Figure 3.15.: Correlation (R) for 3 hour duration.

This leads to the following regression table:

```
                Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)     20.7019816 1.5940835    12.987     < 2e-16
Oekostra        0.3979751  0.0485172    8.203      2.02e-14
Height          -0.0015751 0.0006929    -2.273     0.024

Residual standard error: 3.666 on 218 degrees of freedom
Multiple R-squared:  0.254,Adjusted R-squared:  0.2472
F-statistic: 37.12 on 2 and 218 DF,  p-value: 1.341e-14
```

Lastly, in view of Figure 3.16, we considered a regression with an interaction between the sealevel and our response variable. The estimated model is now give as

$$\widehat{\mathtt{INCA}} = 0.0005h \cdot \mathtt{OEKOSTRA} - 0.004h + 0.56 \cdot \mathtt{OEKOSTRA} + 46.752. \qquad (3.8)$$

The regression table is summarized below.

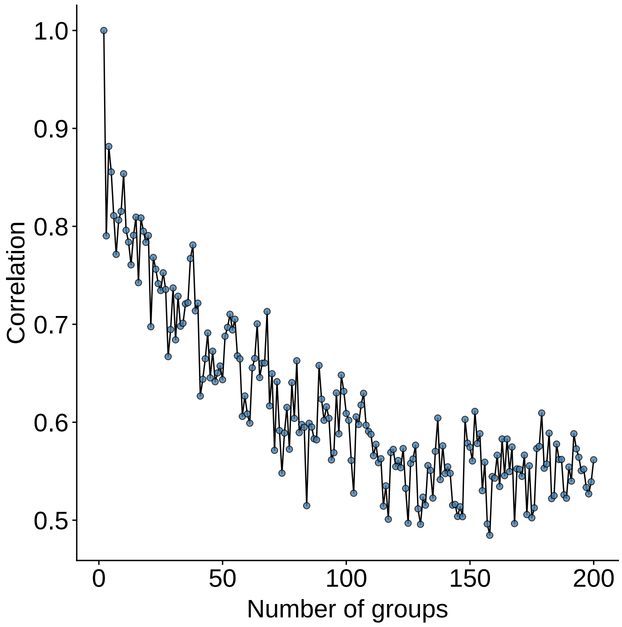


Figure 3.16.: Correlation with the height method for duration of 3 hours.

```
                  Estimate     Std. Error   t value    Pr(>|t|)
  (Intercept)     46.7518691   8.2956463    5.636      5.38e-08
  Oekostra        0.5613296    0.2826970    1.986      0.0121
  Height          -0.0035898   0.0102594   -0.350      <2e-16
  Oekostra*Height 0.0004862    0.0003623    1.342      <2e-16

  Residual standard error: 10.15 on 217 degrees of freedom
Multiple R-squared:  0.2623,Adjusted R-squared:  0.259
F-statistic:  23.4 on 3 and 217 DF,  p-value: 3.659e-13
```

The regression here suggests that in fact the interaction with the sealevel is significant.

4. Theory

4.1. Introduction and basic results

The following chapter about extreme value theory (EVT) is based on Haan and Ferreira, 2006. We will describe the basics of EVT and point out some interesting facts that relate to our problem and its solution.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables. The theory of extreme values is concerned with the behavior of the samples extremes:

$$M_n := \max(X_1, X_2, \ldots, X_n) \quad n \to \infty. \tag{4.1}$$

Since the minima formulation can easily be transformed into maxima formulation we will restrict ourselves to the latter.

As an example let us consider precipitation values. We are interested to find when rainfall exceeds a certain threshold, for example the amount of water a dam can hold. We can think of extreme rainfall as an extreme observation that causes failure.

We are interested in finding a limit distribution for maxima of a random sequence. Assuming that these variables are identically and independently distributed is in many cases oversimplistic, but will be made for the sake of simplicity. Let F be the underlying distribution function and $x^* = \sup\{x : F(x) < 1\}$. We first observe that M_n converges to x^* in probability. This is because

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, \ldots, X_n \leq x) \\ = F^n(x) &\to \begin{cases} 0 & x < x^*, \\ 1 & x \geq x^*. \end{cases} \end{aligned}$$

This shows that the limiting distribution function is degenerate. In order to obtain a nondegenerate limit distribution, a normalisation is necessary. Hence, we consider

$$\frac{M_n - b_n}{a_n}$$

with $a_n > 0$ and $b_n \in \mathbb{R}$.

Suppose that the standardised maximum has a limit distribution $G(x)$, i.e.

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{4.2}$$

for every continuity point x of G and G a nondegenerate distribution function. We call every distribution function G that satisfies the expression (4.2) an **extreme value distribution**. Our main aim at this stage will be to find the possible limiting distributions.

Let us first reformulate (4.2) by taking logarithms left and right. We obtain for each continuity point x for which $0 < G(x) < 1$ holds that

$$\lim_{n\to\infty} n \log F(a_n x + b_n) = \log G(x). \tag{4.3}$$

It follows that $F(a_n x + b_n) \to 1$ for each x, and therefore we get

$$\lim_{n\to\infty} \frac{-\log F(a_n x + b_n)}{1 - F(a_n x + b_n)} = 1,$$

since

$$\lim_{\epsilon\to 0} \frac{-\log(1-\epsilon)}{\epsilon} = 1.$$

Thus

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -\log G(x),$$

or equivalently

$$\lim_{n\to\infty} \frac{1}{n(1 - F(a_n x + b_n))} = \frac{1}{-\log G(x)}. \tag{4.4}$$

To proceed further, we need to introduce generalised inverse functions.

Definition 1. *Let f be any nondecreasing function. We call $f^{\leftarrow}$ the **left continuous inverse** if*

$$f^{\leftarrow}(x) := \inf\{y : f(y) \geq x\}. \tag{4.5}$$

With the following lemma, whose proof can be found in Haan and Ferreira, 2006, page 5, we can see why inverse functions are useful here.

Lemma 1. *Suppose f_n is a sequence of nondecreasing functions and g is a nondecreasing function. Suppose that for each x in some open interval (a, b) which is a continuity point of g,*

$$\lim_{n\to\infty} f_n(x) = g(x). \tag{4.6}$$

Let $f_n^{\leftarrow}, g^{\leftarrow}$ be the left continuous inverses of f_n and g. Then, for each x in the interval $(g(a), g(b))$ which is a continuity point of $g^{\leftarrow}$ we also have

$$\lim_{n\to\infty} f_n^{\leftarrow}(x) = g^{\leftarrow}(x). \tag{4.7}$$

Now observe that the left continuous inverse of the right hand side of (4.4) can be expressed as

$$\inf\left\{y : \frac{1}{-\log G(y)} \geq x\right\} = \inf\left\{y : x \leq \log G(y)\right\} =$$
$$\inf\left\{y : \frac{1}{\log G(y)} \leq \frac{1}{x}\right\} = \inf\left\{y : -\log G(y) \leq \frac{1}{x}\right\} =$$
$$\inf\left\{y : \log G(y) \geq -\frac{1}{x}\right\} = \inf\left\{y : G(y) \geq e^{-1/x}\right\} = G^{\leftarrow}(e^{-1/x}).$$

We can now apply Lemma 1 to (4.4) by selecting

$$U(t) = F^{\leftarrow}\left(\frac{1}{1-t}\right), \quad t > 1, \tag{4.8}$$

and considering

$$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = G^{\leftarrow}(e^{-1/x}) =: D(x), \quad x > 0. \tag{4.9}$$

If we assume that x is a continuity point of D, then for $t \geq 1$

$$\frac{U([t]x) - b_{[t]}}{a_{[t]}} \leq \frac{U(tx) - b_{[t]}}{a_{[t]}} \leq \frac{U([t]x(1+1/[t])) - b_{[t]}}{a_{[t]}} \leq D(x'), \quad (4.10)$$

for $x' > x$ with $D(x') > D(x)$. Since D is continuous at x, we obtain

$$\lim_{t\to\infty} \frac{U(tx) - b_{[t]}}{a_{[t]}} = D(x). \quad (4.11)$$

With our previous observation we can formulate the following theorem that yields useful alternative formulations of our initial condition on the extreme value distribution (4.2).

Theorem 1. *Let $a_n > 0$ and b_n be real sequences of constants and G a nondegenerate distribution function. The following statements are equivalent*

1.
$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$$
for each continuity point x of $G(x)$

2.
$$\lim_{t\to\infty} t(1 - F(a(t)x + b(t))) = -\log G(x) \quad (4.12)$$
for each continuity point x of G for which $0 < G(x) < 1$, $a(t) := a_{[t]}$, and $b(t) := b_{[t]}$ (with $[t]$ the integer part of t)

3.
$$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D(x), \quad (4.13)$$
for each continuity point x of $D(x) = G^{\leftarrow}(e^{-1/x})$, $a(t) := a_{[t]}$, and $b(t) := b_{[t]}$.

4.2. Extreme value distributions

We will now try to identify a class of distributions that can occur as a limit of the relation (4.2), discussed earlier.

Theorem 2 (Fisher and Tippet (1928), Gnedenko (1943))**.** *The class of extreme value distributions is of parametric form $G_\gamma(ax+b)$ with $a > 0, b \in \mathbb{R}$, where*

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}) & 1+\gamma x > 0 \\ \exp(-e^{-x}) & \gamma = 0 \end{cases} \tag{4.14}$$

with $\gamma \in \mathbb{R}$.

Definition 2. *The parameter γ in (4.14) is called the* ***extreme value index****.*

Proof. Let us consider the class of limit functions D in (4.13). First suppose that 1 is a continuity point of D. Then note that for continuity points $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = D(x) - D(1) =: E(x). \tag{4.15}$$

Take $y > 0$ and write

$$\frac{U(txy)-U(t)}{a(t)} = \frac{U(txy)-U(ty)}{a(ty)}\frac{a(ty)}{a(t)} + \frac{U(tx)-U(t)}{a(t)}. \tag{4.16}$$

We claim that $\lim_{t\to\infty}(U(ty) - U(t)/a(t))$ and $\lim_{t\to\infty} a(ty)/a(t))$ exist. Suppose not. Then there are A_1, A_2, B_1, B_2 with $A_1 \neq A_2$ or $B_1 \neq B_2$, where B_i are limit points of $(U(ty) - U(t)/a(t))$ and A_i are limit points of $a(ty)/a(t))$, $i = 1, 2$, as $t \to \infty$. We find from (4.16) that

$$E(xy) = E(x)A_i + B_i \tag{4.17}$$

$i = 1, 2$, for all continuity points x of $E(\cdot)$ and $E(\cdot y)$. For an arbitrary x take a sequence of continuity points x_n with $x_n \uparrow x, n \to \infty$. Then $E(x_n y) \to E(xy)$ and $E(x_n) \to E(x)$ since E is left continuous. Hence (4.17) holds for all $x > 0$ and $y > 0$. Subtracting the expressions for $i = 1, 2$ from each other one obtains

$$E(x)(A_1 - A_2) = B_2 - B_1 \tag{4.18}$$

for all $x > 0$. Since E cannot be constant (because G is nondegenerate) we must have $A_1 = A_2$ and hence also $B_1 = B_2$. Therefore

$$A(y) := \lim_{t\to\infty} \frac{a(ty)}{a(t)} \tag{4.19}$$

exists for $y > 0$ and for $x, y > 0$,

$$E(xy) = E(x)A(y) + E(y).$$

Hence for $s := \log x$, $t := \log y$ $(x, y \neq 1)$ and $H(x) := E(x^x)$, we have

$$H(t+s) = H(s)A(e^t) + H(t), \tag{4.20}$$

which, since $H(0) = 0$, we can write as

$$\frac{H(t+s) - H(t)}{s} = \frac{H(s) - H(0)}{s} A(e^t). \tag{4.21}$$

Since H is monotone, there is certainly one t at which H is differentiable and therefore by (4.21) H is differentiable everywhere and

$$H'(t) = H'(0)A(e^t). \tag{4.22}$$

Write $Q(t) := H(t)/H'(0)$. Note that $H'(0)$ cannot be zero: H cannot be constant since G is nondegenerate. Then $Q(0) = 0, Q'(0) = 1$. By (4.20)

$$Q(t+s) - Q(t) = Q(s)A(e^t),$$

and by (4.21),

$$Q(t+s) - Q(t) = Q(s)Q'(t).$$

Subtracting the same expressions with t and s interchanged we get

$$Q(t)\frac{Q'(s) - 1}{s} = \frac{Q(s)}{s}(Q'(t) - 1),$$

hence (letting $s \to 0$)

$$Q(t)Q''(0) = Q'(t) - 1.$$

It follows that Q is twice differentiable, and by differentiation,

$$Q''(0)Q'(t) = Q''(t).$$

Hence

$$(\log Q')'(t) = Q''(0) =: \gamma \in \mathbb{R},$$

for all t. It follows that (note that $Q'(0) = 1$)

$$Q'(t) = e^{\gamma t}$$

and (since $Q(0) = 0$)

$$Q(t) = \int_0^t e^{\gamma s} ds.$$

This means that

$$H(t) = H'(0)\frac{e^{\gamma t} - 1}{\gamma}$$

and

$$D(t) = D(1) + H'(0)\frac{t^{\gamma} - 1}{\gamma}.$$

Hence

$$D^{\leftarrow}(x) = \left(1 + \gamma\frac{x - D(1)}{H'(0)}\right)^{1/\gamma}. \tag{4.23}$$

Now $D(x) = G^{\leftarrow}(e^{-1/x})$, and hence

$$D^{\leftarrow}(x) = \frac{1}{-\log G(x)}. \tag{4.24}$$

Combining (4.23) and (4.24), we obtain the statement of the theorem. If 1 is not a continuity point of D, follow the proof with the function $U(tx_0)$ with x_0 a continuity point of D. □

Remark 1. *This result shows that the limit distribution functions form a simple explicit one-parameter (γ) family apart from the scale (a_n) and location (b_n) parameters. It also shows that it contains subclasses with quite different features. Consider the subclasses $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$*

1. *For $\gamma > 0$ clearly $G_{\gamma}(x) < 1$ for all x and therefore the right endpoint of the distribution is infinity. Moreover*

$$1 - G_{\gamma}(x) \sim \gamma^{-1/\gamma} x^{-1/\gamma}, \quad x \to \infty.$$

This means that the distribution has a heavy right tail. Furthermore using $G_{\gamma}((x-1)/\gamma)$ and with $\alpha = 1/\gamma > 0$ we obtain another parametrization as follows:

$$\Phi_{\alpha}(x) := \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases} \tag{4.25}$$

*This class is called the **Frechet** class of extreme value distributions.*

2. *For $\gamma = 0$ the right endpoint of the distribution equals infinity. The distribution however is rather light-tailed: $1 - G_0(x) \sim e^{-x}$ as $x \to \infty$. Another parametrisation in this case is obtained selecting*

$$G_0(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}. \tag{4.26}$$

*This class is called the **double-exponential** or **Gumbel distribution**.*

3. *For $\gamma < 0$ the right endpoint of the distribution is $-1/\gamma$ so it has a short tail, verifying $1 - G_\gamma(-\gamma^{-1} - x) \approx (-\gamma x)^{-1/\gamma}$, as $x \downarrow 0$. The parametrisation in this case is obtained by using $G_\gamma(-1(1+x)/\gamma)$ and $\alpha = -1/\gamma > 0$. Then*

$$\Psi_\alpha(x) := \begin{cases} \exp(-(-x)^\alpha) & x < 0 \\ 1 & x \geq 0 \end{cases} \tag{4.27}$$

*This class is called the **reverse-Weibull** class of distributions.*

This next definition is a way to characterize the domain in which (4.2) converges.

