



# Effects of sample size on estimation of rainfall extremes at high temperatures

Berry Boessenkool, Gerd Bürger, and Maik Heistermann

Institute for Earth and Environmental Sciences, University of Potsdam, Potsdam, Germany

Correspondence to: Berry Boessenkool (boessenk@uni-potsdam.de)

Received: 20 May 2016 – Discussion started: 27 June 2016

Revised: 7 August 2017 – Accepted: 22 August 2017 – Published: 25 September 2017

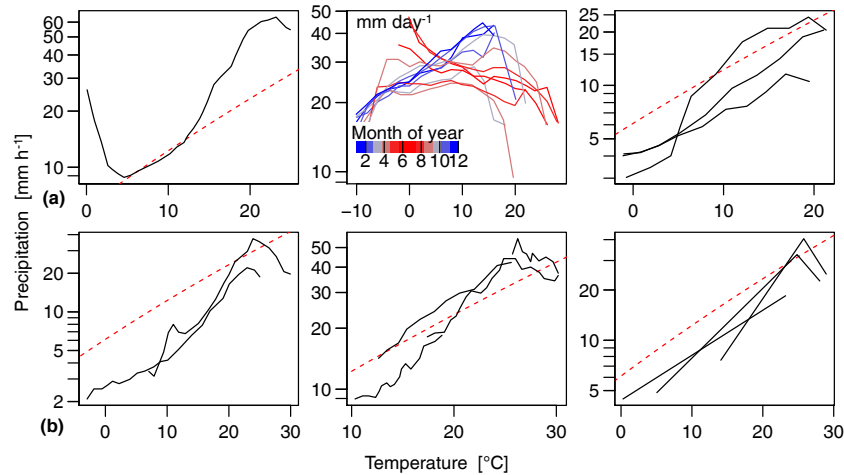
**Abstract.** High precipitation quantiles tend to rise with temperature, following the so-called Clausius–Clapeyron (CC) scaling. It is often reported that the CC-scaling relation breaks down and even reverts for very high temperatures. In our study, we investigate this reversal using observational climate data from 142 stations across Germany. One of the suggested meteorological explanations for the breakdown is limited moisture supply. Here we argue that, instead, it could simply originate from undersampling. As rainfall frequency generally decreases with higher temperatures, rainfall intensities as dictated by CC scaling are less likely to be recorded than for moderate temperatures. Empirical quantiles are conventionally estimated from order statistics via various forms of plotting position formulas. They have in common that their largest representable return period is given by the sample size. In small samples, high quantiles are underestimated accordingly. The small-sample effect is weaker, or disappears completely, when using parametric quantile estimates from a generalized Pareto distribution (GPD) fitted with  $L$  moments. For those, we obtain quantiles of rainfall intensities that continue to rise with temperature.

## 1 Introduction

The atmospheric water holding capacity and thus potential precipitation intensity depends exponentially on air temperature according to the Clausius–Clapeyron (CC) relationship. As empirically documented by several studies, high precipitation quantiles rise with temperature, increasingly so with shorter duration, such as hourly or shorter. This CC scaling describes a log-linear dependence of precipitation intensity on temperature ( $P$ – $T$  relationship) that roughly follows or

exceeds the CC rate of  $7\% \text{ K}^{-1}$  for water vapor. Similarly well documented is a breakdown or even reversal of that relation for temperatures beyond some thresholds, usually somewhere between 15 to 20 °C, as indicated in Fig. 1. This drop was also observed by Brandsma and Buishand (1997), Klein Tank and Koennen (1993), Panthou et al. (2014), and Westra et al. (2014). More details about the methods used in each referenced article can be found in Tables 1 and 2.

Several explanations for this phenomenon have been proposed, such as an increase in the proportion of rainfall stemming from convective events as opposed to large-scale stratiform precipitation (Haerter and Berg, 2009). Other explanations include a slower increase in moisture availability than in moisture storage capacity according to the CC relationship (Berg et al., 2009) or fully saturated conditions lasting less than event duration (Hardwick Jones et al., 2010). There may be several different mechanisms in process at different timescales and locations (Utsumi et al., 2011). The decrease in precipitation intensity at high temperatures coincides with a decrease in the number of observations. The aim of this study is to examine whether this drop could (partly) be a sample size artifact. For this purpose, we contrast two different approaches to estimate very high precipitation quantiles, namely empirical quantiles (which are based on plotting positions), and parametric quantiles (which are derived from fitting the generalized Pareto distribution (GPD) to the data). We compare both estimation methods with regard to their sample size dependency and their effect on the shape of  $P$ – $T$  relationships, using both observed hydrometeorological and synthetic data.



**Figure 1.**  $P$ – $T$  relationships (99 % quantile, hourly intensities) digitized from several figures in the literature on a logarithmic scale. Red dashed lines indicate CC scaling by the August–Roche–Magnus approximation (7 % at 0 °C, 6 % at 20 °C), see Panthou et al. (2014) and Hardwick Jones et al. (2010). Across regions and studies,  $P$  rises with  $T$  but then decreases. **(a)** Berg et al. (2013), Berg et al. (2009) ( $\text{mm day}^{-1}$ ), and Berg and Haerter (2013). **(b)** Lenderink et al. (2011), Hardwick Jones et al. (2010), and Utsumi et al. (2011) (converted from  $\text{mm day}^{-1}$ ). The last two articles use temperature bins of varying width with a semi-constant number of observations per bin. More details on study region and temperature variables can be found in Table 2.

