The Danish Meteorological Institute

DMI Report 21-27 Occurrence of cloud bursts in Denmark obtained from daily precipitation sums.

Final scientific report of the 2020 National Centre for Climate Research Work Package 2.2.2, EkstremRegn

DMI Report
18 January 2021

By Peter Thejll, Flemming Vejen, Torben Schmith and Bo Christiansen

The Danish Meteorological Institute

Colophon

Serial title	DMI Report
Title	DMI Report 21-27 Occurrence of cloud bursts in Denmark obtained from daily precipitation sums.
Subtitle	Final scientific report of the 2020 National Centre for Climate Research Work Package 2.2.2, EkstremRegn
Author(s)	Peter Thejll, Flemming Vejen, Torben Schmith and Bo Christiansen
Other contributors	
Responsible institution	Danish Meteorological Institute
Language	English
Keywords	Cloud burst, precipitation
URL	https://www.dmi.dk/publikationer/
Digital ISBN	978-87-7478-701-3
ISSN	2445-9127
Version	18 January 2021
Website	www.dmi.dk
Copyright	DMI (Image of clouds. Source: Pixabay)

EkstremRegn

Peter Thejll, Flemming Vejen, Torben Schmith og Bo Christiansen

March 15, 2021

Contents

1 Abstract

Using precipitation data from 9 DMI 'synop' stations we statistically investigate whether daily sums of precipitation can be used as a proxy for the occurrence of cloudbursts.

The daily sum of precipitation data are based on a digitisation effort of DMI annual summaries of danish weather conditions (1917-1960s), and already digitally available data from the same, or nearby, stations found in the DMI database. The scope is to eventually have 100-200 such long data records available for studies of precipitation in general, stretching over the past 100 years. The present report deals with 28 such stations with data from 1917 to the modern day.

We first apply univariate logistic regression to the question. We show that a simple approach based on a cutoff in daily precipitation sum near 45 mm can be used as a proxy for an indicator of "50% chance of a cloudburst that day". Robustness of results is demonstrated. This rule is very simple but correctly captures 17% of the actual cloudbursts.

We then introduce one more predictor - the maximum temperature observed each day, and show that additional skill is introduced in a multivariate logistic regression approach, but that the model is not perfect, i.e. there are both 'false positives' and 'false negatives' in the multivariate model predictions.

Using a fixed daily precipitation sum (DS from now on) limit of 45 mm we study the rate of occurrence of such extreme events across Denmark. In future extensions of the project, with a larger dataset, we can look at regional variations in the logistic relationship between DS and cloudbursts and gain further insights into possible geographical variations of this.

2 Resumé

Med hjælp af synopdata fra 9 DMI nedbørstationer undersøger vi statistik om daglige nedbørssummer kan anvendes som proxy for forekomsten af skybrud.

Nedbørsdata er blevet digitaliseret fra DMIs årsbøger for at, til sidst, kunne stille 100-200 daglige nedbørsserier til rådighed, hver ca 100 år lang. Denne rapport handler om analysen af 28 sådanne serier, alle med data siden 1917 til idag.

Først benytter vi univariat logistisk regression på spørgsmålet. Vi viser at man godt kan bruge daglige nedbørssummer, med en grænse ved 45 mm om dagen, som proxy for at der har været et skybrud den dag ved den station. Reglen er meget simpel, men fanger alligevel 17% af rigtige skybrud.

Derefter indfører vi endnu en predictor i den statistiske model - den daglige maximale temperatur, og vi kan vise at yderligere skill opnås med modellen, selvom der stadigvæk er både 'false positives' og 'false negatives' i den multi-variate model's resultater.

Ved brug af den fastsatte grænse på 45mm om dagen studerer vi den gennemsnitlige rate for skybrud i Danmark som helhed. Med et større datasæt hvor alle de potentielle serier blev digitaliseret kunne vi studere den rumlige fordeling af skybrud i Danmark.

3 Introduction

The Danish National Centre for Climate Research (Nationalt Center for Klimaforskning, NCKF) has completed its first year in 2020. It has been a source of funding for the Danish Meteorological Institute and collaborators for climate change related research during this year. The 18 work packages fall under 4 general themes:

1. Arctic and Antarctic Research
2. Climate change in the near future
3. Use of climate data
4. Support for the IPCC

The aim of the NCKF project 'EkstremRegn' is to enable use of a recently digitised set of very long (100 years) and spatially dense (eventually 200 quality-checked station records across Denmark) dataset of daily precipitation.

The approach is that with a trained Machine-Learning system we use the daily-sum data to predict the likely occurrence of cloudburst across Denmark during 100 years in the past. The length of the dataset and the, eventual, high spatial density is then used to look for regional patterns.

We show the results of inspecting precipitation data from 9 DMI 'synop' stations with hourly precipitation observations and daily sum precipitations from the same (or very nearby) stations. We develop insight into the statistical relationship between daily precipitation sums and the daily observed maximum in hourly sums (DM from now on). Logistic regression is applied and we consider whether a skill exists in predicting 'cloudbursts' from daily-sum precipitation observations. We then extend the analysis and consider multivariate models where daily maximum temperature is used, with the idea that cloudburst days are likely indicated by high temperatures whereas days with high levels of frontal rain are not.

In Section 5.1 we inspect the data and apply a simple approach in order to evaluate how well daily precipitation limits can be used to dichotomize the hourly maxima into 'cloudburst vs no cloudburst' cases. Results of this are in 6.1.

In Section 5.2 we apply logistic regression to the combined data too see how we can extend the simplified analysis of section 5.1, and also determine some uncertainty limits for the parameters from the fitted logistic model

('logit' in short) model. This can then be the foundation for the, invariable more 'noisy' approach of logit-fitting individual stations. Results are in section 6.2.

Section 5.3 looks at the properties of logit model fits to individual stations, with particular attention to the stochasticity of various results due to small-numbers statistics. Results shown in Section 6.3.

Multivariate logistic regression is introduced in section 5.4, and results shown in section 6.4.

We summarise findings in Section 7 and discuss and make suggestions for the next steps of this NCKF project.

4 Data

The core of this project are the station data made available by a concentrated digitisation effort, using annual summaries of weather observations published by the DMI from 1917-the 1960s. These data are collated with other, higher-cadence and spatially coincident series of hourly precipitation observations (so-called 'synops'). Each of the selected series can also be extended up to the modern day, using the DMI database of observations.

A set of 11 DMI synopstations, distributed somewhat evenly across the nation, were selected. 9 of these contained data for both day-sum precipitation and the maximum hour-sum of precipitation of high quality (i.e. no data gaps). The stations are listed in Table 1. Abed and Sjælsmark were excluded due to many gaps, and the analysis proceeds with the other 9.

Table 1: The 11 DMI synop-station data selected for analysis of the relationship between daily and hourly sums of precipitation.

Station id	Name	Lat. [°N]	Long. [°E]	h [m asl]
601900	SILSTRUP	56.93	8.64	4
603100	TYLSTRUP	57.19	9.95	13
607200	ØDUM	56.30	10.13	61
608200	BORRIS	55.96	8.62	25
610900	ASKOV	55.47	9.11	62
611601	STORE-JYNDEVAD	54.90	9.12	15
612600	ÅRSLEV	55.31	10.44	49
613500	FLAKKEBJERG	55.32	11.39	32
613600	TYSTOFTE	55.25	11.33	12

4.1 Digitization, and supplementation.

Long time series with daily precipitation observations have been made available for the period 1914-2010 from 28 manual precipitation stations in DMI's climate network. For the period 1914-1960, data were digitised from official monthly and annual meteorological publications (DMI, 1914-1960), while data for the period 1961-2010 were extracted from DMI's climate database.

Many of the data series are not complete for this nearly 100-year period, as there have been several closures of precipitation stations. For some of the stations, there are temporary data interruptions. However, a total of 13 of the series are in practice complete, because these stations were not closed until 2009 or 2010. For the rest of the stations, it has been necessary to supplement with data from nearby precipitation stations however, although in many cases it is only necessary to supplement with 10-15 years of data. Often, the distance to neighbouring stations is close to or less than 10 km, but in one case, however, up to almost 17 km.

Table 2 displays metadata for the 28 precipitation stations.

For the period 1914-1960, it is assumed that data have been subject to quality assurance before publication, even though no documentation for methodology is found, while data for the rest of the period until 2010 were subject to manual quality control.

Until the 1910s, rain was measured by the Fjord rain gauge, while snow was measured with a zinc bucket. In the period 1910-1925 the Fjord gauge was gradually replaced with the Danish Hellmann gauge [3], which was used in the DMI's rain gauge network until 2011.

5 Methods

The basic tool we use is that of logistic regression. Consider a data-set consisting of N pairs of observations - $d = \{x_i, y_i\}_{i=1...N}$. Let x and y both be continuous variables - in our case x are the daily precipitation sums and y the hourly maximum precipitation on the same day.

Categorise the y into two classes designated by 0s and 1s according to whether the value of y_i corresponds to a cloudburst or not:

$$Y_i = \begin{cases} 1, & \text{if } y_i \geq limit \\ 0, & \text{otherwise,} \end{cases}$$

and apply generalized linear regression[1] to the set $D = \{x_i, Y_i\}$.

[1] We use the R package *glm()* from the *stats* library.

Table 2: Metadata for precipitation stations. The table shows the length of the data series for the original station ("data"), length of supplementary data ("neighbour"), and the average distance (in km) to neighbouring stations, as well as the coordinates of the station.