Definition 3. *Let F be a distribution function for which*

$$\lim_{n \to \infty} F^n(a_n x + b_n) = G_\gamma(x),$$

*holds for some $\gamma \in \mathbb{R}$. Then we say F is in the **domain of attraction** of G_γ. We denote this by $F \in \mathcal{D}(G_\gamma)$*

We conclude this section by a useful criteria to check if a distribution function is in the domain of G_γ.

Theorem 3. *Let F be a distribution function and x^* its right endpoint. Suppose $F''(x)$ exists and $F'(x)$ is positive for all x in some left neighborhood of x^*. If*

$$\lim_{t \uparrow x^*} \left(\frac{1-F}{F'} \right)'(t) = \gamma, \tag{4.28}$$

then F is in the domain of attraction of G_γ.

We do not give a proof here, but refer to Haan and Ferreira, 2006, page 15.

Until now we have considered the limiting distribution of the sample maximum. We will consider in the following some asymptotic results for other order statistics. Recall that the order statistics are defined as the ordered sample:

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n},$$

where $X_{i,n}$ is the i-th largest element of our sample $X_1, \dots, X_n$. Hence $M_n = X_{n,n}$. For fixed k we call $X_{k,n}$ and $X_{n-k,n}$ extreme order statistics. For $k \to \infty$ as $n \to \infty$ but $k/n \to 0$ we call $X_{k,n}$ intermediate order statistics.

4.3. Asymptotic distribution of extreme order statistics

From (4.13) and (4.14) we can obtain the relation for $x > 0$:

$$\begin{aligned} \lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} &= D_\gamma(x) = G_\gamma^{\leftarrow}(e^{-1/x}) = \\ \inf\{y : G_\gamma(y) \geq e^{-1/x}\} &= \frac{x^\gamma - 1}{\gamma}. \end{aligned} \tag{4.29}$$

The last equality is because $G_\gamma(\frac{x^\gamma - 1}{\gamma}) = e^{-1/x}$. With this condition we can obtain convergence in distribution for normalized sample maxima as in (4.2). Moreover, it can be used to obtain convergence for the extreme order statistics. We begin by outlining the important special case of exponential distributions.

Let $X_1, \dots, X_n$ be independent and identically distributed exponential random variables with parameter λ and $X_{1,n} \leq \cdots \leq X_{n,n}$ be the order statistics of our sample. Rényi, 1953 showed that

$$(X_{i,n} \colon 1 \leq i \leq n) \overset{d}{=} \left(\frac{1}{\lambda} \sum_{j=1}^{i} \frac{Z_j}{n - j + 1} \colon 1 \leq i \leq n \right), \tag{4.30}$$

where Z_i are independent and identically distributed standard exponential random variables.

For a fixed $k \leq n$ we can therefore deduce that

$$(X_{1,n}, \ldots, X_{k,n}) =^{d} \left(\frac{Z_1}{n}, \ldots, \frac{Z_1}{n} + \frac{Z_2}{n-1} + \cdots + \frac{Z_k}{n-k+1} \right). \qquad (4.31)$$

Hence

$$n(X_{1,n}, \ldots, X_{k,n}) \xrightarrow{d} (Z_1, \ldots, Z_1 + \cdots + Z_k), \qquad (4.32)$$

which suggests that the normalised lower extreme-order statistics converges to a homogeneous Poisson process.

This result can be extended to the high extreme-order statistics as well, with the assumption of exponential variables dropped.

Theorem 4. *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function F. Suppose F is in the domain of attraction of G_γ for some $\gamma \in \mathbb{R}$. Let $X_{1,n} \leq \cdots \leq X_{n,n}$ be the order statistics. Then with the normalizing constants $a_n > 0$ and $b_n \in \mathbb{R}$ from (4.2) and fixed $k \in \mathbb{N}$*

$$\left(\frac{X_{n-j,n} - b_n}{a_n} : j = 0, \ldots, k \right) \xrightarrow{d} \left(\frac{(Z_1 + \cdots + Z_{j+1})^{-\gamma} - 1}{\gamma} : j = 0, \ldots, k \right),$$

where $Z_1, \ldots, Z_k$ are independent and identically distributed standard exponential random variables.

Proof. Note that if Z is a random variable with standard exponential distribution, then

$$U\left(\frac{1}{1 - e^{-Z}} \right)$$

has distribution function F. Hence

$$(X_{n,n}, \ldots, X_{n-k+1,n}) \overset{d}{=} \left(U\left(\frac{1}{1 - e^{-Z_{1,n}}} \right), \ldots, U\left(\frac{1}{1 - e^{-Z_{k,n}}} \right) \right).$$

Next note that

$$\lim_{n \to \infty} \frac{U\left(\frac{1}{1-e^{x/n}}\right) - b_n}{a_n} = \lim_{n \to \infty} \frac{U\left(\frac{n}{n(1-e^{x/n})}\right) - U(n)}{a_n} =$$
$$\frac{(\lim_{n\to\infty} n(1 - e^{-x/n}))^{-\gamma} - 1}{\gamma} = \frac{x^{-\gamma} - 1}{\gamma}.$$

And lastly by applying (4.29) and (4.30) and the fact that for $x \geq 0$

$$\lim_{n\to\infty} n\left(1 - e^{-x/n}\right) = x,$$

we can conclude the proof. □

4.4. Intermediate Order Statistics

Extreme value theory, is about understanding the tail behaviour of the distribution underlying the data. If we want to get a better understanding of the extremal behaviour of our data, we should not solely focus on the maximum. Rather we should incorporate the information of other large order statistics, as this also will contain valuable information about the tail. In particular, if the sample grows, we should take into account more and more large values in our sample. This gives rise to the following definition.

Definition 4. *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. The order statistics $X_{n-k,n}$ with $n \to \infty$, $k = k(n) \to \infty$ and $k(n)/n \to 0$ are then called **intermediate order statistics**.*

The following result shows how the intermediate order statistics are asymptotically behaving.

Theorem 5. *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function F. Recall that $U = (1/(1-F))^{\leftarrow}$. Suppose (4.28) holds for an extreme value distribution G_γ. Then, if $k = k(n) \to \infty$ $k(n)/n \to 0$ as $n \to \infty$,*

$$\sqrt{k}\frac{X_{n-k,n} - U(\frac{n}{k})}{\frac{n}{k}U'(\frac{n}{k})} \xrightarrow{d} N(0,1).$$

The proof can again be found in Haan and Ferreira, 2006, page 41.

4.5. Estimation of the extreme value index

Given a sample $X_1, \dots, X_n$ we would like to estimate the corresponding extreme value index γ. In this section we show different approaches.

We will next focus on the formulation (4.13) for the purpose of statistical analysis in the context of extreme value theory. We have a positive function a such that for all $x \geq 0$

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}. \tag{4.33}$$

We will next have a look at different estimators for the extreme value index γ.

4.5.1. The Hill Estimator

The Hill estimator is one of the earliest established estimators for the extreme value index γ. It serves also as the basis for other estimators of γ and is therefore particularly important. We need the following result.

Theorem 6. *The distribution function F is in the domain of attraction of the extreme value distribution G_γ with $\gamma > 0$ if and only if $x^* = \sup\{x : F(x) < 1\}$ is infinite and*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}, x > 0 \tag{4.34}$$

The proof of Theorem 6 requires a rather large foundation about the theory revolving around domains of attraction and will therefore be omitted here. It can be found in Haan and Ferreira, 2006, page 19.

From (4.34) we know that

$$\frac{1 - F(te)}{1 - F(t)} \leq e^{\epsilon - 1/\gamma},$$

for any $\epsilon > 0$ and sufficiently large t. Therefore

$$\frac{1-F(te^n)}{1-F(t)} = \prod_{k=1}^{n} \frac{1-F(te^k)}{1-F(te^{k-1})} \leq e^{(\epsilon-1/\gamma)n},$$

and hence for all $x > 1$ we obtain

$$\begin{aligned} \frac{1-F(tx)}{1-F(t)} &\leq \frac{1-F(te^{[\log x]})}{1-F(t)} &&\leq e^{(\epsilon-1/\gamma)[\log x]} \\ &\leq e^{(\epsilon-1/\gamma)(1+\log(x))} &&= e^{-1/\gamma+\epsilon}x^{-1\gamma+\epsilon}. \end{aligned}$$

Applying the theorem of dominated convergence on (4.34) we then have that

$$\lim_{t\to\infty} \int_1^\infty \frac{1-F(tx)}{1-F(t)} \frac{dx}{x} = \int_1^\infty x^{-1/\gamma} \frac{dx}{x} = \gamma,$$

or

$$\lim_{t\to\infty} \frac{\int_t^\infty (1-F(tx)) \frac{dx}{x}}{1-F(t)} = \gamma.$$

Now partial integration yields

$$\int_t^\infty (1-F(u)) \frac{du}{u} = \int_t^\infty (\log v - \log t) dF(v).$$

Therefore we have

$$\lim_{t\to\infty} \frac{\int_t^\infty (\log v - \log t) dF(v)}{1-F(t)} = \gamma. \tag{4.35}$$

This asymptotic result motivates the following estimator for γ.

Definition 5 (Hill, 1975). *Let $X_{n-k,n}$ be a intermediate order statistic with $k \leq n$ and F_n the empirical distribution function of F. The **Hill** estimator $\hat{\gamma}_H$ is then defined as*

$$\hat{\gamma}_H := \frac{\int_{X_{n-k,n}}^\infty (\log v - \log X_{n-k,n}) dF_n(v)}{1-F_n(X_{n-k,n})}, \tag{4.36}$$

or

$$\hat{\gamma}_H := \frac{1}{k} \sum_{i=0}^{k-1} \log X_{n-i,n} - \log X_{n-k,n}. \tag{4.37}$$

We will next show that the Hill estimator has some nice asymptotic properties.

Theorem 7. *Let $X_1, \dots, X_n$ be independent identically distributed random variables with distribution function F. Suppose $F \in \mathcal{D}(G_\gamma)$ with $\gamma > 0$. Then as $n \to \infty$, $k = k(n) \to \infty$, $k/n \to 0$*

$$\hat{\gamma}_H \to^P \gamma \tag{4.38}$$

Proof. For this proof we will use Haan and Ferreira, 2006, Lemma 1.2.9 which states that

$$\lim_{t\to\infty} \frac{U(t)}{a(t)} = \frac{1}{\gamma}, \quad \gamma > 0. \tag{4.39}$$

If we combine (4.33) and (4.39) we see that

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma, \quad \gamma > 0, F \in \mathcal{D}(G_\gamma), \tag{4.40}$$

which implies

$$(1-\epsilon)x^{\gamma-\epsilon'} < \frac{U(tx)}{U(t)} < (1+\epsilon)^{\gamma+\epsilon'}, \tag{4.41}$$

and thus

$$\log(1-\epsilon) + (\gamma - \epsilon') \log x < \log U(tx) - \log U(t) \tag{4.42}$$
$$< \log(1+\epsilon) + (\gamma + \epsilon') \log x. \tag{4.43}$$

Let next $Y_1, Y_2, \dots$ be independent and identically distributed, with common distribution $1 - 1/y, y \geq 1$. Since $U(Y_i) =^d X_i$, it is enough to proof (4.38) for

$$\tilde{\gamma}_H := \frac{1}{k} \sum_{i=0}^{k-1} \log U(Y_{n-i,n}) - \log U(Y_{n-k,n}).$$

We can now set $t = Y_{n-k,n}$ and $x = Y_{n-i,n}/Y_{n-k,n}$ and use (4.42) and (4.43) to obtain

$$\log(1-\epsilon) + (\gamma-\epsilon') \log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right) < \log U(Y_{n-i,n}) - \log U(Y_{n-k,n})$$
$$< \log(1+\epsilon) + (\gamma+\epsilon') \log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right),$$

for $i = 0, \ldots, k-1$ and therefore

$$\log(1-\epsilon) + (\gamma - \epsilon')\frac{1}{k}\sum_{i=0}^{k-1}\log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right) < \tilde{\gamma}_H$$
$$< \log(1+\epsilon) + (\gamma + \epsilon')\frac{1}{k}\sum_{i=0}^{k-1}\log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right).$$

The proof then follows from Haan and Ferreira, 2006, Lemma 3.2.3 stating that

$$\frac{1}{k}\sum_{i=0}^{k-1}\log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right) \xrightarrow{d} 1. \tag{4.44}$$

□

4.5.2. A Moment Estimator for $\gamma \in \mathbb{R}$

The moment estimator we consider is an extension of the Hill estimator to a general γ. Notice that previously we required $\gamma > 0$. The Hill estimator was not defined for $\gamma \leq 0$, since in that case $U(\infty) \leq 0$ is possible, but then the logarithm is not defined. By shifting the sample data, it can be assumed that $U(\infty) > 0$, however, this shift will influence the behavior of the estimator.

Lemma 2. *Let $X_1, \ldots, X_n$ be independent identically distributed random variables with distribution function F and suppose $F \in \mathcal{D}(G_\gamma), x^* = U(\infty) > 0$, i.e., for $x > 0$,*

$$\lim_{t\to\infty}\frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}.$$

Define for $j = 1, 2,$

$$M_n^{(j)} := \frac{1}{k}\sum_{i=0}^{k-1}(\log X_{n-i,n} - \log X_{n-k,n})^j. \tag{4.45}$$

Then for $k(n) \to \infty$, $k/n \to 0$, $n \to \infty$,

$$\frac{M_n^{(j)}}{(a(\frac{n}{k})/U(\frac{n}{k}))^j} \xrightarrow{P} \prod_{i=1}^{j}\frac{i}{1 - i\gamma_-}, \tag{4.46}$$

with $\gamma_- = \min(0, \gamma)$.

This lemma can be used to show that the Hill estimator converges to zero for negative γ and therefore is not very informative. However, the lemma can also be used to construct a consistent estimator for $\gamma < 0$.

Definition 6. *Let $M_n^{(j)}$ be defined as in (4.45) then the* ***moment estimator*** *is defined as*

$$\hat{\gamma}_M := M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1} \tag{4.47}$$

This estimator again has some nice properties.

Theorem 8. *Let $X_1, \dots, X_n$ be independent identically distributed random variables with distribution function F. Suppose $F \in \mathcal{D}(G_\gamma)$ with $\gamma \in \mathbb{R}$ and $x^* > 0$. Then as $n \to \infty$, $= k(n) \to \infty$, $k/n \to 0$*

$$\hat{\gamma}_M \to^P \gamma. \tag{4.48}$$

Proof. Since the moment estimator is a combination of the Hill estimator and $(M_n^{(1)})^2/M_n^{(2)}$, which converges due to Lemma 2, we also know that

$$M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1} \xrightarrow{P} \gamma,$$

for $\gamma \in \mathbb{R}$. □

We notice that there are several other estimators, like the **Probability-Weighted Moment Estimator** for $\gamma < 1$ and the **Negative Hill Estimator** for $\gamma < \frac{1}{2}$ with similar asymptotic behavior as the estimators described in Haan and Ferreira, 2006, section 3.6. The probability-Weighted moment estimator even yields a second estimator for the scale function $a(t)$.

4.6. Extreme quantile and tail estimation

The simplest estimator for a quantile is obviously via the empirical quantile. However there are other approaches to estimate the quantile. For this we have to deal with the estimation $U(t)$ and the scale function $a(t)$.

4.6.1. Scale Estimation

An estimator for the scale function $a(\cdot)$ is obtained similarly to the moment estimator in Section 4.5.2. Recall the notation

$$M_n^{(j)} := \frac{1}{k}\sum_{i=0}^{k-1}(\log X_{n-i,n} - \log X_{n-k,n})^j, \quad j = 1, 2,$$

and define

$$\hat{\gamma}_- := 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}. \tag{4.49}$$

An estimator for the scale then is defined as

$$\hat{\sigma}_M := X_{n-k,n}M_n^{(1)}(1 - \hat{\gamma}_-). \tag{4.50}$$

Note that $\hat{\gamma}_- + \hat{\gamma}_H = \hat{\gamma}_M$ where $\hat{\gamma}_M$ is the moment estimator and $\hat{\gamma}_H$ is the Hill estimator.

We can say something about the asymptotic behavior of the scale estimator.

Theorem 9. *Let $X_1, \dots, X_n$ be independent identically distributed random variables with distribution function F. Suppose $F \in \mathcal{D}(G_\gamma)$ with $\gamma \in \mathbb{R}$ and $x^* > 0$. Then as $n \to \infty$, $k = k(n) \to \infty$, $k/n \to 0$*

$$\frac{\hat{\sigma}_M}{a(\frac{n}{k})} \xrightarrow{P} 1. \tag{4.51}$$

The proof for the Theorem 9 can again be found in Haan and Ferreira, 2006, page 130.

5. The Maximum Likelihood Method

Moving forward we will have a different view at our problem. Previously we had an identically and independently distributed sample $X_1, \dots, X_n$, namely our precipitation events, where $X_1 \sim F$ and wanted to infer on the extreme behavior of this sample.

Now we again have an identically and independently distributed sample $Z_1, \dots, Z_m$ where $Z_1 \sim G_\gamma$, with the key difference that our observations $Z_i, i \leq m$ are extreme values of a year. This implies that our sample size is reduced to only one observation per year, the maximum observed precipitation over a fixed duration in that year.

Since our sample $Z_i, i \leq m$ now only features observations of a few years and previous methods relied on asymptotic behavior, they are not necessarily applicable anymore. Thus we will establish a maximum likelihood based method.

Throughout this section we will refer to the location parameter as b, the scale parameter as σ and the shape parameter as γ. Before establishing the maximum likelihood estimators (MLE) we want to introduce a slightly different notation for the family of extreme value distributions.

Definition 7. *Let $a > 0$, $b \in \mathbb{R}$ and $\gamma \in \mathbb{R}$, then the generalized extreme value distribution is given by*

$$G_{a,\gamma,b}(x) = \begin{cases} \exp(-(1+\gamma\frac{x-b}{a}))^{-1/\gamma} & 1+\gamma\frac{x-b}{a} > 0, \gamma \neq 0 \\ \exp(-\exp(-\frac{x-b}{a})) & x \in \mathbb{R}, \gamma = 0. \end{cases} \tag{5.1}$$

We can now give the log-likelihood function of a sample $Z_1, \ldots, Z_m$ of identically and independently distributed generalized extreme value variables for the case $\gamma \neq 0$,

$$\log L(a, \gamma, b) = -n \log a - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{m} \log\left(1 + \gamma \frac{Z_i - b}{a}\right) \\ - \sum_{i=1}^{m} \left(1 + \gamma \frac{Z_i - b}{a}\right)^{\frac{1}{\gamma}},$$

with $1 + \gamma \frac{Z_i - b}{a} > 0$. For the case of $\gamma = 0$ the log-likelihood function is then

$$\log L(a, 0, b) = -n \log a - \sum_{i=1}^{m} \exp\left(-\frac{Z_i - b}{a}\right) - \sum_{i=1}^{m} \frac{Z_i - b}{a},$$

Our estimators with this method are then obtained by calculating

$$(\hat{a}, \hat{\gamma}, \hat{b}) = \max_{a,\gamma,b} \log L(a, \gamma, b). \tag{5.2}$$

Since the support of G depends on the unknown parameter values, the usual regularity conditions underlying the asymptotic properties of maximum likelihood estimators are not satisfied. However, in the case of $\gamma > -\frac{1}{2}$, the usual properties of consistency, asymptotic efficiency and asymptotic normality hold. For more details about this we refer to Haan and Ferreira, 2006, section 3.4.

6. Estimation of return levels

6.1. Introduction

This chapter we will apply some of the theory outlined previously, in order to obtain return level estimates for $l(s, d, r)$. The sample data used in this Chapter will be the raw `INCA` dataset outlined in Section 2.2.1. Although the data is complete, the major downside is that it only features the years 2004 to 2017, where the last observation year is restricted to the first four months. Consequently, we have only 13 years of complete data.

We recall that we have precipitation amounts in 15 minutes intervals, meaning that each day has 96 sample points, and therefore each year contains approximately $96 \cdot 365 = 35040$ sampling points at each spatial location.

This dataset is obviously highly correlated, because rainfall usually occurs over a larger period than 15 minutes. In addition to this, a seasonal trend is observable. See Figure 6.1.

The theory we have seen above is built around the assumption that the sample consists of i.i.d. observations, which is definitely not realistic in our context. Nevertheless, many of the results we have seen can be extended to a certain extent to more complex settings. See e.g. Tawn, 1988. Motivated by this, we are hence going to assume that for each year $R_{\max(s,d,i)}$—the maximum amount of precipitation in year i at location s of duration d—is distributed according to a generalised extreme value distribution.

The objective is to estimate the unknown parameters of this distribution from a sample

$$\{R_{\max(s,d,i)} : 1 \leq i \leq m\}.$$

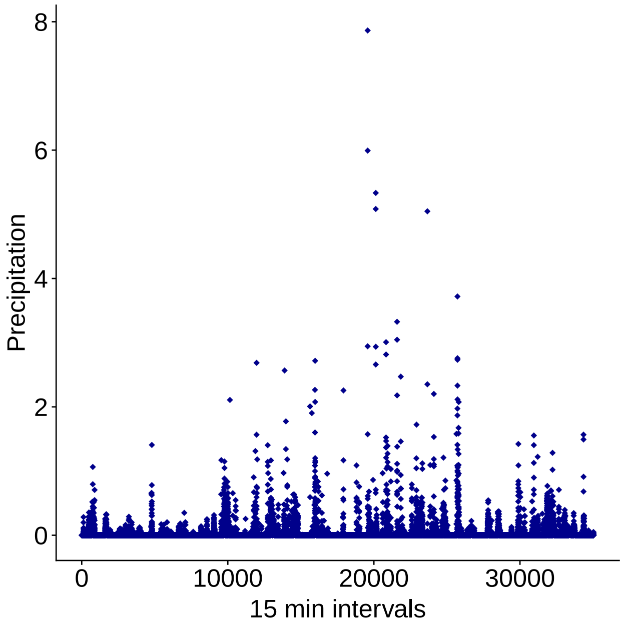


Figure 6.1.: Precipitation in 15 minutes intervals for the year 2010 at the station *Kittsee*.

Although the spatial locations will be correlated we will treat them independently, i.e. we fit a separate model to each spatial location. Later we are going to fix s and d and then use Z_i instead of $R_{\max(s,d,i)}$.

6.2. Demonstration for the station Kittsee

Since the calculations are completely analogue for each duration and each station we will choose for the purpose of demonstration a duration of $d = 24$h and the station *Kittsee*, which is the station listed first in our dataset. (See Figure 6.2.)

We will demonstrate the application of three different approaches to obtain estimates for the location, scale and shape parameters and therefore also the return levels on the station *Kittsee*, with the following three methods:

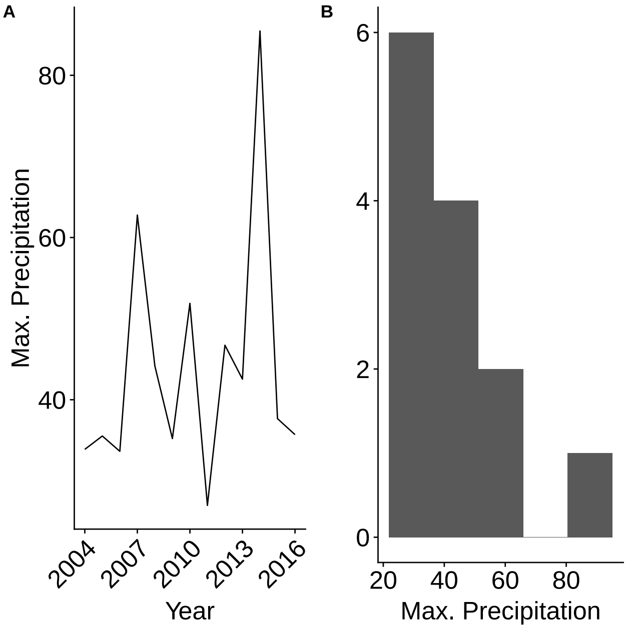


Figure 6.2.: (A) Precipitation-levels for each year, (B) histogram of precipitation-levels in *Kittsee* for the duration of 24h.

1. *The* `QQ`*-method* - A method based on comparing empirical with theoretical quantiles.
2. *The moment-method* - A method based on the moment estimator from Section 4.5.2.
3. *The* `MLE`*-method* - A method based on the Maximum Likelihood estimator from Section 5.

6.2.1. The `QQ`-method

Let $z_1, \dots, z_m$ be the observed annual maximum precipitation levels. (In our application $m = 13$.) We assume they are i.i.d. realisations from random variables whose distribution (appropriately scaled and centered) is given as

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}) & 1+\gamma x > 0 \\ \exp(-e^{-x}) & \gamma = 0. \end{cases}$$

The idea is now to match the quantiles of such a distributions with the empirical quantiles of our data, in the spirit of QQ-plot. To this end we first invert $G_\gamma(x)$ and obtain the p-quantile by

$$Q_\gamma(p) = \begin{cases} \frac{\left(\frac{1}{\log(1/p)}\right)^\gamma - 1}{\gamma} & \gamma \neq 0 \\ \log(\log(1/p)^{-1}) & \gamma = 0. \end{cases} \tag{6.1}$$

The corresponding QQ-plot consists of the pairs

$$\left(Q_\gamma\left(\frac{1}{m+1}\right), z_{1,m}\right), \ldots, \left(Q_\gamma\left(\frac{m}{m+1}\right), z_{m,m}\right). \tag{6.2}$$

Note that we did not use the normalized form of the generalized extreme value distribution as in Definition 5.1, since our estimator of γ here is not depending on a or b. A linear point pattern indicates a good fit of the distribution. Since the generalized extreme value distribution function is bounded from above for $\gamma < 0$ and we did not assume any bounds on the precipitation levels, we assume that $0 \geq \gamma$. Futhermore we will limit ourselves to $\gamma < 5$ to prevent unnecessary long computations. The goal is hence to tune $0 < \gamma < 5$, such that the correlation between empirical and theoretical quantiles is maximised.

For the station *Kittsee* this yields the value $\hat{\gamma} = 0.63$.

Next, we will use $\hat{\gamma}$ to obtain an estimate for a return level. With our quantile estimator we obtain an estimation for the return level r with

$$\hat{U}(r) = \hat{Q}(1 - 1/r) = \hat{b} + \hat{a}\frac{(-\log(1 - 1/r))^{-\hat{\gamma}} - 1}{\hat{\gamma}}, \tag{6.3}$$

where $\hat{a}$ and $\hat{b}$ are the slope and intercept of the model fitted in the quantile-quantile plot. For the station *Kittsee* this yields the values $\hat{a} = 7.54$ and $\hat{b} = 35.6$. (Figure 6.4.)

Plugging $\hat{\gamma}$, $\hat{a}$ and $\hat{b}$ into (6.3) we then obtained return level estimates for the return periods $2, 3, 5, 10, 100$ years, which can be found in Table 6.2.1 under in row named *Estimation-QQ*. This table also includes the corresponding

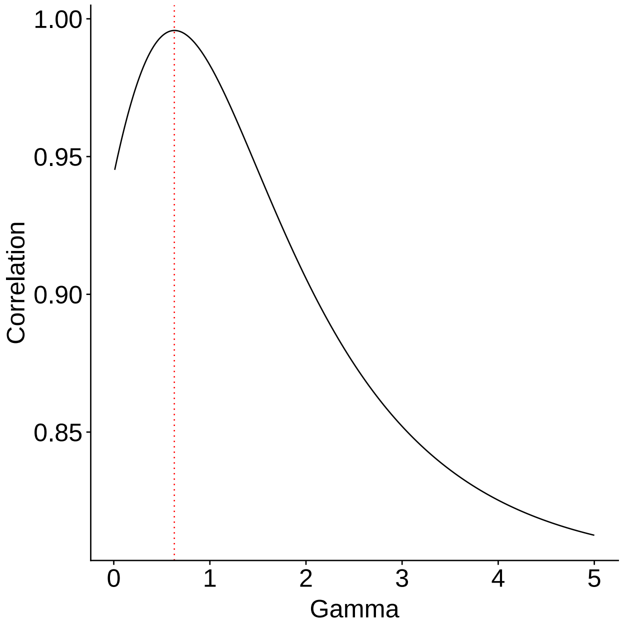


Figure 6.3.: Correlation values for different γ values. The maximum is marked by the red line.

years	2	3	5	10	100
Estimation-QQ	38.711	44.77	54.43	73.04	240.74
Quantile	37.67	44.16	49.82	60.59	82.74
INCA	49.00	57.6	67.21	79.27	-
INCA 95%	(38.8,59.2)	(44.8,70.4)	(50.9, 83.5)	(58.1,100.5)	-
OEKOSTRA	42.1	45.8	50.4	56.7	77.6

Table 6.1.: Return level estimates for *Kittsee* with empirical quantiles and INCA-estimates with 0.95 confidence intervals and OEKOSTRA-estimates for comparison.

quantiles, the INCA return levels as well as the provided 0.95 confidence intervals and the OEKOSTRA return levels.