Station	data	neighbour	latitude	longitude	dist.
Aakirkeby	1914-2005	2006-2010	55.074999	14.906029	10.3
Aarslev	1914-2010	-	55.307257	10.444377	-
Abed	1914-1960	1961-2010	54.828584	11.325035	1.9
Askov	1914-2010	-	55.471288	9.110088	-
Bækmarksbro	1914-2009	2009-2010	56.413013	8.308747	12.2
Bogø	1914-1970	1971-2010	54.928915	12.047868	6.7
Borris	1914-2010	-	55.957145	8.626332	-
Christiansø	1914-2009	-	55.321469	15.187529	-
Faxinge Sanatorium	1914-2006	2007-2010	55.132647	12.007733	8.7
Folkekuranst. V. Hald	1933-2009	-	56.412200	9.350214	-
Gaardbogaard	1914-2009	2009-2010	57.585913	10.349413	10.5
Grindsted Pl.	1914-1960	1961-2010	55.800805	8.916720	16.7
Gudum	1914-2010	-	55.442504	11.377927	-
Hofmansgave	1914-1991	1991-2010	55.537046	10.482184	11.8
Høgildgaard	1914-2009	2010	56.062473	8.974132	13.3
Hornum	1917-2002	2002-2010	56.833369	9.436567	5.0
Landbohøjskolen	1914-1997	1997-2010	55.681372	12.540288	10.0
Ll. Dyrehavegaard	1914-1995	1996-2010	55.940068	12.304821	9.0
Mejlgaard	1914-1956	1962-2010	56.300000	10.666700	4.6
Næsgaard	1914-1994	1994-2010	54.869402	12.115357	5.4
Norringure	1914-1988	1989-2010	56.257064	10.035784	5.0
Ribe	1914-2010	-	55.328986	8.735823	-
Skanderborg	1917-2003	2003-2010	56.044383	9.922684	8.3
Søndersted	1914-1983	1983-2010	55.626386	11.603678	12.1
Tvingstrup	1914-2005	2005-2010	55.913873	9.928382	8.1
Tylstrup	1914-2010	-	57.184903	9.954229	-
Vestbirk	1914-2010	-	55.972129	9.706898	-
Vestervig	1914-2010	-	56.763663	8.320653	-

The logistic model expresses the probability of Y being 1, p, as a function of x. A linear relation is assumed between the logarithm of the odds, $p/(1-p)$, and x:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot x_i. \tag{1}$$

Interpreting the β is not as straightforward as in ordinary least square regression, but β_0 is referred to as the 'Intercept' and β_1 as the 'Slope'. Interpretation of probabilities from the fitted model proceeds via the definition of the log odds. For instance, the value on the x-axis, $x_{1/2}$, that corresponds to a 50% chance of x_i corresponding to the 0s class is

$$x_{1/2} = -\frac{\beta_0}{\beta_1}. \tag{2}$$

In this point the slope obtains its maximum value $\beta_1/4$. Note also that the logistic curve is symmetric about this point, $p(x_{1/2}+z) = 1 - p(x_{1/2}-z)$. Figure 1 shows an example of a logistic regression. The points are the Y_i values, here slightly 'jittered' for illustration purposes. The blue curve is the fitted model. The red vertical line indicates the position of $x_{1/2}$. The basic idea of applying logistic regression to these data is to generate a proxy for cloudbursts. It will be used on datasets consiting only of daily precipitation sums in a probabilistic manner – i.e. any day-sum over $x_{1/2}$ will be designated as a cloudburst day and studies regarding trends and geographic patterns will be based on this proxy and its classification of days.

5.1 A first look at the synop data

Logistic regression enables answers to questions such as 'is it likely that a cloudburst occurred on this day'? It requires a dichotomous data-set (cloudburst/no cloudburst) as the outcome and something else as the input (e.g. "the total precipitation on this day").

We proceed as follows:

- Since the data we have are several years worth of daily pairs of 'sum of precipitation this day' and 'largest 1-hour sum of precipitation observed' we do not really have data required for identifying 'cloudbursts' since the DMI definition of this quantity relates to whether at least 15 mm of precipitation was collected over a half-hour interval. We shall assume that useful results can still be obtained by simply assuming that "15 mm or more during one hour" also is an interesting 'almost cloudburst-like' quantity.

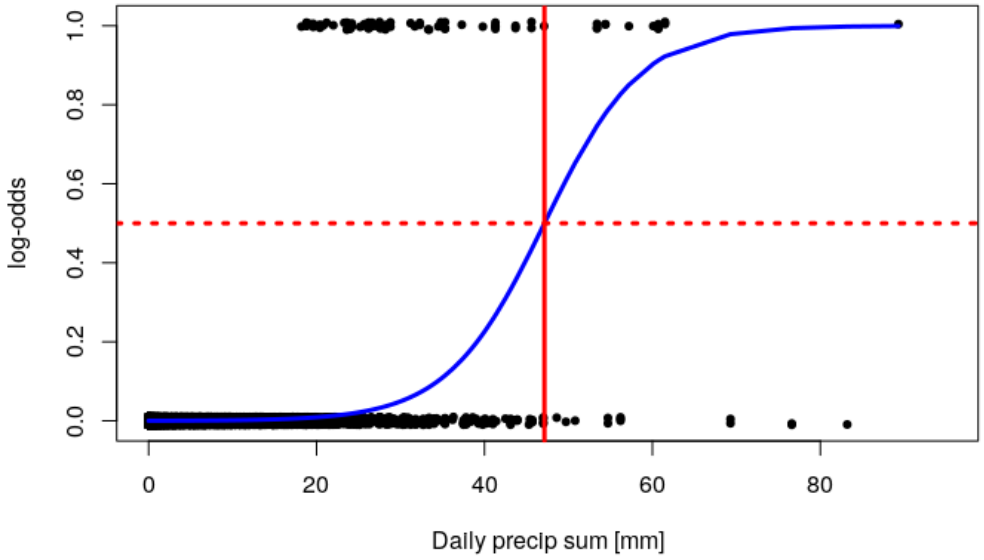

Figure 1: Example of a logistic regression on synop data, y values have been classified as 0s or 1s (jittered slightly for clarity) on the basis of some user-specified limit (here hourly maximum precipitation of 15 mm). The logit model has then been fitted to the x_i, Y_i values, and the resulting parameter estimates used to calculate $x_{1/2}$. The result is that the logistic regression approach suggests that a cutoff in daily precipitation at $x_{1/2}$ is a proxy for the occurrence of cloudbursts.

- All data from all selected stations are combined in order to limit stochasticity
- Perform a preliminary investigation of how a limit set in daily-sum precipitation can be used to project likely occurrences of cloudbursts.

Cloudbursts in Denmark mainly occur during the warmer months (May-September) so we inspect the month-distribution of 'large daily precipitation sums' to see if this is realistically distributed. Figure 2 shows the distribution over months for daily precipitation sums for various limits.

5.2 Logistic regression on the combined synop data

We next use the combined data, from all the synop stations, to perform a logistic regression, in order to see how it is done, but also to provide a reference against which to compare similar results from individual stations which are bound to be more stochastic.

We proceed as follows

- Using the 15 mm limit on precipitation hourly maximum sum we dichotomize all the data from the combined synop station into 0's and 1's
- we apply logistic regression, and inspect the logit model coefficients and their uncertainties
- Above, we showed that the 50%-point on the fitted logit model can be calculated from the model parameters (see section 5), and we shall consider this parameter, and its sampling uncertainty as it has direct bearing on the possibility of using a cutoff level in daily precipitation sum as a proxy indicating possibility of cloudbursts

5.3 Application to individual synop stations

We consider the application of the logistic model determined above to individual stations' data, as well as taking a first look at the spread in logit model parameters across Denmark.

- Fit the logit model to data from each station
- Show the logit model parameters on maps.

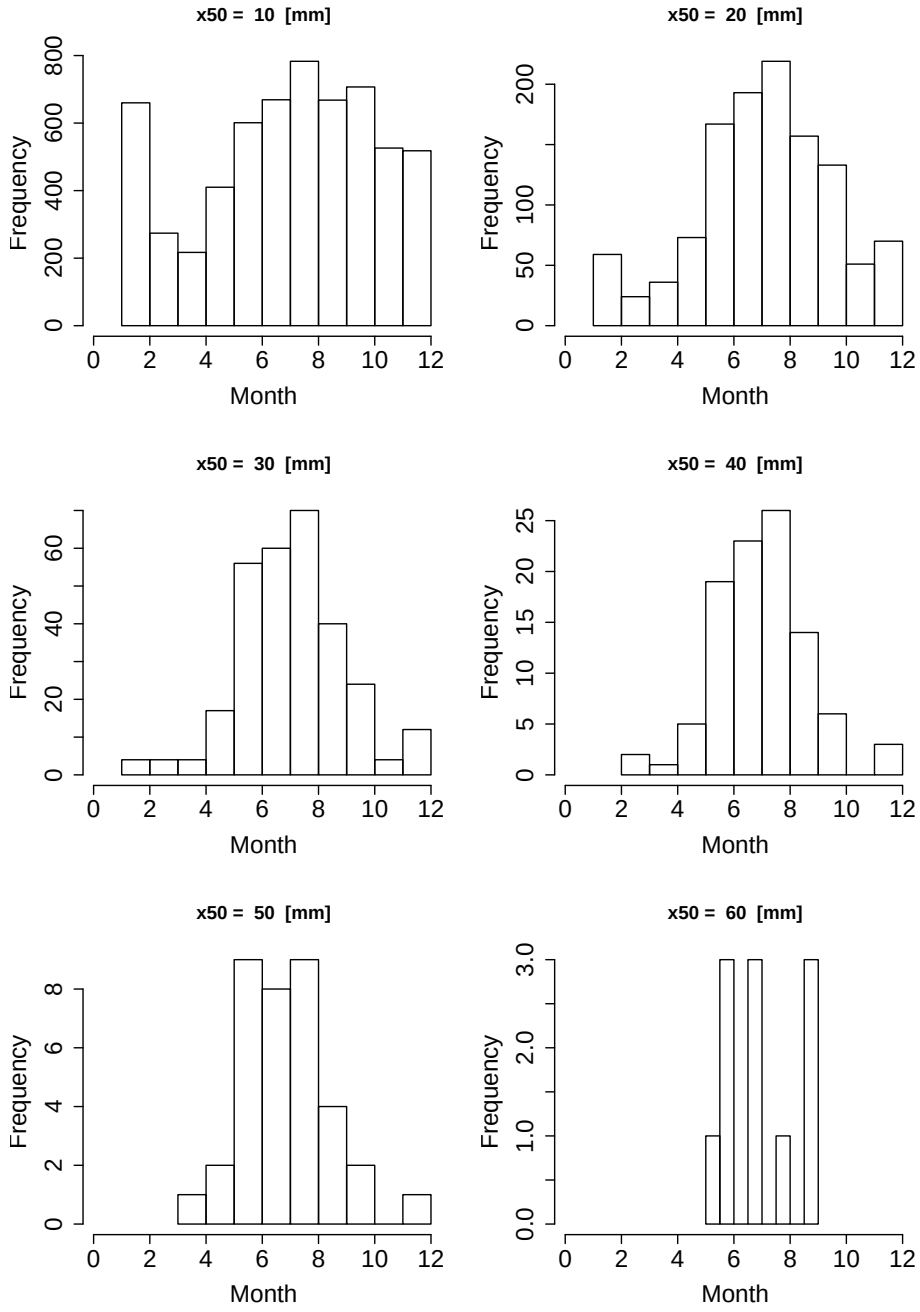

Figure 2: Distribution of the month of occurrence of daily precipitation sums over different limits. For a limit of 30 and above, the distribution of 'large day-sum precipitations' is distributed similarly to how cloudbursts are distributed. See Figure 3 for the climatology of cloudbursts in Denmark. "x50" indicates the limit on precipitation day-sum used in each panel.

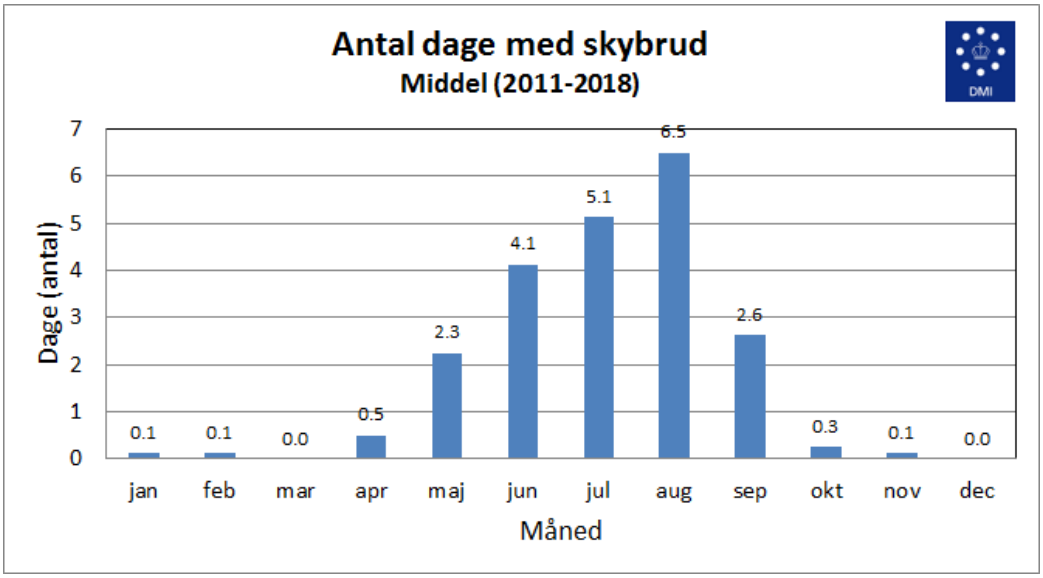

Figure 3: The climatology of observed, real, cloudbursts in Denmark. From DMI.

5.4 Multivariate logistic regression

We assume that the predictor variables can be related linearly to the log-odds l of a combination of the response variables x and z:

$$l = \log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot z \tag{3}$$

Consider again the crossover-point where $p = 0.5$ – there $l = ln\frac{0.5}{0.5} = 0$. Solving, we find:

$$x_{1/2} = -\frac{\beta_0 + \beta_2 \cdot z}{\beta_1} = -\frac{\beta_0}{\beta_1} - \frac{\beta_2}{\beta_1} \cdot z, \tag{4}$$

which is the equation of a line in the $x_{1/2}, z$-plane.

6 Results

6.1 The preliminary analysis

Figure 4 is a first look at the combined data from the 9 synop stations. The vertical red line is at the 15mm/hour limit. The blue points indicate all

the data above this limit. The blue line is a robust[2] linear regression to the blue points. The thick dashed blue line illustrates how a rule for projecting cloudbursts could be formulated. The red and thick dashed blue line intersect at 15, 24 mm. One could thus say 'daily sums above 24 mm indicate a likely cloudburst on that day', but inspection of the distribution of black and blue points shows that not all such days experienced an hourly maximum precipitation even close to something like a cloudburst - many days with a lot of precipitation are simply days where it rained all day at a moderate rate. In this report we shall explore other cutoffs on the daily hourly max precipitation – see Figures 8 and 5 below. We shall then use the thin dashed blue line in Figure 4 as the cutoff on daily-sum precipitation that designates a cloudburst. We therefore ask, 'given various limits of daily precipitation, what fraction of those days experienced an hourly maximum sum above 15 mm?'. Figure 5 shows the results. The error bars are generated by data re-sampling (with replacement). The graph indicates that at a day-sum cutoff near 45 mm (but probably stretching from just above 40 and upwards) there is a 50% chance that the days thus indicated also experienced a maximum hourly precipitation sum of 15 mm or above - or a cloudburst. We do note, however, that the results of our 'first look' analysis on the combined data have quite a stochastic uncertainty – it is actually only possible to say, using all the data, that we see a lower limit in daily precipitation, near 45 mm, that indicates a 50% chance of a cloudburst that day.

We note (Figure 2 that the distribution over months of large daily precipitation sums becomes similar to the distribution of real cloudbursts when a limit no lower than 30 mm is used. This is encouraging for the basic strategy of using day-sum as a proxy for cloudbursts.

6.2 Results of the logistic regression on the combined synop data

Table 3 shows re-sampling results for the logit model fitted to the combined synop station data. We see that using a cutoff on daily sums of precipitation near 45 mm is indeed validated. This was an outcome of the simple analysis in section 5.1. Now we have estimates of the uncertainties to expect due to sampling.

Pursuing the theme of stochasticity we inspect how large a fraction of the data correspond to a cloudburst, at the 15 mm limit. To obtain the standard deviation we collect re-sampling results. Figure 6 shows that about 0.18% of the data is in the cloudburst category with a standard deviation of about

[2]For robust regression we use the *rlm()* package in the *stats* library

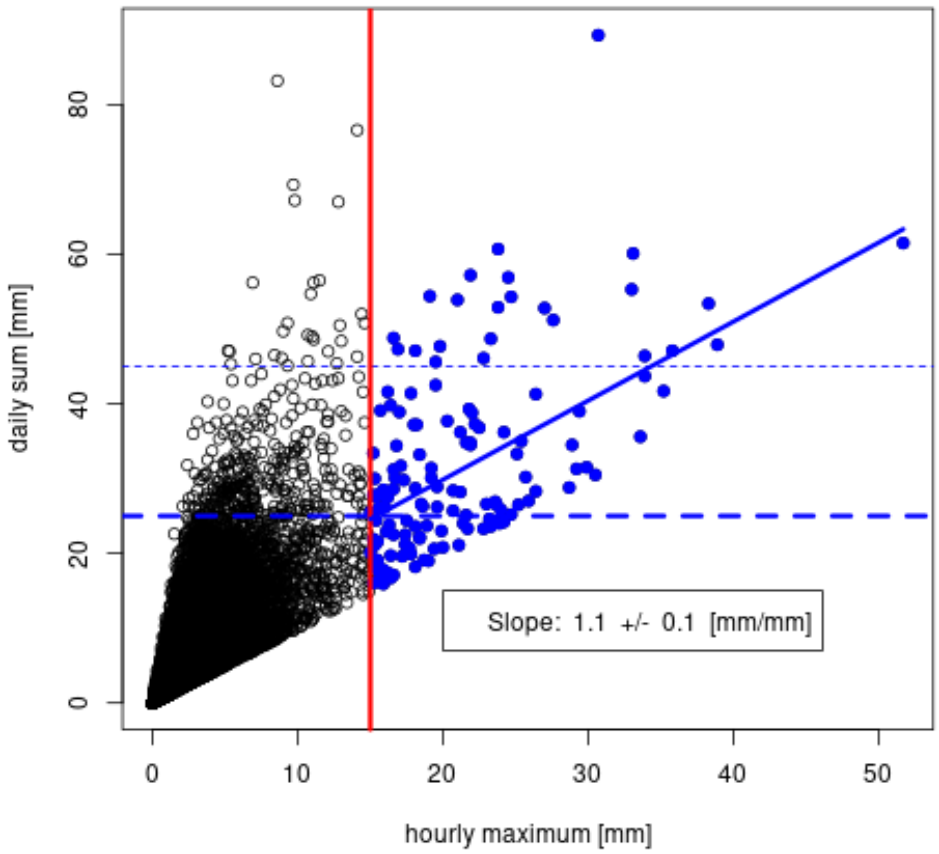

Figure 4: A limit at 15 mm day hourly precipitation corresponds to about 24 mm daily precipitation. Not all days with daily precipitation above 24 mm are cloudbursts, however. In this figure all 9 synop-station data have been combined.

0.02 percentage points. We shall inspect how this number changes as we move the analysis to individual stations.

In Figure 7 we consider the spread on the parameter $x_{1/2}$. This quantity is calculated from the logit model parameters, as explained in section 5.2, and we shall be examining the variability and uncertainty of this parameter

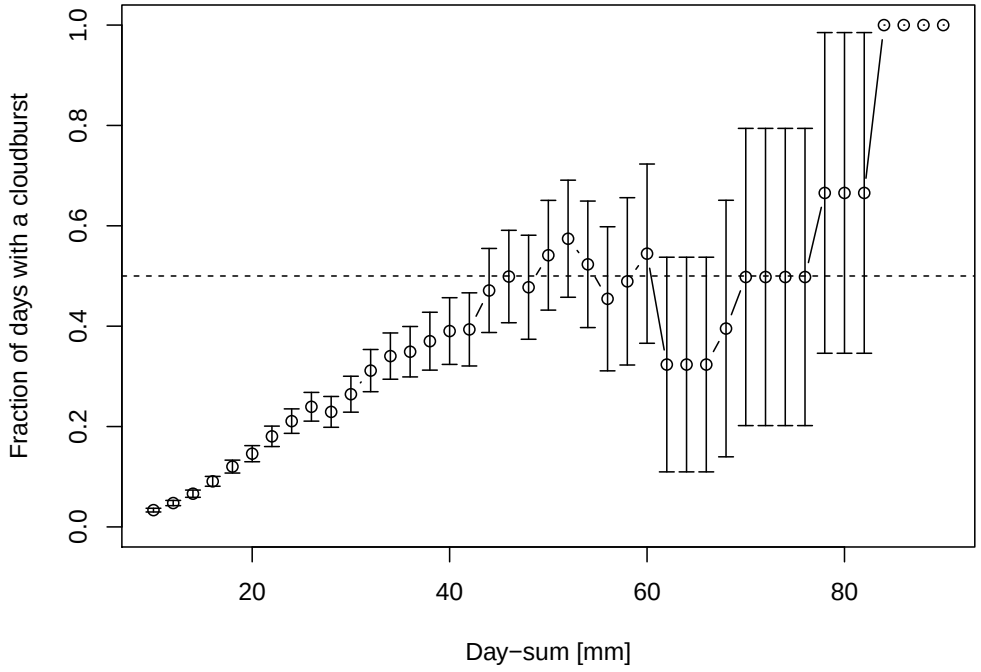

Figure 5: The fraction of days where the hourly precipitation sums exceeds 15 mm as a function of the day-sum used as cutoff. The error bars indicate $\pm$ 1 standard deviation. In this figure all 9 synop-station data have been combined.

when we go to analysis of individual stations, below.