Note that when estimating the return level for a return period r we are just estimating the $(1-1/r)$-quantile. For example the return period of 2 years is an estimation for the median of our sample. This simple estimator is also included in our Table 6.2.1. Estimating the return period $r = 100$

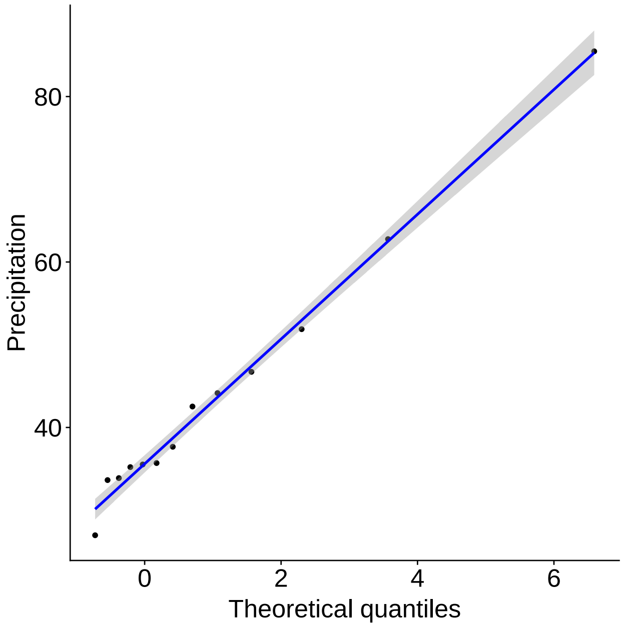


Figure 6.4.: The q-q-plot for *Kittsee* with $\gamma = \hat{\gamma}$.

is questionably when we have only 13 observations, but it was included nevertheless. In Section 6.7 we are going to assess the impact of the sample size in the estimation procedure by means of a simulation study.

The newly derived estimates are not within the provided 0.95 confidence intervals for lower return periods of `INCA`. What is interesting is that at this station our estimates compare better with the `OEKOSTRA` values than the INCA return levels which were included in the data files. We are going to further analyse this in Sections 6.4 and 6.5, in order to see if this holds true more generally, i.e. across stations.

6.2.2. The moment-method

While Our `QQ`-method seems to produce acceptable results, we have no theory to support this approach. Thus, we aim to use a method based on the more established theory from Section 4.5.2. Despite the fact the moment estimator relies on asymptotic behavior and applying intermediate sequences with only 13 observations is rather ill-advised, we tried using it for our return level calculation nevertheless.

While the results we obtained for the station *Kittsee* look acceptable at first glance, we will later show in the simulations from Section 6.7, that indeed the moment method performs much worse then the other methods. Thus, we will only show the application of the moment method only for the station *Kittsee* and afterwards omit this method for our further computations.

Before we continue with the moment estimator we will derive a way to use the estimator for the shape parameter obtained from this method. To calculate the return level we need an estimation for the shape and location parameter but we won't have a `QQ`-plot and corresponding intercept or slope parameters available.

Consider again the alternate formulation of our extreme value condition (4.33) and rewrite it as

$$U(x) \approx b(t) + a(t)\frac{(x/t)^{\gamma} - 1}{\gamma}, \quad x > t. \tag{6.4}$$

which can be used to obtain an estimation for a small p

$$U(y) \approx U(\frac{n}{k}) + a(\frac{n}{k})\frac{(\frac{yk}{n})^{\gamma} - 1}{\gamma}, \quad y = \frac{1}{p} \tag{6.5}$$

where k is an intermediate sequence. This means we need to estimate our location parameter $U(n/k)$, the scale parameter $a(t)$ and the shape parameter γ. The location parameter can be estimated by the order statistic $X_{n-k,n}$, but estimators for the shape and the scale parameter need to be constructed.

The difficulty here lies with choosing the parameter k. Since k is an intermediate sequence it can in practice only be chosen to be $k \in [1, n-1]$ where n is the sample size.

Calculating the moment estimations $\hat{\gamma}_M$ yield different values for all the possible k values, shows that for very low values of k the parameter is likely out of range of the real γ and stabilizes as it approaches higher values. (Figure 6.5)

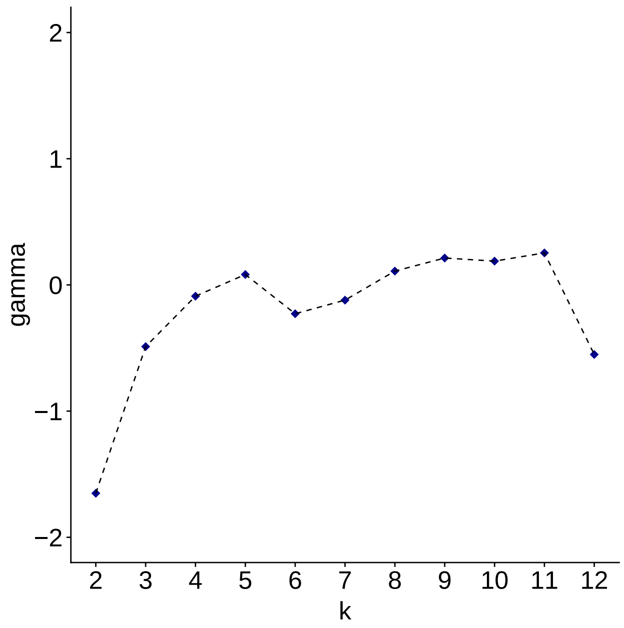


Figure 6.5.: $\hat{\gamma_M}$ for different values of k

Since the 0.95 confidence interval for $\hat{\gamma}$ in the last section was $[0.07, 0.24]$, and from visualizing $\hat{\gamma}$ for different k values, we can assume that $k = 9, 10$ with $\hat{\gamma} = 0.21, 0.19$ are good choices.

Lastly we need to estimate the scale parameter $a(t)$. This is simply done by calculating the estimator in (Equation 4.50).

Combining all our estimators we can calculate the return level for the return periods $2, 3, 5, 10$ and 100 years.

$$\hat{U}(r) = X_{n-k,n} + \hat{\sigma}_M \frac{(\frac{rk}{n})^{\hat{\gamma}} - 1}{\hat{\gamma}}, \quad k = 10, 11. \tag{6.6}$$

For $k = 9$ and $k = 10$ the results are very similar as showed in Table 6.2.2

years	2	3	5	10	100
Estimation k=9	38.97	44.04	51.08	61.94	112.05
Estimation k=10	39.00	44.23	51.40	62.30	110.65
Quantile	37.67	44.16	49.82	60.59	82.74
INCA	49.00	57.6	67.21	79.27	-
INCA 95%	(38.8,59.2)	(44.8,70.4)	(50.9, 83.5)	(58.1,100.5)	-
OEKOSTRA	42.1	45.8	50.4	56.7	77.6

Table 6.2.: Return level estimates with the moment method for *Kittsee* with empirical quantiles and INCA-estimates with 0.95 confidence intervals and OEKOSTRA-estimates for comparison.

The return levels for $r = 100$ in all cases are substantially higher than the OEKOSTRA estimates, which is no surprise since we only have 13 observations from whom only one is above 80 for *Kittsee*. The OEKOSTRA data might have featured far more observations in a similar magnitude and hence higher return periods were probably estimated much lower.

6.2.3. The MLE-method

Since our sample size is quite small with $m = 13$ our third and last approach will be according to the theory of Section 5. As stated earlier the Maximum Likelihood estimators do not depend on asymptotic theory, but rather on solving the equation (5.2). We used the R-package *extRemes* [1] to solve the equation and obtain the MLE estimators.

The MLE-method produces very similar results as the moment-method for the station *Kittsee*, as seen in Table 6.2.3. Note that the MLE-method is the only one of our three methods, the QQ-method, the moment-method and the MLE-method, that matches the INCA confidence intervals for every return period, meaning that the estimators of the MLE-method lie within the INCA confidence intervals for each considered return period ($r = 2, 3, 5, 10$).

[1] https://cran.r-project.org/web/packages/extRemes/extRemes.pdf

years	2	3	5	10	100
Estimation-MLE	39.87	45.04	51.68	61.55	107.51
Quantile	37.67	44.16	49.82	60.59	82.74
INCA	49.00	57.6	67.21	79.27	-
INCA 95%	(38.8,59.2)	(44.8,70.4)	(50.9, 83.5)	(58.1,100.5)	-
OEKOSTRA	42.1	45.8	50.4	56.7	77.6

Table 6.3.: Return level estimates with the MLE method for *Kittsee* with empirical quantiles and INCA-estimates with 0.95 confidence intervals and OEKOSTRA-estimates for comparison.

6.2.4. Comparison of all methods

Let us summarize again the results of the different estimation methods for the station *Kittsee*. We consider the return periods $r = 2, 3, 5, 10$. We are omitting $r = 100$, since the INCA data files didn't contain return levels for $r = 100$. Now we reference our estimates with the INCA and OEKOSTRA return levels which were provided in the data files. The results are displayed in Figure 6.6. On the left hand graph we see the return level estimates for INCA and OEKOSTRA. And on the right hand side we see the return level estimates from our 3 estimation methods using the INCA raw data. The main reference value in both frames are the empirical quantiles (in black).

Despite the QQ-method, which appears to overestimate the return levels for longer return periods, we observe that the estimates are close to the empirical quantiles. Surprisingly, the return levels from moment-method, the MLE-method and the OEKOSTRA dataset are very close to the quantiles, while INCA seems to be far off. We will see in Section 6.7, that indeed our estimates match the return level estimates from OEKOSTRA much better, than the INCA return levels.

Lastly, while the MOMENT-method seems to produce good results for the *Kittsee* station, we will drop this method from further consideration since our sample size of 13 observations is too small to consider an approach that relies on asymptotic behavior.

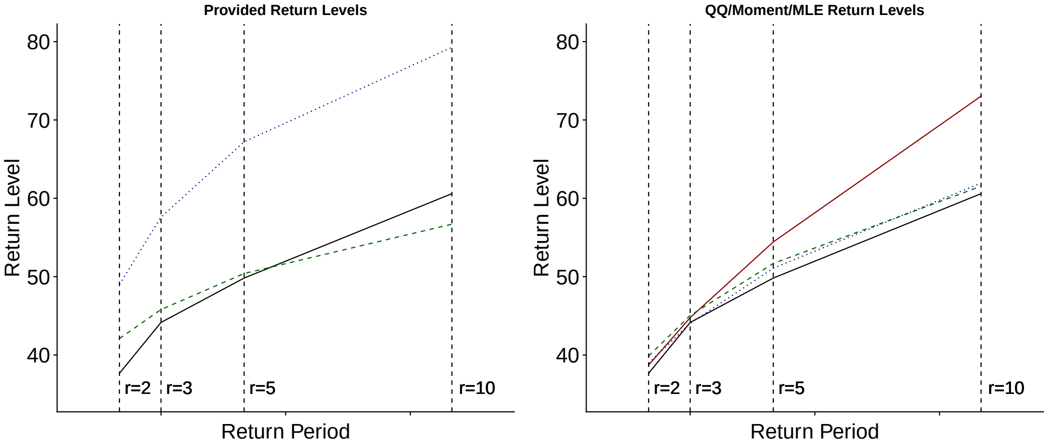


Figure 6.6.: Comparison of our three described estimation methods applied to the station *Kittsee*. The empirical quantiles are shown in both frames (black, solid). On the left side we see the INCA return levels (blue, dotted) and the OEKOSTRA return levels (green, dashed). On the right side we see the QQ-method (red, solid), the moment-method (green, dashed) and the MLE-method (blue, dotted).

6.3. Bootstrapping

Since we can construct confidence intervals for the moment-method and the MLE-method from theory but not for the QQ-method, we constructed confidence intervals for the QQ-method by resampling our sample, a procedure know as bootstrapping.

We draw from our sample at random with replacement T times and repeat this process B times. In each iteration we calculate $\hat{\gamma}$ as done in the previous sections. The pseudo-code for the procedure is given by Algorithm 1.

The Algorithm 1 can easily be modified to provide confidence intervals for the shape parameter γ, by just skipping the calculation of the return levels and instead calculating quantiles over the shape parameters of each step. The resulting $\hat{\gamma}_1, \ldots, \hat{\gamma}_B$, when applying the Algorithm 1 to the *Kittsee* sample, are distributed as shown in (Figure 6.7).

A 0.95 bootstrapping confidence interval for γ is $(0.01, 1.28)$. The estimated return levels for the return periods $2, 3, 5$ and 10 years with $\hat{\gamma} = med(\hat{\gamma}_i) = 0.45$ are shown in Table 6.4.

Algorithm 1 Bootstrapping algorithm for return levels

```
Given $Z = (Z_1, \dots, Z_m), B \in \mathbb{N}$
for $i = 0, i < B$ do
    Construct sample $Z_1^*, \dots, Z_m^*$ by drawing from $Z$ with replacement
    Calculate $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$
    Use $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$ to calculate the return level $r_i^*$
    i++
Calculate the quantiles $q_{0.25}$ and $q_{0.975}$ from $r_1^*, \dots, r_B^*$
The return level confidence interval is then given by $(q_{0.25}, q_{0.975})$
```

years	2	3	5	10	100
Estimate	39.4	46.2	56.0	72.6	180.5
Quantile	37.67	44.16	49.82	60.59	82.74
INCA	49.0	57.6	67.2	79.3	-
INCA 95%	(38.8,59.2)	(44.8,70.4)	(50.9, 83.5)	(58.1,100.5)	-
OEKOSTRA	42.1	45.8	50.4	56.7	77.6

Table 6.4.: Return level estimates with bootstrapping for *Kittsee* with INCA-estimates with 0.95 confidence intervals and OEKOSTRA-estimates for comparison.