We counted the rate of correct identification of our 45 mm cutoff on daily sum precipitation, and can use Figure 4 for that. Of the cases (blue dots) to the right of the vertical red line at 15 mm hourly max precipitation, 17% lie above a horizontal cut at 45 mm in the daily sum. This is the true-positive rate of our simplified counting method.

6.3 Results of the logistic regression on individual station data

We fitted a logit model to individual stations, using a set of cutoffs for the dichotomisation of the days into cloudburst and not-cloudburst. The cutoffs

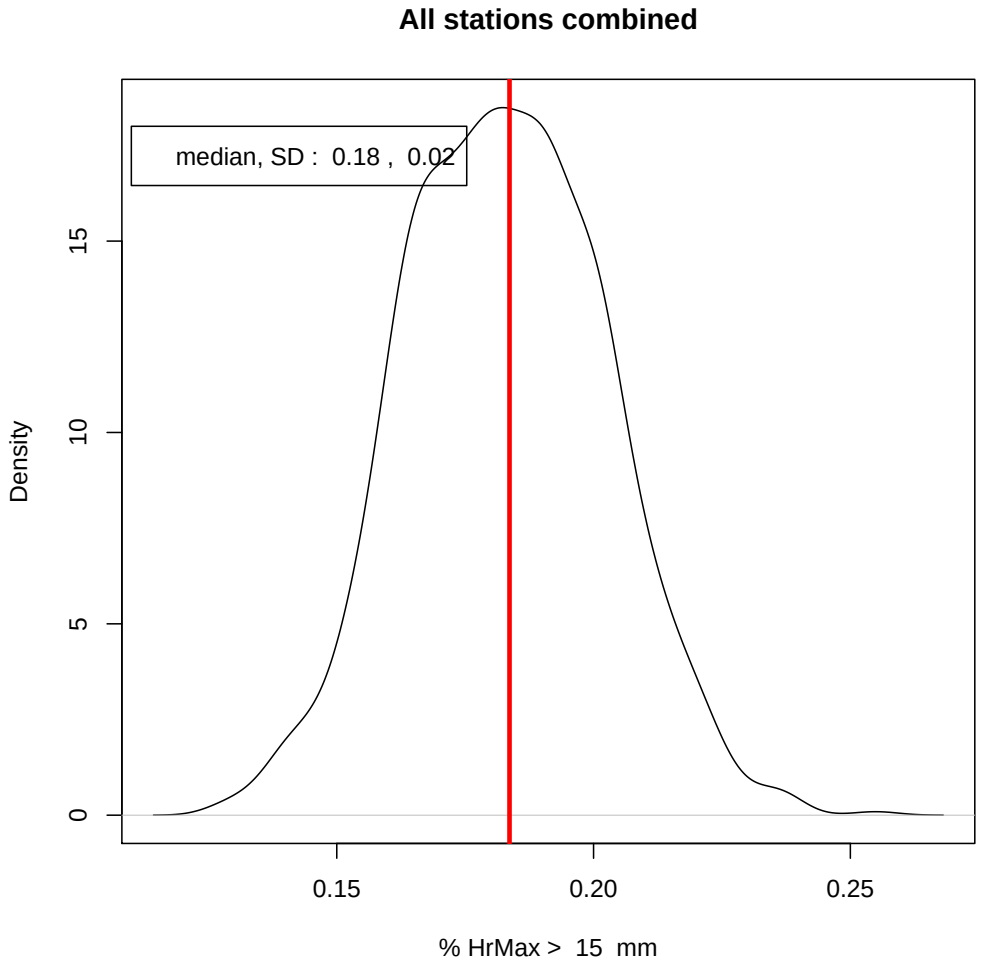

Figure 6: Resampling distribution of the fraction of days (when all 9 synop-station data are combined) with cloudbursts as defined by the ”15 mm in an hour” limit.

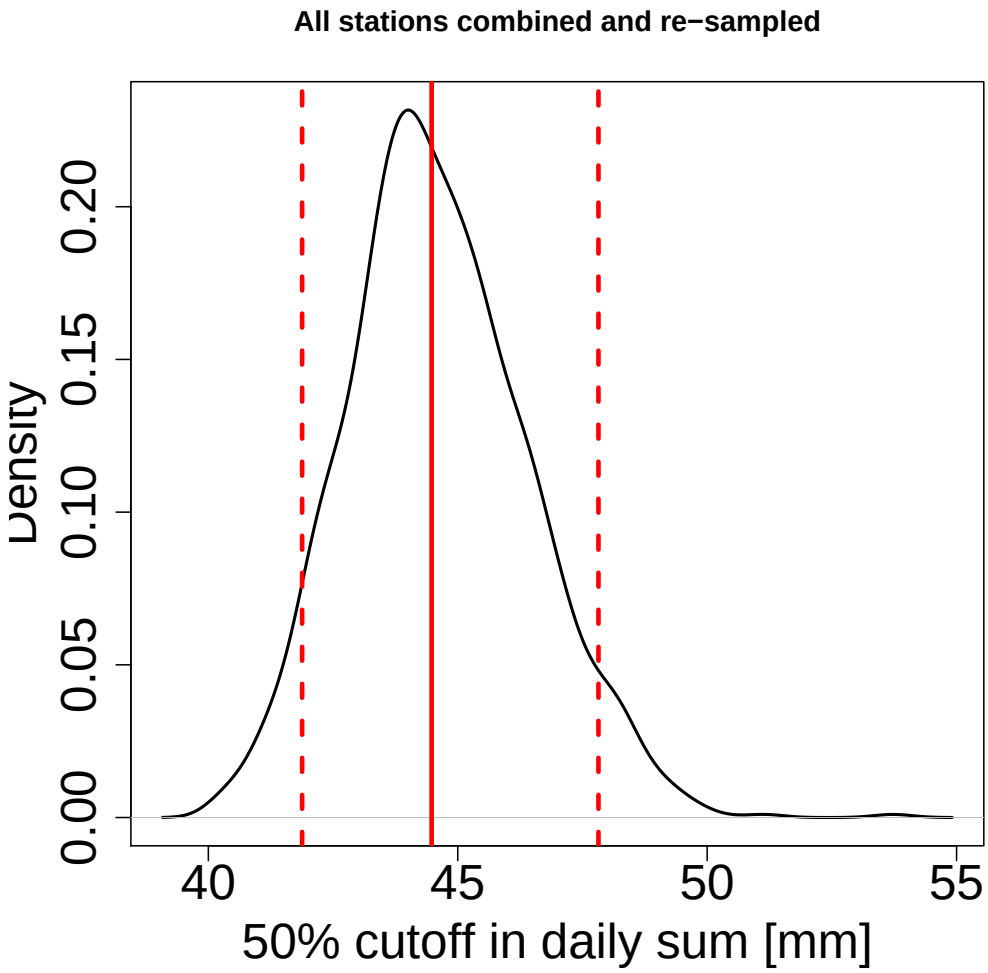

Figure 7: Re-sampling of all station data combined. The density plot of the 50% value in daily precipitation sum is shown, as calculated from the logit-model parameters. See text for more details on this calculation. The vertical lines show the 5, 50 and 95%-iles. Compare to the results of individual-station determinations of the 50% cutoff in Figure 14 .

Table 3: logit model parameters and the uncertainties. Shown are the 5, 50 and 95%-iles of the logit model *intercept, slope and 50% point in daily sum*. Re-sampling, with replacement, was performed to reveal the sampling error. Data from all 9 synop stations were combined, and days dichotomized into 0's (no cloudburst) and 1's (cloudburst) using the 15 mm maximum hourly-sum limit. A relevant daily-sum limit is, apparently, near 45 mm for a p=0.5 cutoff.

Name	
Intercept	-8.6, **-8.2**, -7.9
Slope	0.17, **0.18**, 0.20
50% point in daily sum [mm]	41.8, **44.6**, 48.1

were chosen to span the official limit that defines a cloudburst. Having no access to 30-minute data we proceed as if this was not the case, but sample a wide span of possible cutoffs. Figure 8 shows typical results.

Clearly, the logit fits depend on the cutoff used, and as the cutoff rises, fewer and fewer points are in the cloudburst category, which influences the accuracy of the fits.

We show, in Figure 9 that for a selection of cutoffs in hourly maximum precipitation a sequence of logit model parameter pairs arise. We note three things: there is a relationship between the two fitted logit model parameters with large intercepts generally implying small slopes; we also note that scatter around the regression lines increases with the cutoff in maximum hourly precipitation, being largest for the cutoff at 15 mm; finally, there appears to be a regularity in the location of the stations on the parameter-pair sequence even across a range of cutoffs.

We explore whether stations with particularly values of the logit model fit parameters are located in particular parts of Denmark. We use maps for this – See Figure 11, 12 and 10. It appears that while some parameter-values 'stand out' in a given location across the choice of maximum hourly precipitation cutoff this is not the case for all values of the cutoff. For the Intercept the stations on Fyn and near Helsingør may be outliers; for the Slope stations near Korsør and Tylstrup seem special; and for $x_{1/2}$ only Tylstrup is notable.

6.4 Results of multivariate logistic modelling

As the second regressor we now add T_{max} data from a single station in Denmark, that at Landbohøjskolen. Guided by the results above we now

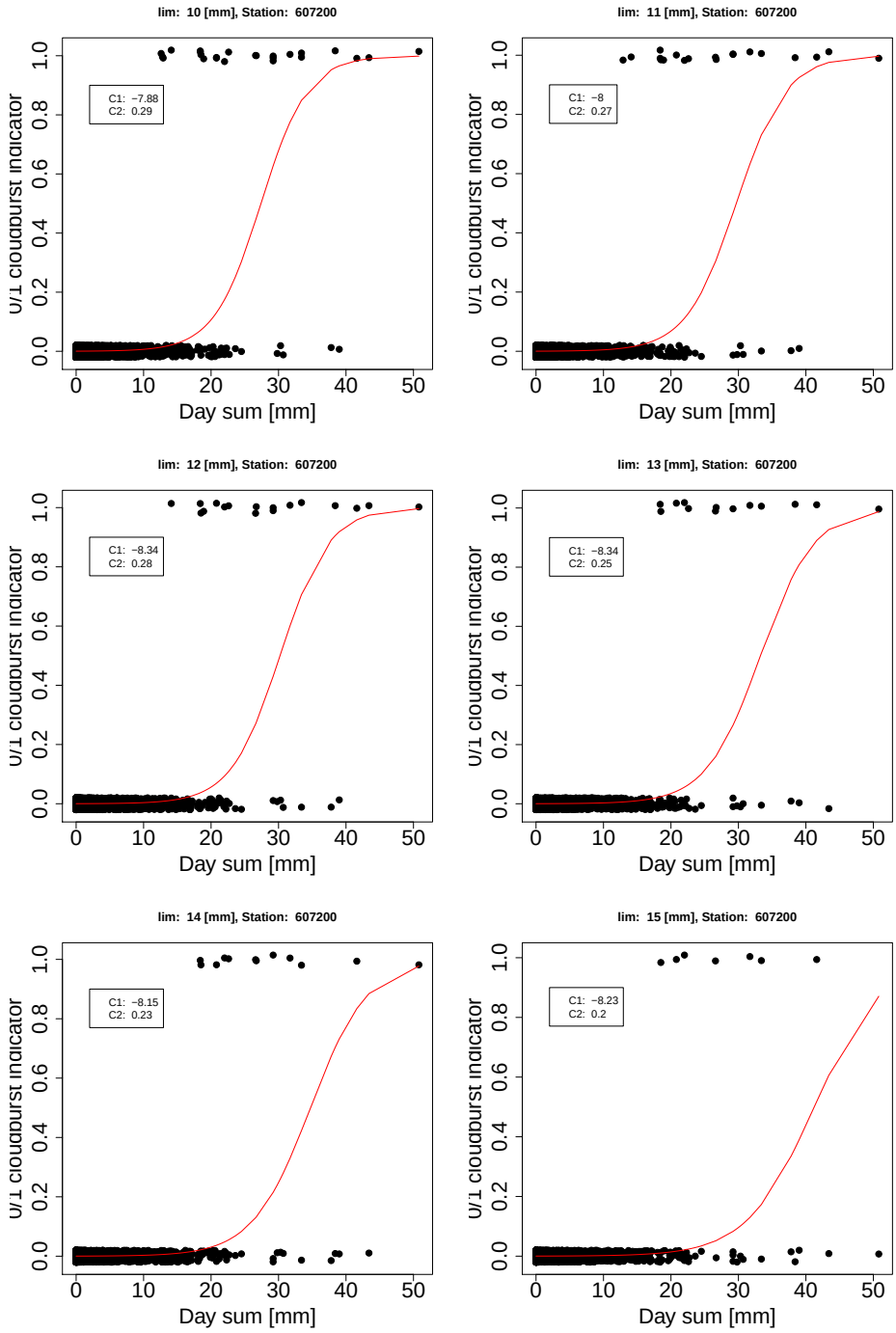

Figure 8: Logistic model fits for various cutoffs on the maximum hourly precipitation sum for a single station (607200, Ødum). The legend shows the logit model parameters found. 0/1 points have been 'jittered' for illustration purposes.

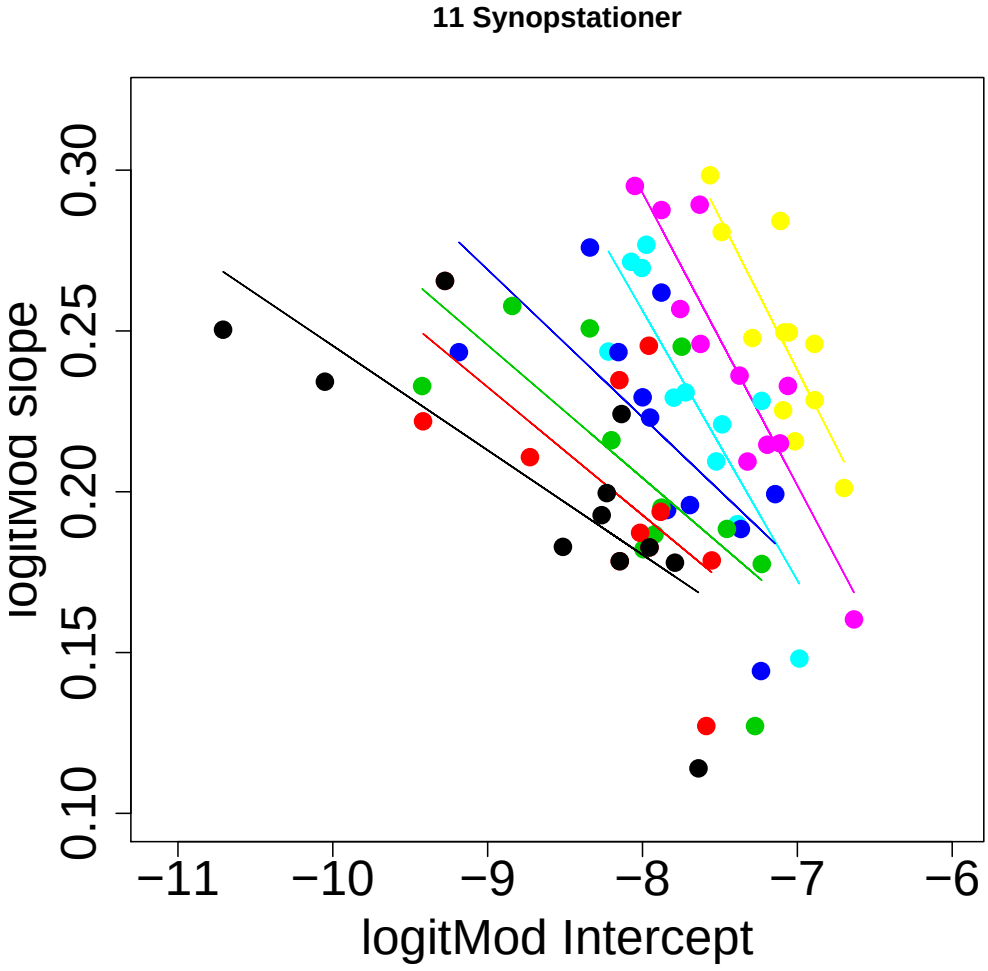

Figure 9: logit model parameters plotted, for individual stations and a selection of cutoffs for dichotomization in the hourly maximum precipitation, with yellow being for cutoff 9 mm and upwards in steps of 1 mm for the colours shown, right to left (yellow, purple, turqoise, blue, etc). Thus black corresponds to the cutoff at 15 mm. The six points at lower right all represent the same station - Tylstrup, in Northern Jutland (603100).

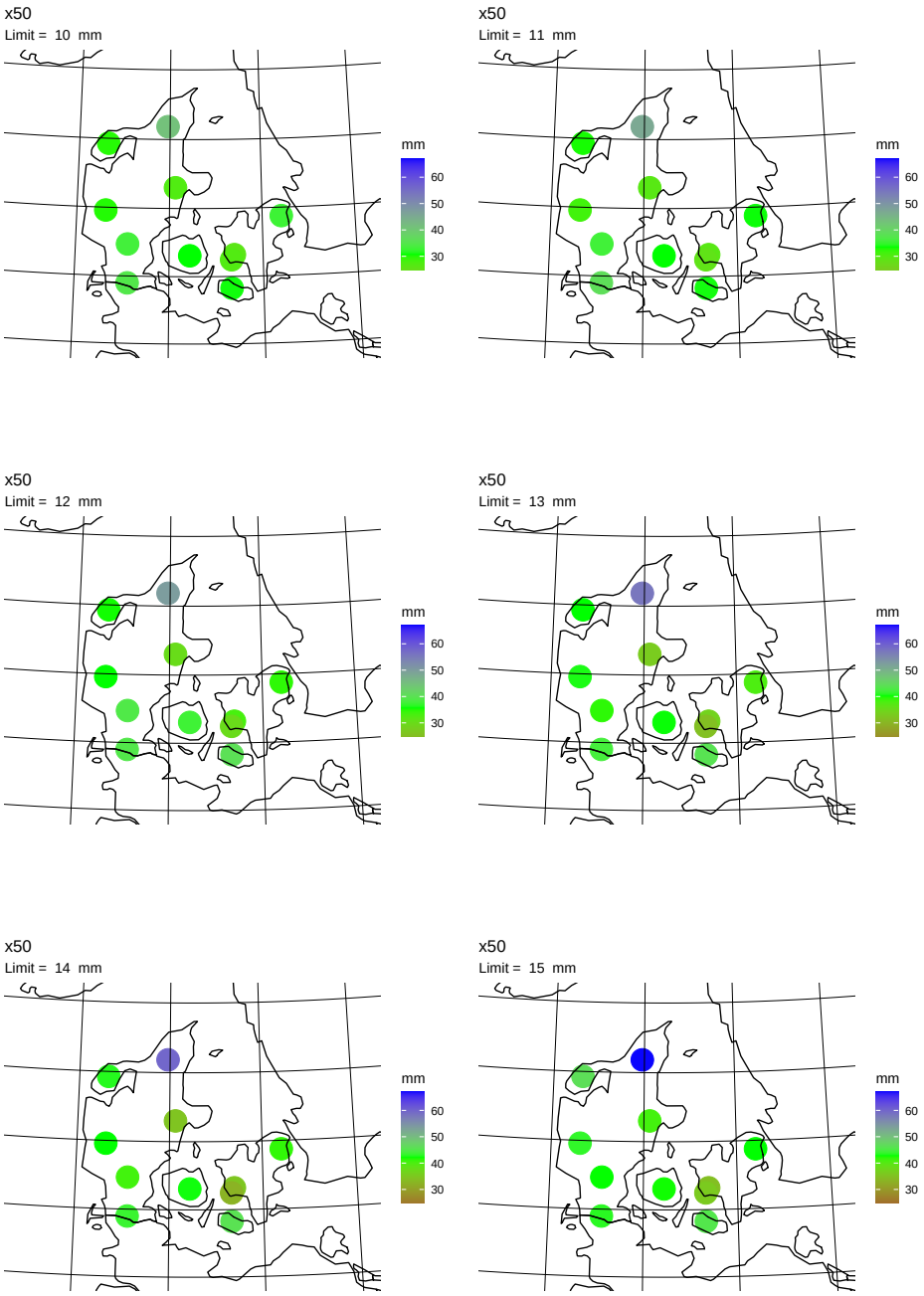

Figure 10: Geographic distribution of the $x_{1/2}$ parameter derived from the logit model parameters (see Section 5.2 for details), for 11 stations, for 6 different hourly maximum precipitation cutoffs.