This time our estimate lies within the 0.95 confidence interval of the INCA return level estimates. This approach estimates lower return periods similarly as the OEKOSTRA estimates but overestimates higher return periods. The return levels for $r = 10, 100$ seem much more reasonable with this approach.

We will use Algorithm 1 to calculate confidence intervals for the return levels later. For now we move on to applying our methods seen in this section for every station in Austria.

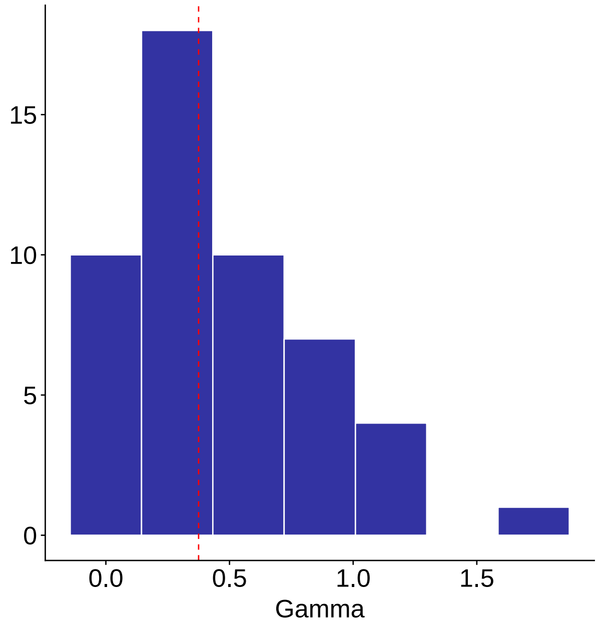


Figure 6.7.: Bootstrapping data for $\hat{\gamma}$ with 100 iterations with sample size 13. The red line represents the median

6.4. Comparison with INCA and OEKOSTRA for $d = 24$h

In this section we expand our return level estimation to every station in Austria and compare the results with the INCA and OEKOSTRA estimates. We will, as argued in Section 6.2.2 Section 6.2.4 proceed only with the QQ-method MLE-method and drop the MOMENT-method. The main reasoning again is that the MOMENT-method on one hand relies on asymptotic behavior and we only feature 13 observations.

Next, the general idea for calculating the return levels for each station is provided with the Algorithm 2.

Note that we will focus mostly on the evaluation and interpretation of the connection between our results, which are calculated from the raw INCA

Algorithm 2 Return level for each station

```
1: Given stations $1, \dots, N$
2: for $i = 0, i < N$ do
3:     Calculate yearly maxima $Z_{i1}, \dots, Z_{im}$
4:     Calculate $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$ from $Z_{i1}, \dots, Z_{im}$
5:     Use $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$ to calculate the return level $r_i$
6:     i++
```

data, and the provided OEKOSTRA return level estimates, which we consider as our benchmark.

6.4.1. The QQ-method

We compute the return levels $\text{INCA}^{\text{QQ}} = \{r_1, \dots, r_N\}$ from the raw INCA data, by applying the QQ-method in the 4th step of Algorithm 2. The resulting return levels r_i can be seen in Figure 6.8. For a duration of $d = 24\text{h}$ and a return period of $r = 2$ years we obtain correlation values of $R = 0.81$ and $R = 0.78$ as seen in Figure 6.9 and Figure 6.10 with INCA and OEKOSTRA, respectively. Hence our estimates perform in this sense equally well as when INCA was compared to OEKOSTRA.

We again performed regression analysis to obtain the following model (6.7).

$$\widehat{\text{INCA}}^{\text{QQ}} = 0.699 * \text{OEKOSTRA} + 15.226, \tag{6.7}$$

with

```
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)    15.22569   1.12807      13.50     <2e-16
Oekostra       0.69883    0.01791      39.01     <2e-16

Residual standard error: 7.569 on 914 degrees of freedom
Multiple R-squared:  0.6248,Adjusted R-squared:  0.6244
F-statistic:  1522 on 1 and 914 DF,  p-value: < 2.2e-16
```

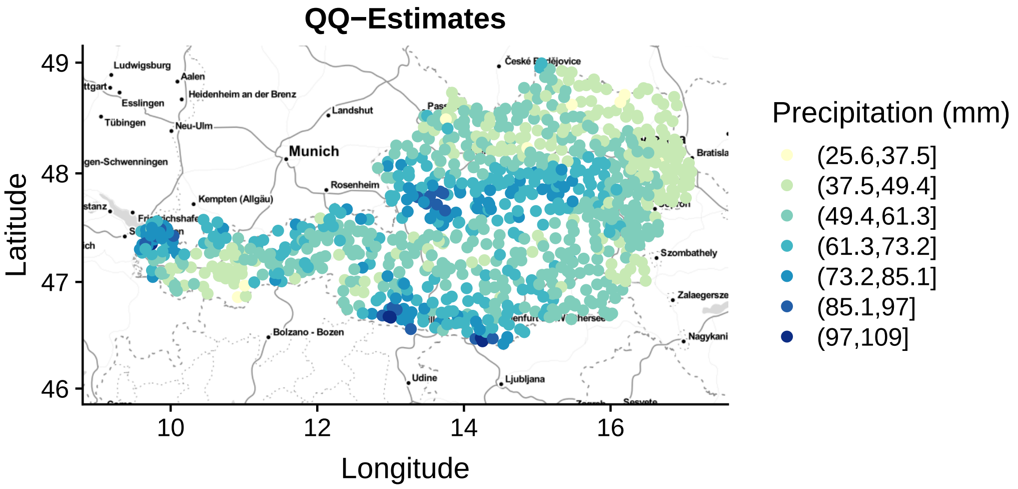


Figure 6.8.: Estimates for return level 2 years and duration 24h by the QQ-method.

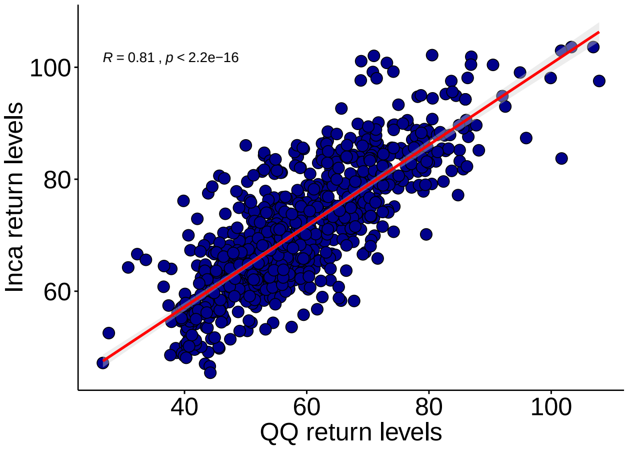


Figure 6.9.: QQ-estimates vs INCA for return level 2 years and duration 24 hours

This linear model has a much lower intercept, but a bigger slope than the model from our preliminary evaluations (3.1). Considering that the return level estimates from the QQ-method correlated quite well with results from INCA and OEKOSTRA we can conclude that the QQ-method provides an acceptable way of calculating return levels in this case.

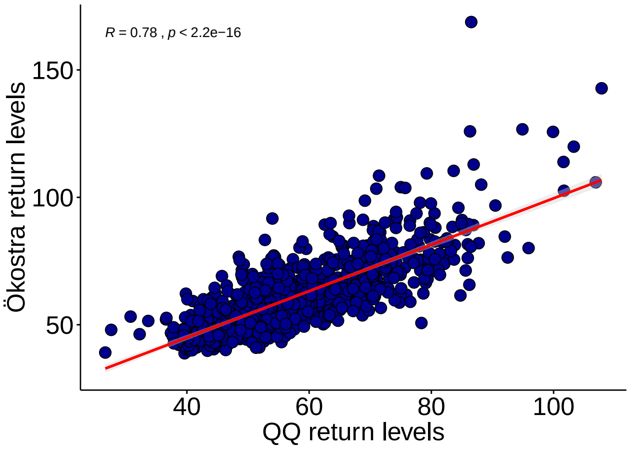


Figure 6.10.: QQ-estimates vs OEKOSTRA for return level 2 years and duration 24 hours

Including the sealevel in our model (see (6.8)) we observe that it is only significant with interaction between sealevel and OEKOSTRA. This could imply that the return levels are differentiated among sealevels:

$$\widehat{\texttt{INCA}}^{\texttt{QQ}} = 0.00014 * h + 0.6554 * \texttt{OEKOSTRA} + 17.12, \tag{6.8}$$

```
               Estimate    Std. Error  t value    Pr(>|t|)
(Intercept)    1.712e+01   1.056e+00   16.22      <2e-16
Oekostra       6.554e-01   1.776e-02   36.90      <2e-16
Height         1.044e-04   6.153e-04   0.17       0.865

Residual standard error: 7.278 on 914 degrees of freedom
Multiple R-squared:  0.6285,Adjusted R-squared:  0.6277
F-statistic: 773.1 on 2 and 914 DF,  p-value: < 2.2e-16
```

and

$$\widehat{\texttt{INCA}}^{\texttt{QQ}} = -0.000094h * \texttt{OEKOSTRA} + 0.0039 * h + 0.234 * \texttt{OEKOSTRA} + 13.97, \tag{6.9}$$

```
               Estimate    Std. Error  t value    Pr(>|t|)
(Intercept)    1.397e+01   1.278e+00   10.934     < 2e-16
```

```
Oekostra        2.234e-01   2.212e-02   10.100    < 2e-16
Height          3.905e-03   1.493e-03   2.615     0.009058
Interaction     -9.371e-05  2.429e-05   -3.858    0.000122

Residual standard error: 7.13 on 822 degrees of freedom
Multiple R-squared:  0.6485,Adjusted R-squared:  0.6472
F-statistic: 505.5 on 3 and 822 DF,  p-value: < 2.2e-16
```

6.4.2. The MLE method

Next, we tried applying the the MLE-method from Section 5 to obtain return level estimations for every station by estimating the parameters of the extreme value functions with the MLE method in the 4th step of Algorithm 2.

However, we did not obtain results for every station since the underlying numerical algorithm of finding the solutions of the MLE-equation (5.2) did not converge for every station. Since finding the solution of (5.2) is not trivial we choose to omit stations where the MLE-method didn't converge. As a short remark we want to point out that we did experiment with supplying initial values for the parameters to be estimated by the MLE-method, such as the estimators provided by the QQ-method, which seems to cause the process to converge for some stations but we did not further investigate this approach.

We therefore obtain estimations for return levels for 827 stations, which we denote by $\texttt{INCA}^{\texttt{MLE}}$.

When we regress $\widehat{\texttt{INCA}}^{\texttt{MLE}}$ onto OEKOSTRA we get the following results:

$$\widehat{\texttt{INCA}}^{\texttt{MLE}} = 0.662 * \texttt{OEKOSTRA} + 16.939, \tag{6.10}$$

```
                Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)     16.93908    1.10383     15.35     <2e-16
Oekostra        0.66239     0.01749     37.87     <2e-16

Residual standard error: 7.256 on 824 degrees of freedom
Multiple R-squared:  0.635,Adjusted R-squared:  0.6346
```

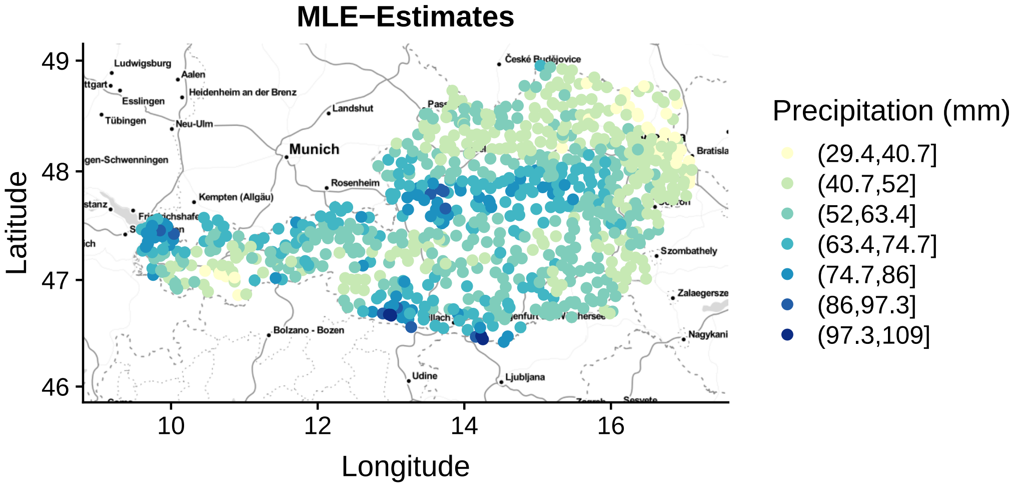


Figure 6.11.: mle-method estimates for return level 2 years and duration 24 hours.

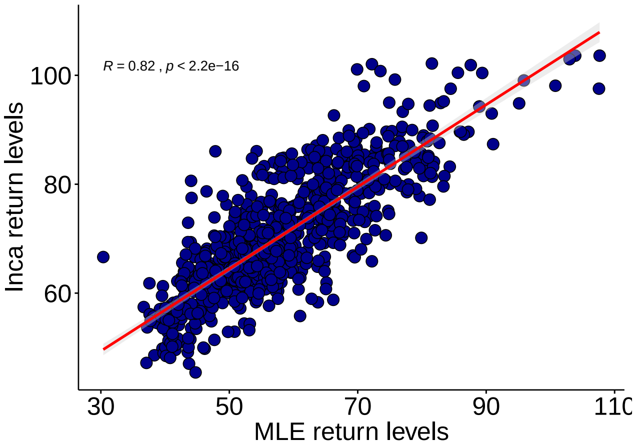


Figure 6.12.: MLE-estimates vs INCA for return level 2 years and duration 24 hours.

```
F-statistic:  1434 on 1 and 824 DF,  p-value: < 2.2e-16
```

The correlation between INCA^{MLE} and OEKOSTRA increased to 0.8, showing now slightly better correspondence as it was with INCA^{QQ}.