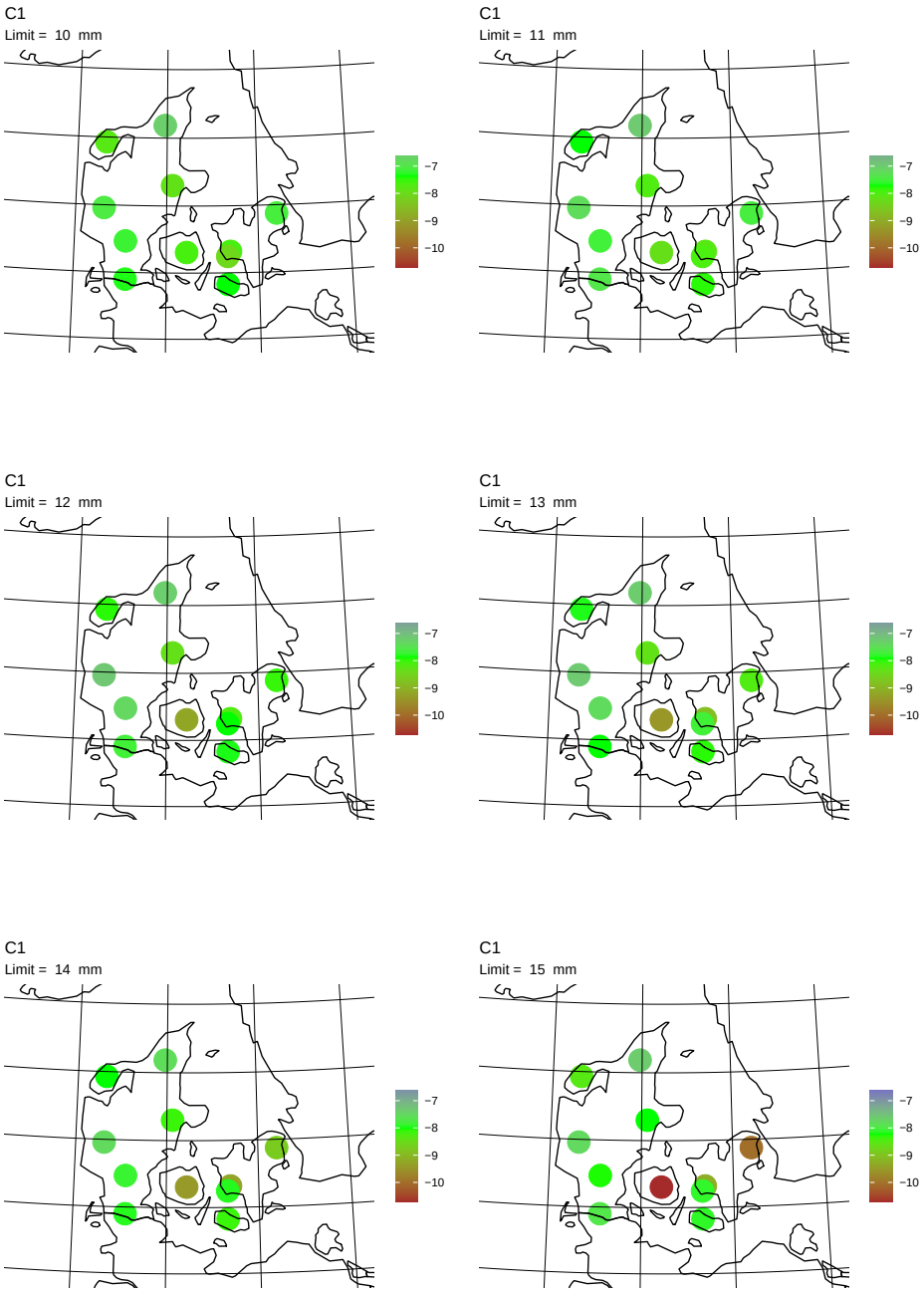

Figure 11: Geographic distribution of the Intercept logit model parameter for 11 stations, for 6 different hourly maximum precipitation cutoffs.

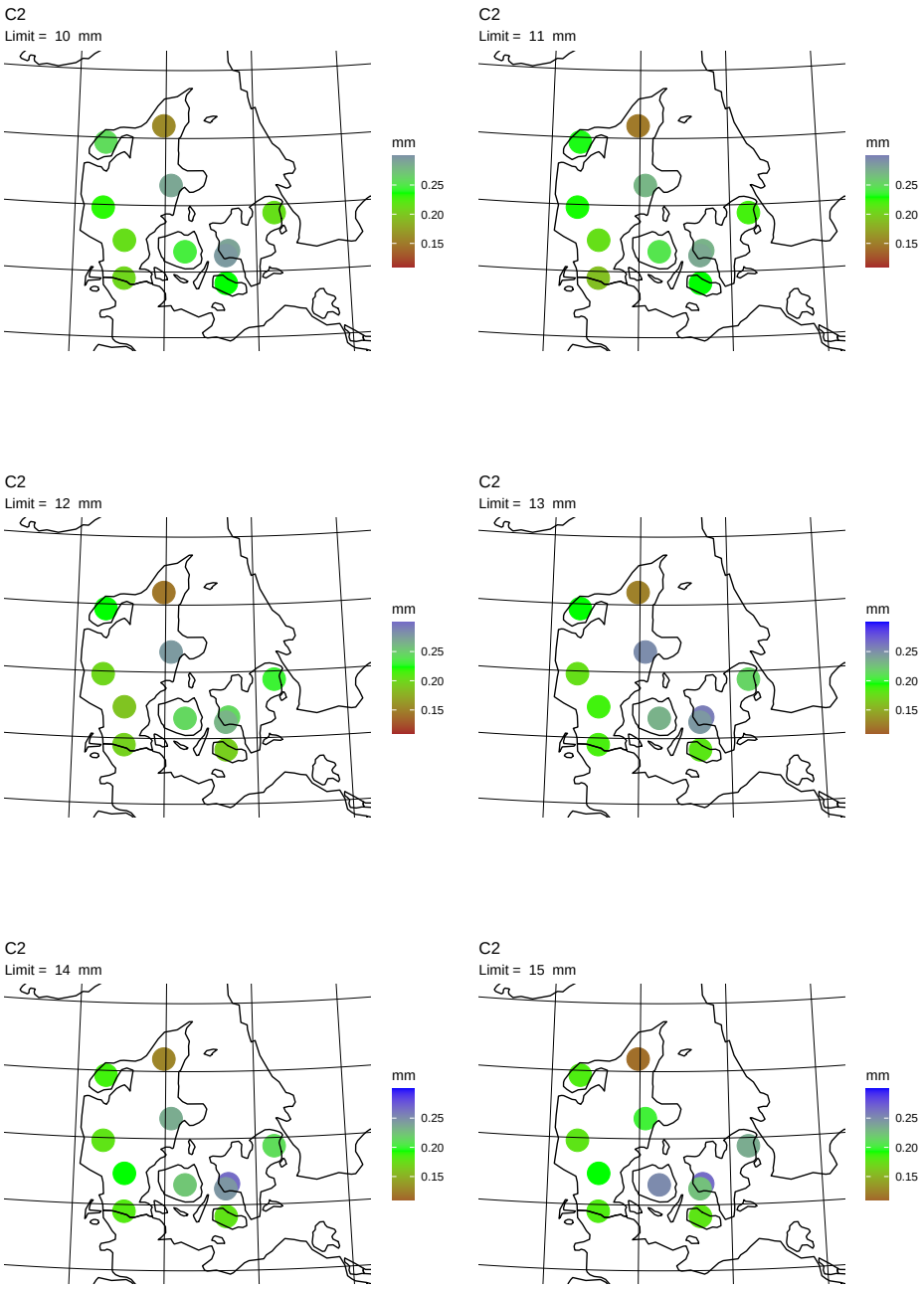

Figure 12: Geographic distribution of the Slope logit model parameter for 11 stations, for 6 different hourly maximum precipitation cutoffs.

select for analysis only days where the daily precipitation sum is at least 15 mm.

We perform the regressions between binary values ($Y_{1/0}$) of maximum hourly precipitation, and the daily precipitation sum as x, and T_{max} as z.

We get the coefficients listed in Table 4.

The coefficients are used to calculate $x_{1/2}$, which for the univariate regression is a single number, and for the multivariate depends on the T_{max} on the day. The multivariate model indicates that one degree increase in T_{max} lowers $x_{1/2}$ by 0.5 mm. This seems interesting from a physical point of view - on really hot days with cloudbursts, the daily sum of precipitation can be lower than on cooler days with cloudbursts.

We also inspect the sampling error in these results by resampling all data with replacement, and calculating the AIC (Akaike's Information Criterion, [1]), a well-known, mainstream method for comparing different models fitted to the same data[3].

In Figure 13 the bivariate model p-values are shown along with events with and without cloudbursts (using the 15 mm definition on max hourly precip, and only using data for which $DS > 15$ mm). There are quite a few 'false positives', i.e. days that the model designates as a cloudburst but which do not have hourly max precipitation above the DMI definition, at 15 mm (brown triangles). There are also 'true cloudbursts' below the p=0.5 black line. At a chosen high level of p these latter events would not be detected by this logistic model.

Table 4: Results of univariate and multivariate regression applied to binary values for cloudbursts (using a defining cutoff at 15 mm) and daily sum precipitation and daily T_{max} ($= z$) in degrees C (here taken from the series at Landbohøjskolen). AIC indicates that the multivariate model fit is better than the univariate one: the smaller AIC indicates the better model, and differences of about 3 in AIC indicate differences at the p=0.95 level. The large difference in AIC seen here clearly indicates that the multivariate model is much better than the univariate model.

Uni/Multi	β_0	β_1	β_2	$x_{1/2}$ [mm]	AIC
Uni	-8.4 ($\pm$0.2)	0.18 ($\pm$0.01)	-	47.5	505
Multi	-7.98	0.172	0.087	46 - 0.5$\cdot$ z	464

[3] *smaller* AIC values indicate the *better* model. Since, arguably, 'the same data' are not used for the univariate and the multi-variate model we need to consider appropriate method-selection procedures in future work.