Again, including the sealevel in our model (see (6.11)) we observe that it does

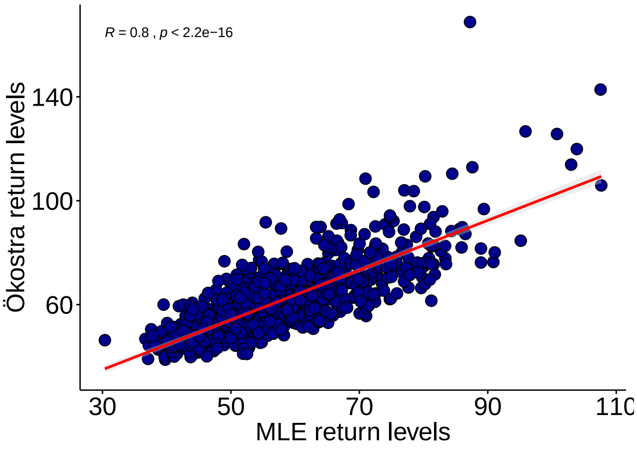


Figure 6.13.: MLE-estimates vs OEKOSTRA for return level 2 years and duration 24 hours.

not appear to be significant this time. However, if we include the interaction term (6.12) we obtain indeed strong significance of the variables.

$$\widehat{\texttt{INCA}}^{\texttt{MLE}} = 0.0004 * h + 0.579 * \texttt{OEKOSTRA} + 16.93, \tag{6.11}$$

```
                Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)     1.693e+01   1.104e+00   15.326    <2e-16
Oekostra        6.579e-01   1.879e-02   35.019    <2e-16
Height          4.379e-04   6.672e-04   0.656     0.512

Residual standard error: 7.259 on 823 degrees of freedom
Multiple R-squared:  0.6352,Adjusted R-squared:  0.6344
F-statistic: 716.6 on 2 and 823 DF,  p-value: < 2.2e-16
```

and

$$\widehat{\texttt{INCA}}^{\texttt{MLE}} = -0.00023h * \texttt{OEKOSTRA} + 0.014 * h + 0.839\texttt{OEKOSTRA} + 6.53, \tag{6.12}$$

```
                Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)     6.563e+00   2.153e+00   3.048     0.00238
```

```
Oekostra        8.389e-01   3.737e-02   22.451    < 2e-16
Height          1.425e-02   2.564e-03   5.557     3.71e-08
Interaction     -2.313e-04  4.151e-05   -5.571    3.43e-08

Residual standard error: 7.13 on 822 degrees of freedom
Multiple R-squared:  0.6485,Adjusted R-squared:  0.6472
F-statistic: 505.5 on 3 and 822 DF,  p-value: < 2.2e-16
```

6.5. Comparison for $d = 3$h

Since the discrepancy between the INCA and the OEKOSTRA return level estimates for the duration $d = 3$h is one of the foci of this thesis, we are now moving on to apply Algorithm 2 to every station in Austria for both the QQ-method and the MLE-method, for the case that $d = 3$h.

6.5.1. The QQ-method

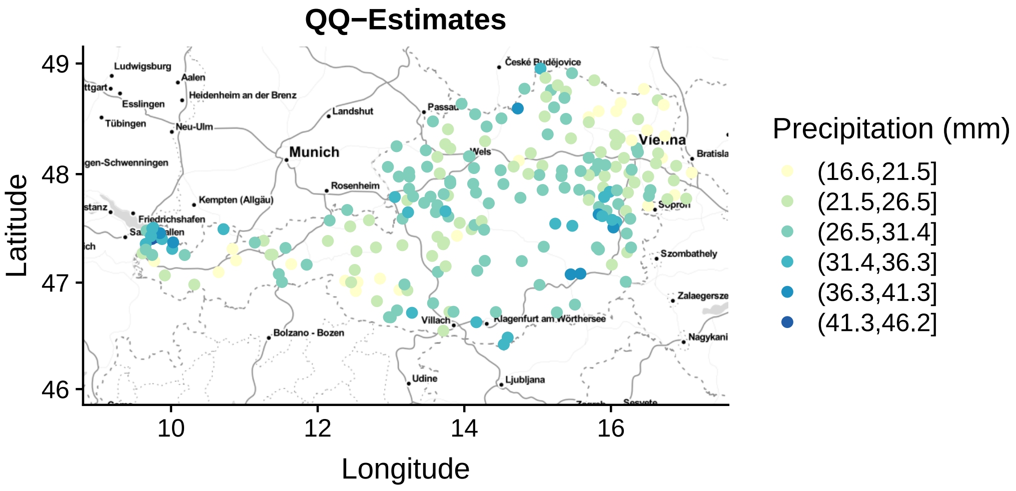


Figure 6.14.: QQ-method estimates for return level 2 years and duration 3 hours.

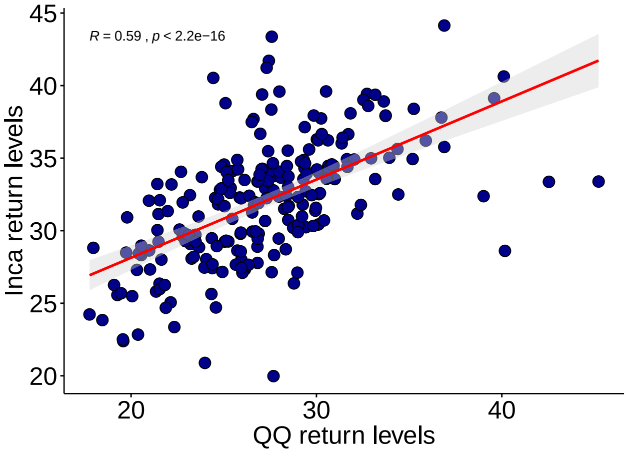


Figure 6.15.: `QQ`-estimates vs `INCA` for return level 2 years and duration 3 hours.

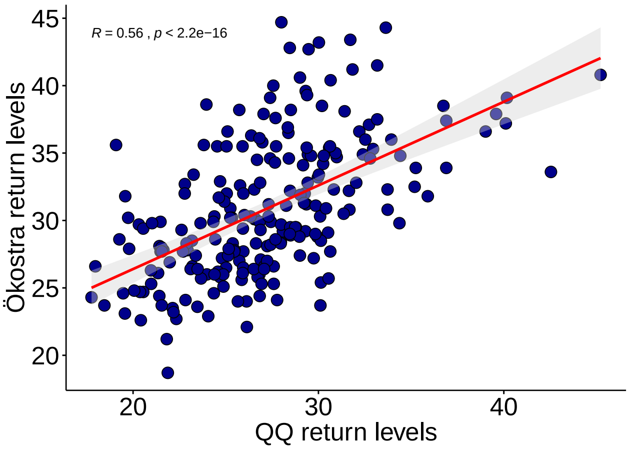


Figure 6.16.: `QQ`-estimates vs `OEKOSTRA` for return level 2 years and duration 3 hours.

With a correlation of $R = 0.56$ Figure 6.16 suggests that our approach to compute $\texttt{INCA}^{\texttt{QQ}}$ provides a significantly higher correlation to `OEKOSTRA` as compared to `INCA`, where we had $R = 0.49$. We computed the regression models again and obtained similar results as before, for the case $d = 3$.

$$\widehat{\texttt{INCA}}^{\texttt{QQ}} = 0.505 * \texttt{OEKOSTRA} + 11.599, \tag{6.13}$$

with

```
              Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)   11.59870    1.57995     7.341     4.1e-12
Oekostra      0.50512     0.05053     9.997     < 2e-16

Residual standard error: 3.821 on 219 degrees of freedom
Multiple R-squared:  0.3133,Adjusted R-squared:  0.3102
F-statistic: 99.93 on 1 and 219 DF,  p-value: < 2.2e-16
```

As seen in (6.13), a linear regression model with OEKOSTRA yields similar results as in the preliminary evaluation but fits the data slightly better. We again included the sea level as response variable.

$$\widehat{\texttt{INCA}}^{\texttt{QQ}} = 0.0014 * h + 0.509 * \texttt{OEKOSTRA} + 10.58, \tag{6.14}$$

```
                Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)     1.058e+01   1.651e+00   6.41      8.87e-10
Oekostra        5.095e-01   5.024e-02   10.14     < 2e-16
Height          1.427e-03   7.175e-04   1.99      0.0479

Residual standard error: 3.796 on 218 degrees of freedom
Multiple R-squared:  0.3256,Adjusted R-squared:  0.3194
F-statistic: 52.62 on 2 and 218 DF,  p-value: < 2.2e-16
```

and

$$\widehat{\texttt{INCA}}^{\texttt{QQ}} = 0.00048h * \texttt{OEKOSTRA} - 0.012 * h + 0.18 * \texttt{OEKOSTRA} + 19.95, \tag{6.15}$$

```
                Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)     19.9527555  3.0186615   6.610     2.94e-10
Oekostra        0.1776302   0.1028692   1.727     0.08563
Height          -0.0120181  0.0037332   -3.219    0.00148
Interaction     0.0004833   0.0001318   3.666     0.00031

Residual standard error: 3.692 on 217 degrees of freedom
Multiple R-squared:  0.3649,Adjusted R-squared:  0.3561
F-statistic: 41.56 on 3 and 217 DF,  p-value: < 2.2e-16
```

Surprisingly, the model (6.15) shows that the significance of OEKOSTRA dropped substantially. The overall R-squared increases substantially when including the sealevel as interaction but is generally low with 0.3649 in the best case, which suggests that a linear model might not be a good fit.

6.5.2. The MLE-method

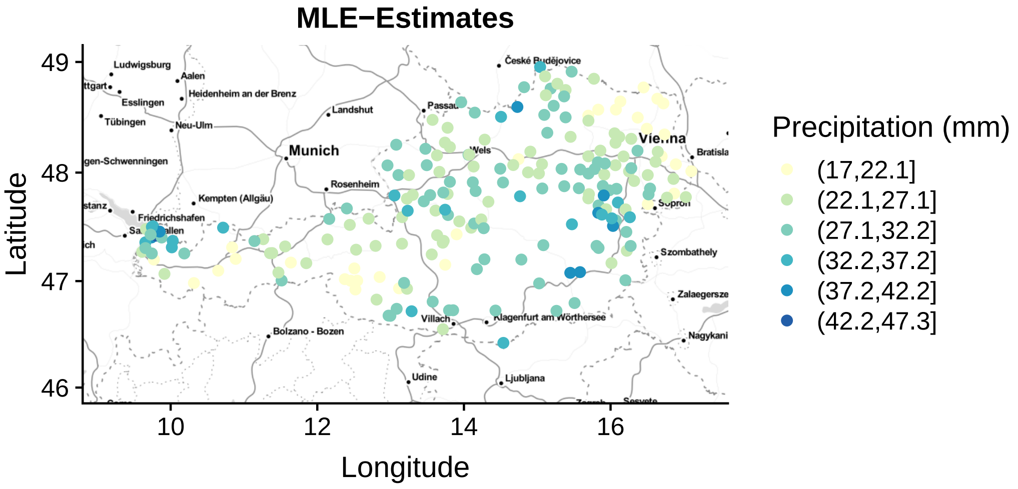


Figure 6.17.: mle-method estimates for return level 2 years and duration 3 hours.

The return levels calculated from MLE-method have a slightly higher minimum and maximum precipitation value as can be seen comparing the legend from Figure 6.14 and Figure 6.17. The correlation with the OEKOSTRA return levels is with $R = 0.55$ equally good as the correlation obtained from the QQ-method and again better then the correlation between INCA and OEKOSTRA in the preliminary evaluation.

The regression analysis for this method yielded very similar results as for the QQ-method and did not bring new insights. We therefore chose to omit them here.

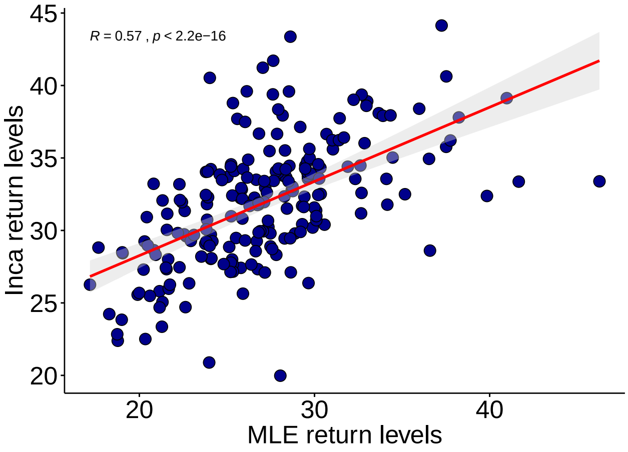


Figure 6.18.: mle-estimates vs `INCA` for return level 2 years and duration 3 hours.

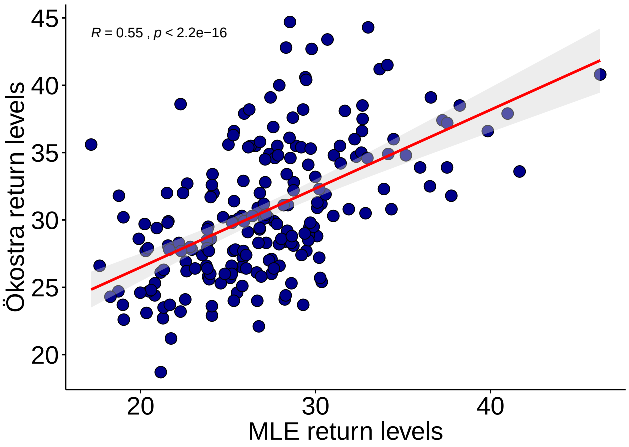


Figure 6.19.: MLE-estimates vs `OEKOSTRA` for return level 2 years and duration 3 hours.

6.6. Confidence intervals for bootstrapped data

Considering that we achieved good results with our methods we will dedicate this section towards constructing confidence intervals for bootstrapped data as in Section 6.3. We will assume the common $\alpha = 0.05$ significance level, resulting in 0.95 percent confidence intervals throughout this section. The general idea is to repeat Algorithm 1 for every station.

6.6.1. Comparison of confidence intervals from the QQ and the MLE-method

To create confidence intervals we simply apply Algorithm 6.3 for every station and obtain the confidence interval $(\hat{q}_\alpha, \hat{q}_{1-\alpha})$. The methods used in step 4 of the Algorithm 6.3 are again the MLE or the QQ-method as reflected in Table 6.6.1.

method	duration	INCA	OEKOSTRA
MLE	24h	56.1 %	83.5 %
MLE	3h	65.1 %	72.4 %
QQ	24h	32.1 %	72.1 %
QQ	3h	52.6 %	63.8 %

Table 6.5.: Amount (in percent) of stations whose 2-year INCA-estimates and OEKOSTRA-estimates are within the confidence intervals of our estimates.

We observe that the MLE-method overall matches the provided INCA-estimates and OEKOSTRA-estimates more often than the QQ-method but both provide acceptable results, especially for OEKOSTRA.

A very important finding here is that we cover OEKOSTRA more often then INCA, although we use INCA data to construct our confidence intervals.

Additionally we derived the length of the confidence intervals we created and compared those lengths to the confidence intervals provided by INCA. The results can be seen in Figure 6.20.

The corresponding mean and median interval lengths can be found in Table 6.6.1.

	MLE	QQ	INCA
Mean	24.53	17.77	20.90
Median	22.29	16.28	20.58

From this we can deduce that the QQ-method constructs more compact confidence intervals and therefore also matches less return stations in the comparison. The MLE-method in contrast constructed the widest confidence

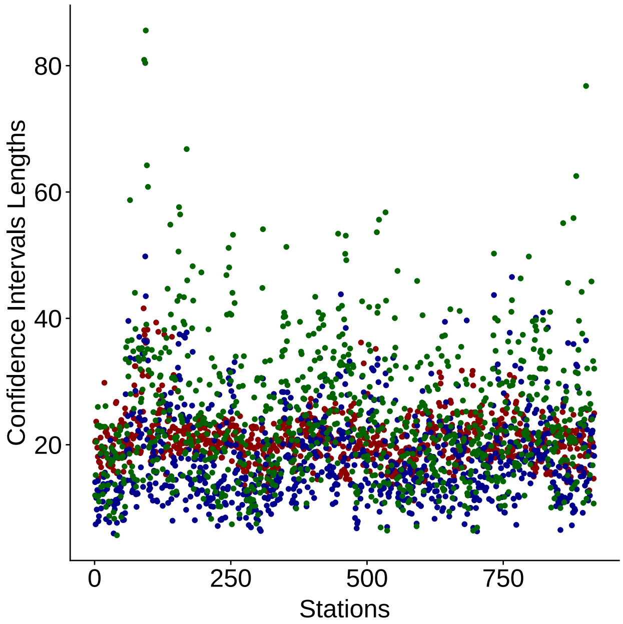


Figure 6.20.: Interval lengths of confidence intervals for `INCA` (red), `QQ`-method (blue) and the `MLE`-method (green)

intervals which surely impacted on the fact that they matched the most return levels. Some of the confidence intervals obtained from the `MLE`-method seem unreasonably wide.

6.7. Analysis on the impact of the sample size

In this last section we focus our attention towards simulated data in order to get a sense of how well our methods perform with different sample sizes. We revert to our toy data, namely the station *Kittsee*.

We discussed the estimation of the shape, location and scale parameter of the extreme value function for the maximized data obtained in *Kittsee* in Section 6.2.1. We use the estimators obtained from the bootstrapping method, $\hat{\gamma} = 0.17$, $\hat{a} = 9.13$ and $\hat{b} = 35.83$ and generate observations from the extreme value distribution function $G_{\hat{\gamma},\hat{a},\hat{b}}$, see (5.1).

We then simulate data according to the Algorithm 3.

Algorithm 3 Simulation Algorithm A

```
Given $\hat{\gamma}, \hat{a}, \hat{b}$ and $N_0, N_1 \in \mathbb{N}, N_0 < N_1$
for $i = N_0, i < N_1$ do
    Construct random sample $X_1, \dots, X_i \in G_{\hat{\gamma},\hat{a},\hat{b}}$
    Calculate $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$
    Use $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$ to calculate the return level $r_i^*$
    i++
```

Looking at Figure 6.21 we notice that the `QQ` method and the `MLE` method estimate the location, scale and shape parameter of the given extreme value distribution $G_{\hat{\gamma},\hat{a},\hat{b}}$, quite well, especially with a greater sample size. The moment-method however overestimates the location and the scale parameter while it slightly underestimates the shape parameter.

While the estimates for the parameters of the `MLE` and the `QQ` method of the extreme value function vary for smaller sample sizes, they approach the real parameters as the sample size grows.

Surprisingly, however, every method yields quite good results for the return level estimates (Figure 6.22). As expected the estimates approach the corresponding know quantile of the extreme value distribution, as the sample size grows toward the return period r.

Figure 6.22, graph (A) shows that for the return period $r = 2$ the estimates are close to the actual value even for smaller sample size while we can observe from graph (C) and (D) that estimation of higher return levels performs poorly for smaller sample sizes.

We now fix the amount of how many times we repeat the experiment $N = 100$ and try estimating the precipitation levels for $r = 2, 3, 5, 10$ with sample sizes $T = 20, 50, 100, 200$, generated from $G_{\hat{\gamma},\hat{a},\hat{b}}$, according to Algorithm 4.

As expected we notice in Figure 6.23 that the boxes shrink with growing sample size, as the estimation is less prone to sampling errors. Additionally we observe that the `QQ`-method indeed tends to overestimate the precipitation in all cases.

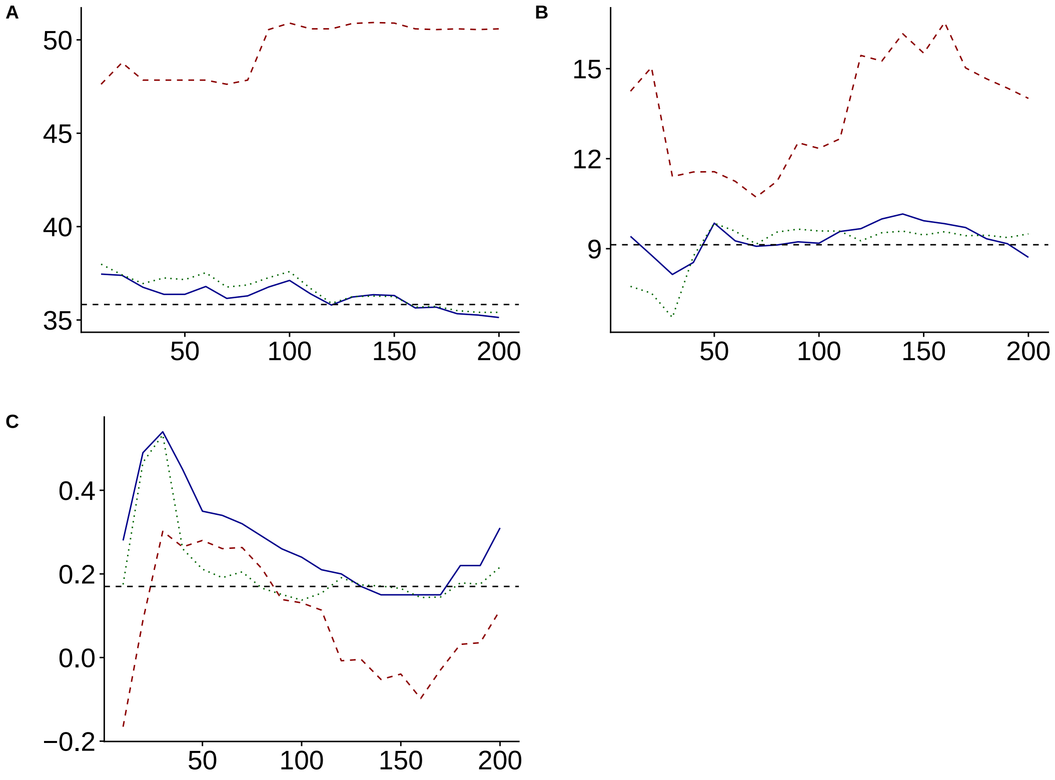


Figure 6.21.: Line-charts of parameter estimates for location (A), scale (B) and shape (C). The different methods are distinguished as QQ-method (blue, solid), as moment-method (red, dashed) and MLE-method (green, dotted). The known location, scale and shape parameters are marked by the horizontal lines. The x-axis shows the sample size of the randomly generated sample.

Lastly we conducted a similar simulation where we fix the amount of how many times we repeat the experiment $N = 100$ and try estimating the location, scale and shape parameters for $r = 2, 3, 5, 10$, with sample sizes $T = 20, 50, 100, 200$, generated from $G_{\hat{\gamma},\hat{a},\hat{b}}$.

Again we observe in Figure 6.24 that with growing sample size our estimations are within a smaller interval. We can also see that the overestimation of the QQ-method is due to an overestimation of the location parameter, since the shape and scale parameter do not differ much between the QQ-method

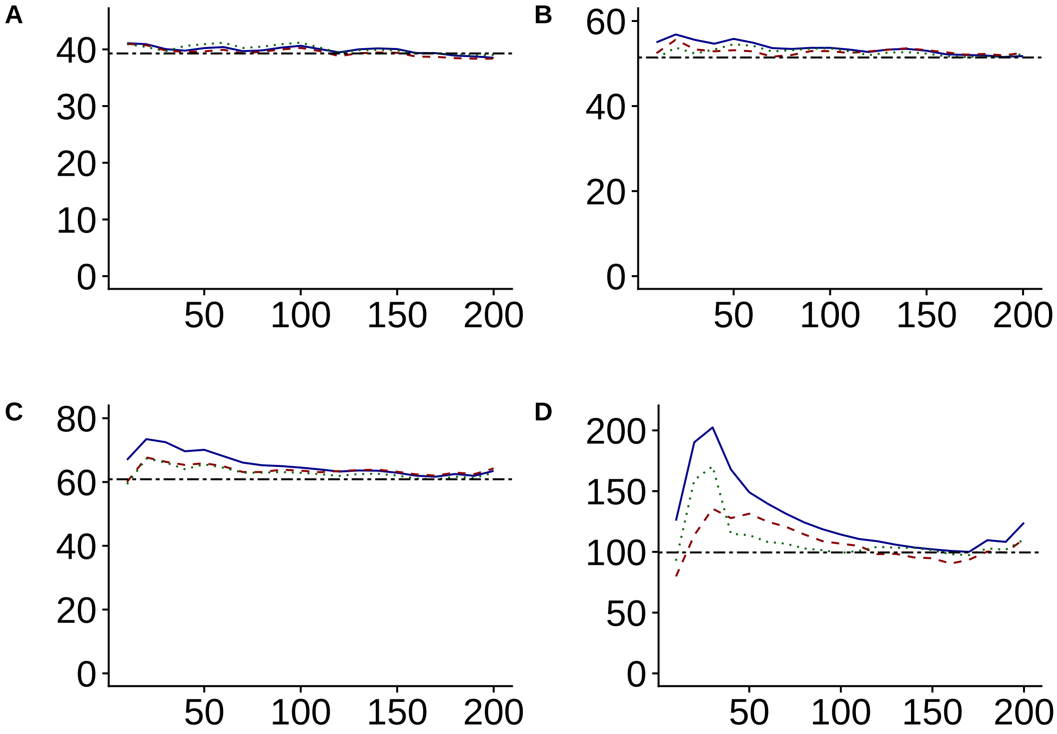


Figure 6.22.: Line-charts of return-level estimates for return periods $r = 2$ (A), $r = 5$ (B), $r = 10$ (C) and $r = 100$ (D). The different methods are distinguished as QQ-method (blue, solid), as moment-method (red, dashed) and mle-method (green, dotted). The corresponding quantiles of the known extreme value distribution are marked by the horizontal line. The y-axis shows the precipitation levels and the x-axis the sample size of the randomly generated sample.

and the MLE method.

Algorithm 4 Simulation Algorithm B

```
Given $\hat{\gamma}, \hat{a}, \hat{b}$ and $T, N \in \mathbb{N}$
**for** $i = 0, i < N$ **do**
    Construct random sample $X_1, \ldots, X_T \in G_{\hat{\gamma},\hat{a},\hat{b}}$
    Calculate $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$
    Use $\hat{\gamma}_i, \hat{a}_i, \hat{b}_i$ to calculate the return level $r_i^*$
    i++
Calculate the boxplot from either $r_i^*$ or one of the parameters
```

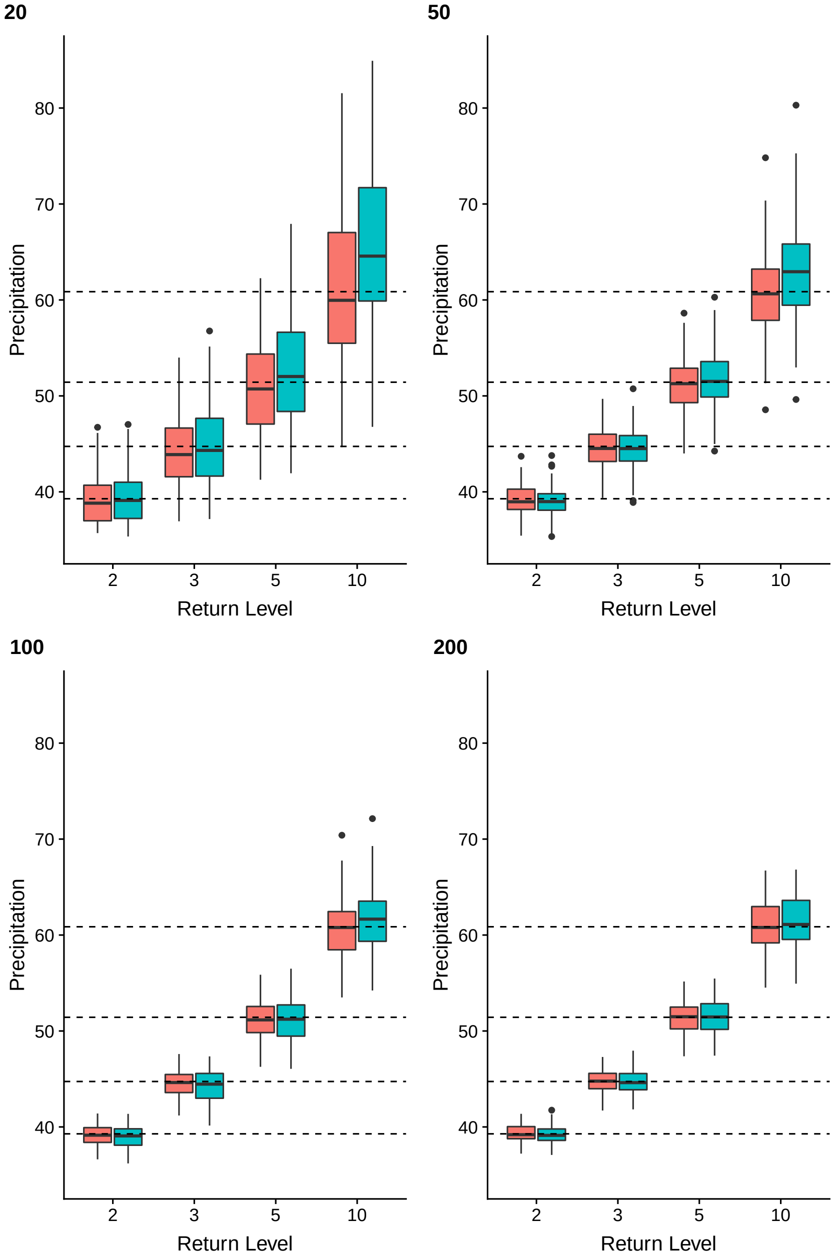


Figure 6.23.: Boxplot-pairs for the MLE method (left,red) and the QQ-method (right, green) for return levels r=2,3,5,10 repeated N = 100 times with sample sizes n = 20,50,100,200.