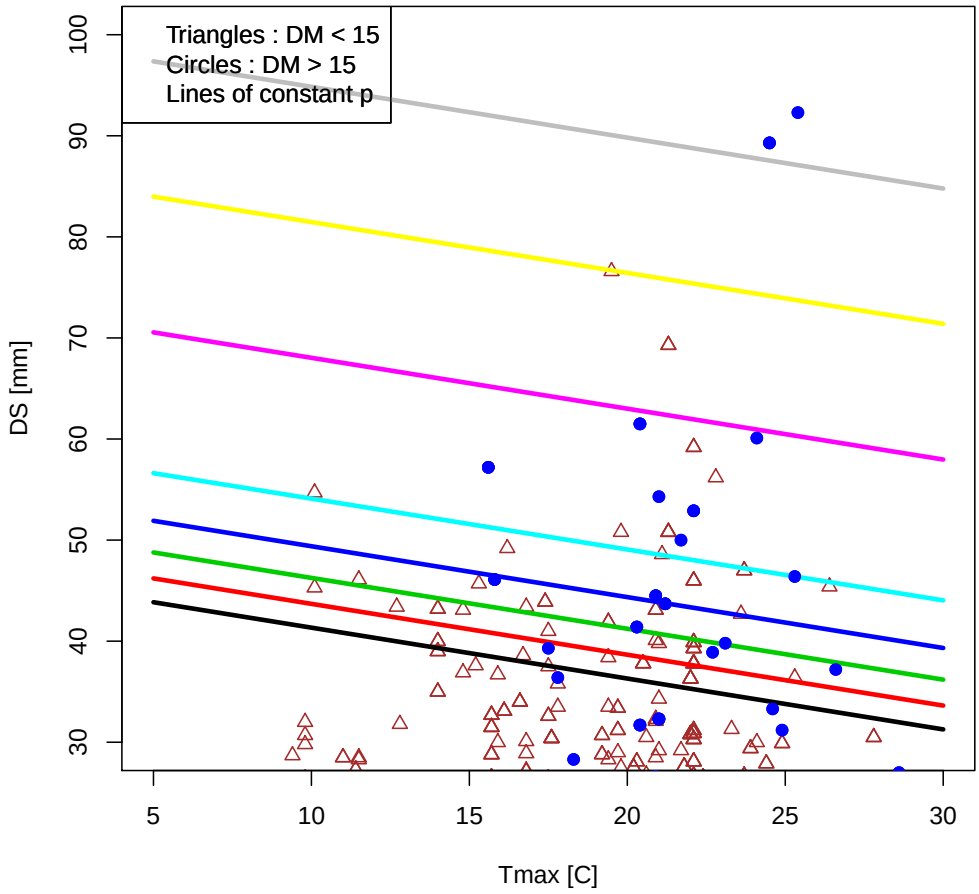

Figure 13: Values of observed daily sum of precipitation (DS) and observed daily maximum temperature (T_{max}) plotted for cloudbursts (blue filled circles) and non-cloudbursts (brown triangles). The cutoff for cloudbursts is $DM > 15$mm. The lines are for constant p-values, using Equation 3, starting at black with p=0.5 and then for p=0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 0.9999 at grey. Diagram based on data for which $DS > 15$mm, only.

7 Preliminary Summary and Discussion

The results from the first-look analysis above suggested that a cutoff in the daily sum of precipitation could be used as a probabilistic proxy for the occurrence of cloudbursts. The results of a deeper quantitative look at the combined station data suggested that the cutoff, or proxy, could be defined at 45 mm and did correspond, in the logistic sense, to a limit on the maximum hourly precipitation near 15 mm which would be close to the official definition of a cloudburst, although defined on half-hourly precipitation sums.

As we moved to analysis of individual stations we came up against low numbers of data to work with and saw (compare Figures 14 and 7) that the range of the $x_{1/2}$ parameter, which is our suggested proxy for a 50% chance of a cloudburst that day, was from 42 to 48 mm (mean 44.5 mm) in the case all station data were combined, but from 30 to 70 mm (mean 43 mm) if we looked at the combination of results for that parameter determined for individual stations. While the mean of these two intervals is near one another (44.5 and 43 mm) the distribution is much wider when individual station results are combined than if all data are combined and then analysed. This makes good sense in terms of what we would expect based on stochasticity and small-numbers effects.

Supporting this result is the observation that if the defining cutoff for a cloudburst is lowered from 15 mm to 12 mm an even narrower range from 35 to 37 mm between the 5% to the 95%-iles is found (when all station data is combined at the outset). For cutoffs in the full range from 10 to 15 mm we get the relationship shown in Figure 15. We see a rise in $x_{1/2}$ of 10 mm when the cutoff defining a cloudburst is raised by about 4 mm.

It is, of course, a bit arbitrary if we use 15 mm or another lower limit as the definition of a cloudburst since we still do not have 30-minute data available - any reasonable choice such as 12 or 15 mm will still lead to a useful proxy $x_{1/2}$ for use on daily-sum precipitation data – the results will still help us analyse 'extreme precipitation' events in lower-resolution data, and will enable studies of trends as well as the geographic patterns of this cloudburst proxy.

We saw that some stations 'stood out' when their logit model parameters were plotted on a map. With more data providing joint daily and hourly data we could investigate whether these individual oddities are part of a general pattern.

Based on the robustness of the results above we should try to recommend what definition of a cloudburst to use - the lower cutoff will lead to conclusions based on larger, more stable, amounts of data.

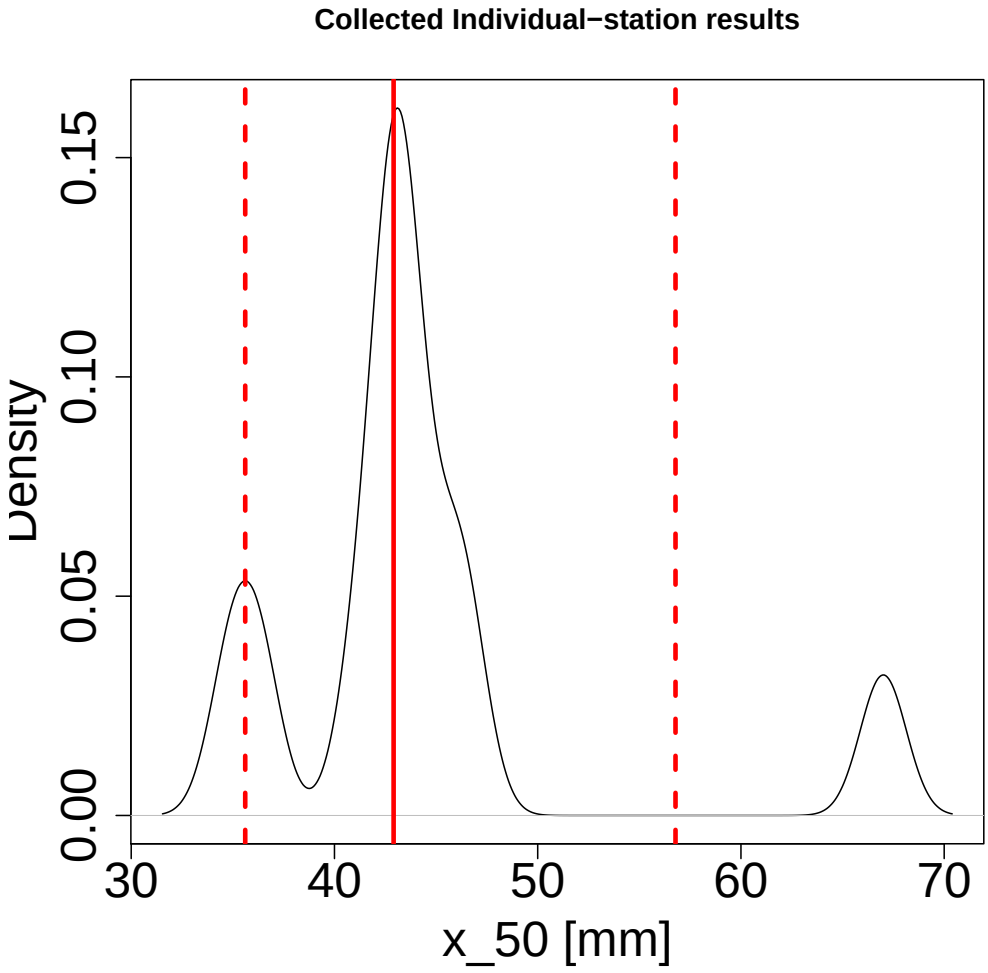

Figure 14: Density distribution (smoothed) of the 11 individual stations' $x_{1/2}$ parameter - that is, the cutoff in daily precipitation sum that corresponds to the 50% chance of a designated cloudburst.The vertical lines show the 5, 50 and 95%-iles. The spread due to station outliers is evident. Compare to determinations of $x_{1/2}$ using all data combined, in Figure 7.

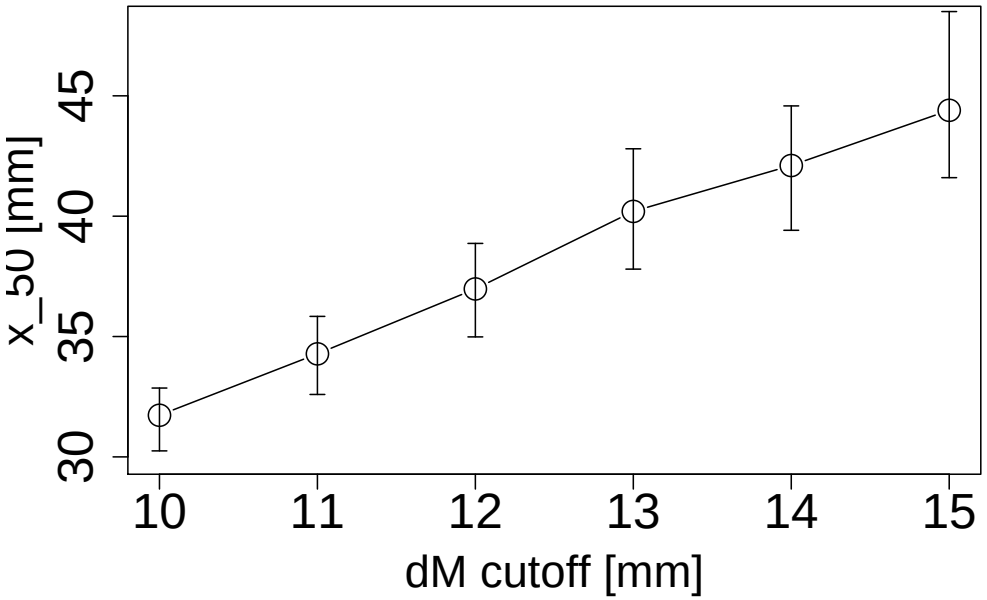

Figure 15: The relationship between the cutoff in hourly maximum precipitation used as definition for 'a cloudburst' and the corresponding, logit model predicted, cutoff in daily sum precipitation $x_{1/2}$.