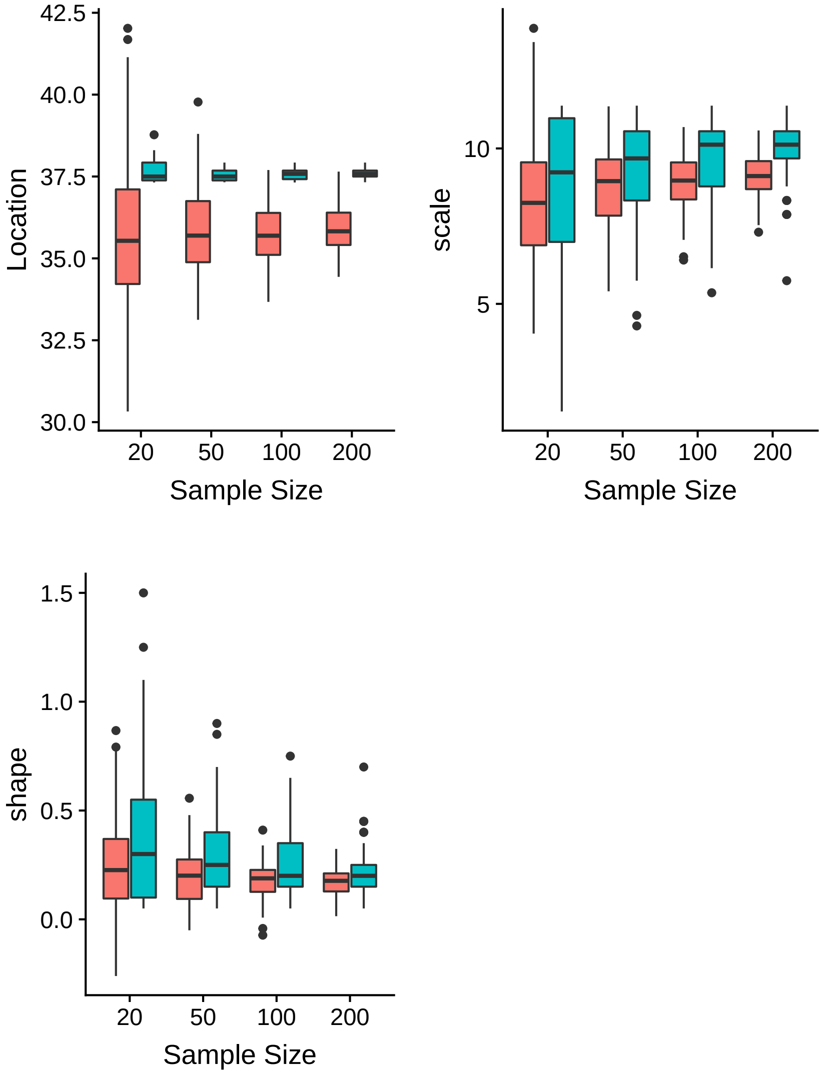


Figure 6.24.: Location, shape and scale boxplot-pairs for the MLE method (left,red) and the QQ-method (right, green) repeated N = 100 times with sample sizes n = 20,50,100,200.

7. Conclusion

In this thesis we have looked at several different approaches to compare the precipitation return levels from INCA with levels from OEKOSTRA and then to generate and assess our own estimates with raw data obtained from the INCA system. Contrary to our first assumption we were able to show that the maps and return levels provided to us were not as badly correlated with each other as expected. Even for smaller durations, where estimation is more prone to errors, the maps were fairly similar.

Next we approached the estimation of the return levels on our own and obtained comparable results. Our estimates correlated very well with both the INCA and the OEKOSTRA and we even obtained a higher correlation between OEKOSTRA and our return level estimates for the duration $d = 3$h, than the correlation from the given INCA return level estimates.

We choose a bootstraping approach to construct confidence intervals for our estimates in order to validify our results. To our surprise the intervals we obtained covered INCA estimates less often than the return level estimates from OEKOSTRA. Considering that our estimators are based on the raw data from INCA and not from OEKOSTRA, this is an intriguing finding and something that should be investigated further.

Lastly, we showed with simulations and bootstrapping, that our methods for calculating the return levels are indeed performing reasonably well, even for smaller sample sizes, but we also saw that the precise estimation of the real parameters of the underlying extreme value functions can be quite difficult with the short observation period of 13 years.

Appendix

Appendix A.

Code documentation

A.1. Overview

This chapter will give an overview of the R code used to perform all the calculations in the previous chapters. For ease of use the code was split into several R-script files, where each such file aims to encapsulate only one method of calculation/evaluation. Additionally a utility file for miscellaneous helper functions was created and a main file where the rest of the files are imported and then executed.

The code is highly dynamic and can be executed for every duration d or return level r that can be specified in the *Main.R* file and will automatically generate and export all according plots.

The R-scripts are as follows:

- *Main.R*: The entry point for all calculations. It imports all the other scripts and functions and executes them. Important variables such as the return period r and the duration d are defined here.
- *BasicPlots.R*: Provides maps, scatterplots, regressionplots and histograms that have been used in the beginning of our empirical analysis.
- *Util.R*: Contains helper functions for importing the map of Austria, importing libraries, importing all other R-scripts and server small helpful features.

- *Stations.R*: Calculates the closest INCA points to all the stations from the OEKOSTRA data and, since the calculation takes several minutes, exports it into a *txt* file. (See Section 2.1)
- *StationsFromOrography.R*: Calculates the closest INCA points from the the orography file (See Section 2.2.1) to all the stations from the OEKOSTRA data and is again exported into a txt file. Note that this was necessary since the raw precipitation data was organized differently then their respective return level estimates.
- *IncaReturnLevels.R*: Imports return level estimates from the data explained in Section 2.2.2, given duration and return period.
- *RadiusAnalysis.R*: Smooths the data for a given radius as described in Section 3.2.
- *ClusterAnalysisHeight.R*: Groups data into sea-level groups as described in Section 3.3.
- *GridAnalysis.R*: Constructs a grid over Austria as described in Section 3.4

A.2. Code examples

A.2.1. Empirical analysis

The *Main.R* script defines parameters such as colors, exportPaths, duration and returnLevel and generates plots with the help of the other files

```r
# Generate data.frame
df = loadData(oekoFilePath, incaFilePath, duration, returnPeriod)
df.complete = df[complete.cases(df), ] # dataframe without NAs

# Export basic plots
durations = c(3,9,12,18,24,48,72,96)
for (duration in durations) {
  print(p("Plots for: [", returnPeriod, ",", duration, "]"))
  df = basicPlots(duration, returnPeriod)
}
```

After the basic plots the main script continues with the approaches from Section 3.1.

```
# Height method
groups = 200
exportHeightCorrelationPlot(df.complete, exportPath, p("completeHeightMethod",returnPeriod,"J_",duration,"h_",groups,"grp"), groups)

for (group in 1:groups) {
  customHeightPlot(df.complete, exportPath, p("completeHeightMethod",returnPeriod,"J_",duration,"h_",groups,"grp"), group)
}

# Radius method
max = 1
step = 0.01
exportRadiusCorrelationPlot(df, exportPath, p("completeRadiusMethod",returnPeriod,"J_",duration,"_max",max,"_step",step), max, step)

for (i in 1:(max/step)) {
  current = i * step
  customRadiusPlot(df, exportPath, p("completeRadiusMethod",returnPeriod,"J_",duration,"_max",max,"_step",step), current)
  customRadiusMap(df, exportPath, p("completeRadiusMethod",returnPeriod,"J_",duration,"_max",max,"_step",step), current)
}
```

The core of the radius analysis is calculating the stations within range, which is done by this method:

```
# df is the dataframe containing all data and col is the specified column (array) that contains the precipitation levels.
getClosest <- function(radius, col, df) {
  lat <- df$Latitude
  lon <- df$Longitude
  res <- c()
  for (i in 1:length(lat)) {
    closest <- c()
    for (j in 1:length(lat)) {
      eval <- (((lat[j] - lat[i])^2 + (lon[j] - lon[i])^2) <= radius)
      if(eval) {
        closest <- c(closest, col[j])
      }
    }
    res <- c(res, mean(closest))
  }
  return(res)
}
```

The core of the sea-level analysis is calculating the groups, which is done by the method:

```
# df is the dataframe containing all data, col is the specified column (array) that contains the precipitation levels and n is the amount of groups.
getCluster <- function(df, n, col) {
  len <- length(df$Height)
  step <- len/n
  cluster.df <- data.frame(min = integer(), mean = integer(), max = integer(), minHeight = integer(), maxHeight = integer())
  for (i in 0:(n-1)) {
    clusterPre <- c()
    clusterHeight <- c()
    for (j in (step * i):(step * (i+1))) {
      if (j != 0) {
        clusterPre <- c(clusterPre, col[j])
        clusterHeight <- c(clusterHeight, df$Height[j])
      }
    }
    newRow <- data.frame(min = min(clusterPre), mean = mean(clusterPre), max = max(clusterPre), minHeight = min(clusterHeight), maxHeight = max(clusterHeight))
    cluster.df <- rbind(cluster.df, newRow)
  }
  return(cluster.df)
}
```

The core of the grid analysis is calculating the grid-segments, which is done by the method:

```r
# i and j are the indices of the grid, df is the dataframe containing all data and n is the dimension of the grid.
getGridPoint <- function(i, j, n, df) {
  res <- c()
  lonDomain <- range(df$Longitude)
  latDomain <- range(df$Latitude)
  lonStep <- (lonDomain[2] - lonDomain[1]) / n
  latStep <- (latDomain[2] - latDomain[1]) / n

  lowerLon <- lonMin + lonStep * i
  upperLon <- lonMin + lonStep * (i+1)
  lowerLat <- latMin + latStep * j
  upperLat <- latMin + latStep * (j+1)
  for (k in 1:length(df$Station)) {
    lon <- df$Longitude[k]
    lat <- df$Latitude[k]
    if (lon <= upperLon && lon >= lowerLon) {
      if (lat <= upperLat && lat >= lowerLat) {
        res <- c(res, k)
      }
    }
  }
  return(res)
}
```

A.2.2. Estimation of return levels

The code for the estimation of the return levels was mostly decoupled from the *main.R* and all its inherent scripts, because it uses the raw data described in Section 2.2.1. We only needed the other datasets in a final step to compare our results to the given `INCA` and `OEKOSTRA` estimators.

The core functionality for estimating return levels, that is implementing the estimators is done in the *Estimators.R* and the *QQPlot.R* script.

Building on those two files the *ReturnLevelFunctions.R* contains more complex routines that calculate all the return levels and confidence intervals that we described in Section 5.

Everything is then put together in the *ReturnLevels.R* script where functions from *Estimators.R*, *QQPlot.R* and *ReturnLevelFunctions.R* are executed.

One of the most important and most expensive steps is to calculate the maximum precipitation that occurred during a year over a specified duration *d*.

```
getPrecipitationWithDuration <- function(durationMin, data) {
  k = durationMin / 15
  n = length(data)

  res = 0
  for (i in 1:k) {
    res = res + data[i]
  }

  curr_sum <- res
  for (i in (k+1):n) {
    curr_sum = curr_sum + data[i] - data[i - k]
    res <- max(res, curr_sum)
  }
  return(res)
}
```

The functions to calculate the moment estimator simply implement the formulas stated in the theory Section 3.

```
  M <- function(data,n,k,j) {
    sum <- 0
    for (i in 0:(k-1)) {
      x1 <- getOrderStatistic(n-i,data)
      x2 <- getOrderStatistic(n-k,data)
      sum <- sum + (log(x1) - log(x2))^j
    }
    res <- 1/k*sum
    return(res)
  }

  getMomentEstimator <- function(data,k) {
    n <- length(data)
    m1 <- M(data,n,k,1)
    m2 <- M(data,n,k,2)
  est <- m1 + 1 - 1 / 2 * (1 - (m1^2) / m2)^(-1)
  return(est)
}
```

The qq-method was calculated by the following piece of code

```
getExtremeQuantile <- function(gamma,p) {
  if (gamma == 0) return(log(1/log(1/p)))
  return((( 1 / log(1 / p) )^gamma - 1) / gamma)
}

getQQPlotQuantiles <- function(n, gamma) {
  res <-c()
  for (i in 1:n) {
    p <- i/(n+1)
    res <- c(res,getExtremeQuantile(gamma,p))
  }
  return(res)
}
```

A.2.3. Simulation

The code for the simulation is separated into two files, *Simulation.R* and *SimulationFunctions.R*. The structure here is similar as before, main steps are executed in the *Simulation.R* and helper functions are loaded from *SimulationFunctions.R*.

Most functions here build on existing ones, for example to calculate return levels and estimate the parameters and are only there to rearrange the data in a useful way for our simulations. The only main addition is the generation of simulated data from an extreme value distribution, for which we used the *extRemes* package and its *revd* function.

A.2.4. Database

In order to work with around 60 GB worth of precipitation data, we had to introduce some helpful files to speed up our calculations. We have therefore extracted the raw 15 minutes interval precipitation data into folders named ″yyyymm″, where ″yyyy″ is a year from 2004 to 2017 and ″mm″ is a month from 01 to 12. Each of those folders contains 917 csv files that enlist the precipitation that occurred at a specific station in the year ″yyyy″ and the month ″mm″.

We have also exported data into several files that contain data such as the nearest `INCA` points to each station, the maximum precipitation over a duration d for each year, return level estimates with different methods, bootstrapped data and so on.

Bibliography

Beck Bica, Schellander and Zingerle (2013). *Hochaufgelöste Starkregentabellen*. Zentralanstalt für Metetrologie und Geodynamik (cit. on pp. 2, 10).

Haan, Laurends de and Ana Ferreira (2006). *Extreme Value Theory - An Introduction*. Springer (cit. on pp. 31, 33, 39, 41, 42, 44–47, 50).

Rényi, Alfréd (1953). *On the theory of order statistics* (cit. on p. 39).

Tawn, Jonathan A. (1988). *An extreme-value theory model for dependent observations*. URL: https://www.sciencedirect.com/science/article/pii/0022169488900376 (cit. on p. 51).