If we can access data from more stations with both the maximum hourly and the daily-sum precipitation then we could investigate the possibility of regional patterns in e.g. the $x_{1/2}$ proxy, but we can also simply use one national average value (say, 45 mm) and then use *this* on all data and investigate patterns in *that* set of results.

This suggestion is largely based on the monotonicity we have shown.

Using T_{max}-values from stations nearer to where the precipitation was observed should improve the above multivariate experiment where data from one station only (Landbohøjskolen) was used. Using local observations would also enhance the geographic interpretability of the model fits. Use of even more relevant meteorological parameters should be explored.

8 Adding more data

While the above analysis was conducted, digitisation of the annual books continued and relevant data were drawn from the DMI database, enabling the creation of a dataset of daily-sum observations for 28 stations extending from 1917 to the present day (this can be further extended in the future, perhaps to more than 100 station series). The 28 stations in this report are listed in Table 2, and are shown as red dots on the map in Figure 16.

We now proceeded to investigate the geographical distribution of the rate of 'designated cloudbursts' for these stations by use of the 28 stations worth of data. In Figure 17 we show, as colour-coded dots, the rate of designated cloudbursts in the extended dataset of daily sum observations. The designation is based on identifying how many times a daily sum of precipitation exceeded 45mm divided by the number of years of data available in the series. Figure 18 shows the distribution of designated cloud burst rates in the data from the 28 station series. The median rate appears to be close to 0.15 events per year, or a return period of 7 years. This is comparable to the 5-year rain-flux curve in Figure 5 of [2].

The 28 stations do not seem to map Denmark densely enough to discern any robust patterns in where the high cloudburst-rate stations are. Most stations have cloudburst rates near the mean.

9 Project Summary and Discussion

In this two-step pilot project we have explored the relationship between cloud bursts as observed in hourly-sum precipitation series and the daily

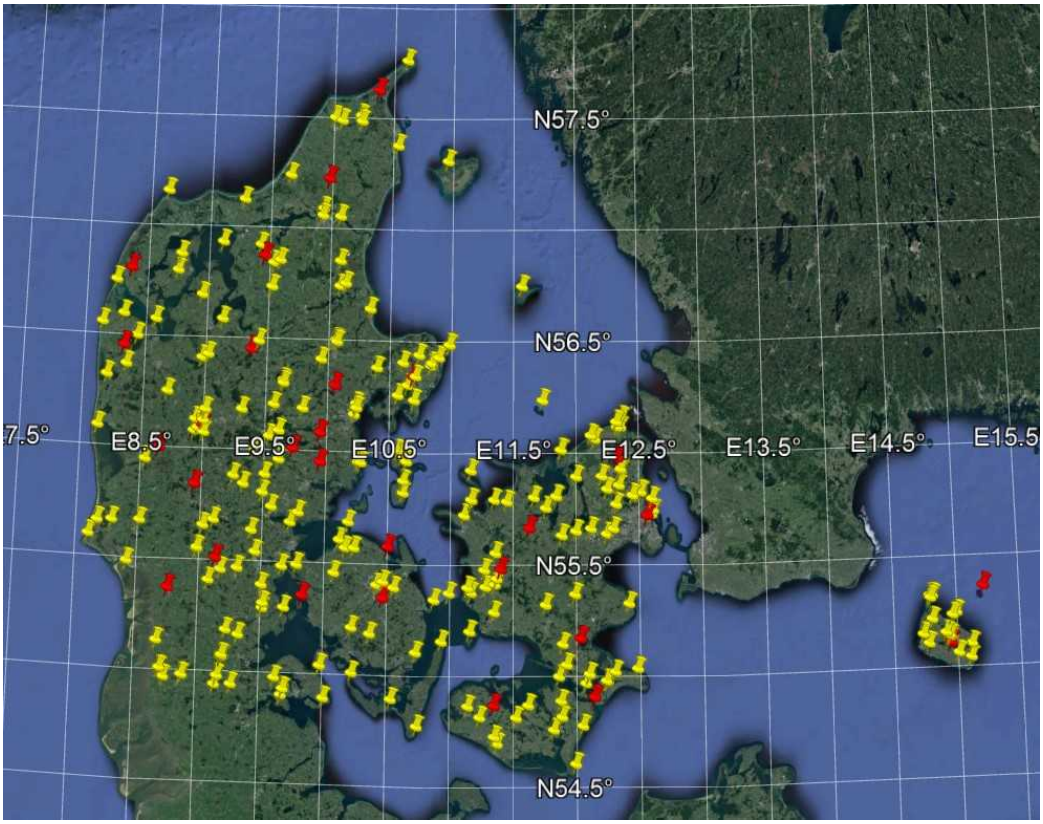

Figure 16: Map showing the location of 28 digitized stations with DS data covering 1914-1960s. Red markers indicate stations we look at in this report, while the yellow show other stations that have long records but so far have not been considered.

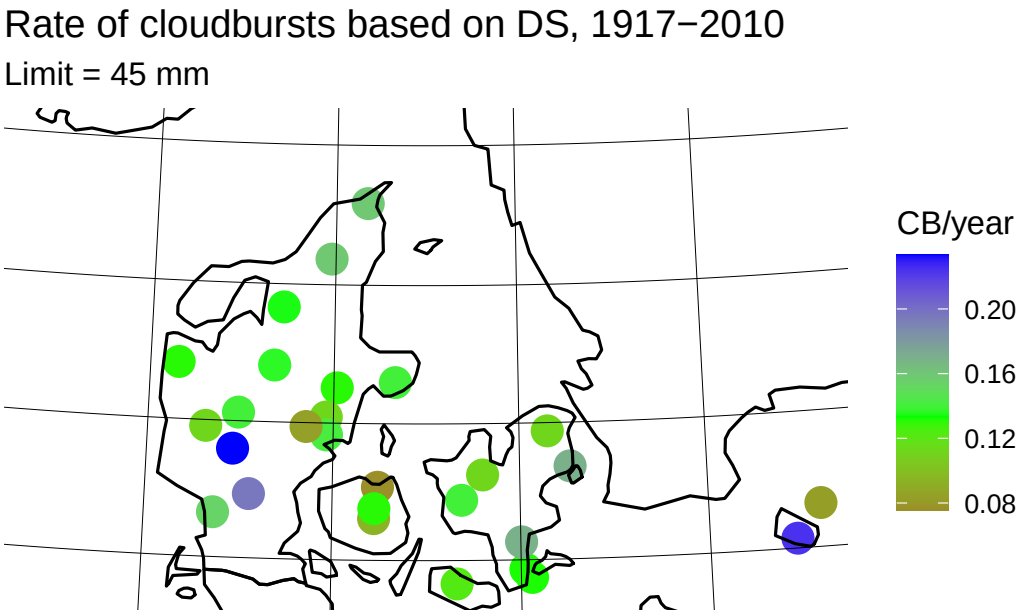

Figure 17: Rate of cloudbursts in events/year at 28 stations in Denmark, using data from 1917-2010. The proxy limit for a cloudburst used is 45 mm/day in the daily precipitation sum, as determined by logistic analysis in this work.

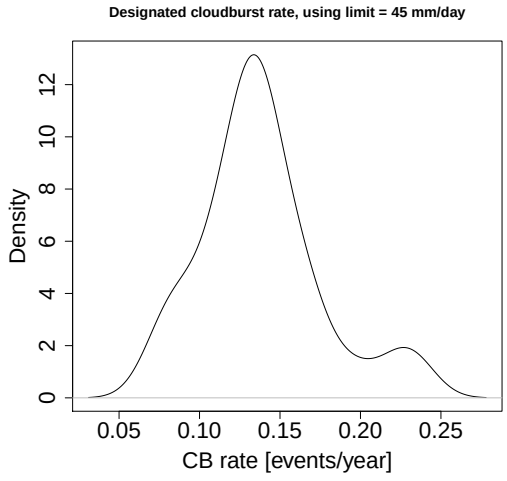

Figure 18: Histogram of the designated cloudburst rate using 45 mm/day in the daily sum as the designating limit.

sum in synop data, using logistic regression. We found that a limit near 45 mm/day in the daily sum is the best value if all data are taken together. This rule allows correct classification of 17% of the candidate cloudbursts as actual cloudbursts. We have explored whether the logistic regression can be improved by adding co-variates and found an indication that adding the daily maximum temperature can improve prediction of cloudbursts. Finally we approached our dataset of long digitized daily-sum-only data covering 100 years each so that we now have 28 century long time series of the daily precipitation sum in Denmark. Using the detected limit for the occurrence of a cloudburst we counted the rate of 'designated cloudbursts' across Denmark, and plotted the rate. We did not see any clear sign of geographical signatures in the positions of the stations with highest and lowest cloudburst rates, but suggest that adding more stations from the digitised set, and prolonging them to the present day, will yield results of interest for the question of whether cloudbursts occur more in some places than others, in Denmark. The question of the distribution of cloudbursts geographically over Denmark is important for efforts to first of all describe local hydrographic issues of interest for authorities tasked with water management in streams and on fields.

In addition to suggesting that the dataset used here is allowed to grow so that potentially 100 or 200 century-long series become available for analysis, we suggest that adding uncertainty estimates (e.g. based on resampling of the data) to the present analysis will help us understand how robust the results are and allow rigorous conclusions to be drawn.

We have shown that use of daily values for T_{max} significantly improves the skill of the regression model. We have so far used only one such series of T_{max} values – that of the Landbohøjskole in Frederiksberg. Because Denmark is a small country, a hot day in one place is therefore also likely to be a hot day in all of Denmark, this seems to work well, but we would like to suggest, in a possibly extension of this NCKF project, that T_{max} for regionally distributed locations are extracted from the database and used as shown.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, December 1974.

[2] Kim Sarup. DMI Report 20-03 Drift af Spildevandskomitéens Regnmålersystem Årsnotat 2019. Technical report, DMI, Copenhagen, 2020.

[3] M. L. Brandt. The North Atlantic Climatological Dataset (NACD). Instrumenter og rekonstruktioner. En illustreret gennemgang af arkivmateriale. DMI Technical Report 94-19. Technical report, DMI, Copenhagen, 1994.

10 Previous reports

Previous reports from the Danish Meteorological Institute can be found on: https://www.dmi.dk/publikationer/