**Table 1.**  $P$ – $T$  analysis methods used in the cited literature.

Article	Bin width	Min $n_{\text{bin}}$	Quantile + estimation method
Klein Tank and Koennen (1993)	2 °C	unknown	mean amount
Brandsma and Buishand (1997)	2 °C	unknown	mean amount
Lenderink and Meijgaard (2008)	2 °C	unknown	75 + 90 + 99 + 99.9 %, emp. + GPD top 5 and 10 %
Berg et al. (2009)	2 °C	300	99 %, GPD top 20 %; $\text{mm day}^{-1}$ (for each month)
Hardwick Jones et al. (2010)	variable	median 233	99 %, empirical order stats <sup>1</sup>
Lenderink et al. (2011)	2 °C (overlap: 1 °C steps)	200 <sup>1</sup>	90 + 99 + 99.9 %, empirical + GPD top 4 % <sup>2</sup>
Utsumi et al. (2011)	variable, avg. 2 °C	150	99 %, unknown, presumably empirical
Berg and Haerter (2013)	5 °C (overlap: 3 °C steps)	300	99 %, unknown, presumably empirical
Berg et al. (2013)	1 °C	unknown	95 + 99 %, unknown, presumably empirical
Panthou et al. (2014)	2 °C (overlap: 1 °C steps)	100	90 + 95 + 99 %, emp.: Cunnane unbiased estimator <sup>1</sup>
Westra et al. (2014)	–	–	as in Lenderink et al. (2011)

<sup>1</sup> Personal communication per email. <sup>2</sup> As in [climexp.knmi.nl](http://climexp.knmi.nl) (p).

## 2 Data and methods

### 2.1 Climate data

We analyzed publicly available time series of precipitation, temperature, and relative humidity from 142 stations across Germany from the German Weather Service (DWD, 2016). The stations are selected based on the length of available hourly time series. All selected datasets contain at least 15 years of observations, mostly 20 years. The R code for data selection, download, and analysis is available at <https://github.com/brry/pretemp>.

To analyze only the nonzero precipitation records that are actually of interest for this article, values below  $0.5 \text{ mm h}^{-1}$  are omitted. This cutoff is in line with the cited literature and

is suitable because measurements of very low rainfall intensities have a high relative uncertainty. The values are then logarithmized to enable a comparison of rates of precipitation change across temperatures. Because of the very skewed nature of rainfall values, this also allows for better distribution fits.

### 2.2 Temperature binning

Throughout this paper, event dew-point temperature is used as an integrated measure of air temperature and water vapor saturation (or moisture supply). It is defined as the average dew-point temperature of the 5 h preceding each rainfall hour, similar to the procedure by Lenderink et al. (2011). Dew-point temperature is calculated with the Magnus for-

**Table 2.** Regions and temperatures used in the literature cited in Fig 1.

Article	Region	Temperature variable
Berg et al. (2013)	SW Germany	Refers to Lenderink et al. (2011)
Berg et al. (2009)	Western Europe	Surface temperature
Berg and Haerter (2013)	Germany	Presumably air temperature
Lenderink et al. (2011)	Hongkong + NL	Dew-point temperature
Hardwick Jones et al. (2010)	Australia	Surface temperature
Utsumi et al. (2011)	Japan	Presumably air temperature

mula based on observed relative humidity and air temperature at 2 m height (see Buck, 1981).

Following the analysis method of Lenderink and Meijgaard (2008) and Berg and Haerter (2013), we partition the hourly precipitation depths according to the event dew-point temperature. We use moving temperature bins with a fixed width of 2 K. Bin midpoints increase in  $0.1^\circ$  steps.

### 2.3 Empirical quantiles

Empirical quantiles are estimated by a monotonic mapping of the ordered sample to sample-size-specific probabilities called plotting positions. This can be done in a variety of ways as reviewed by Hyndman and Fan (1996). Common to all is the fact that the portion to the right of the sample maximum is left unresolved (no extrapolations) and receives the same probability as the maximum. Quantiles representing return periods larger than the sample length are consequently mapped to that maximum. They are therefore underestimated – a fact apparently too trivial to have warranted any publication. The empirical quantiles used in this article are computed based on the  $\frac{k-1/3}{n+1/3}$  plotting positions ( $n$  = sample size,  $k = 1, \dots, n$ ; see Hyndman and Fan, 1996).

### 2.4 Parametric quantiles

The parametric quantile estimates are obtained in a peak-over-threshold approach, where the generalized Pareto distribution is fitted to the top 10 % of the sample. Quantiles are calculated from the fitted GPD.

We use the method of  $L$  moments to fit the GPD parameters. They are analogous to the conventional statistical moments (mean, variance, skewness, and kurtosis) but “robust [and] suitable for analysis of rare events of non-normal data.  $L$  moments are consistent and often have smaller sampling variances than maximum likelihood in small to moderate sample sizes.  $L$  moments are especially useful in the context of quantile functions” (Asquith, 2016, 2011; Hosking, 1990).

To obtain the quantile from the fitted distribution, the given probabilities must be scaled with the conditional probability of the truncation. For example, if the 99 % quantile (Q0.99) is to be computed from the top 10 % of the data, Q0.90 of the truncated sample must be used. We refer to Q0.99 as the

“censored 99 % quantile”. Because five values are required to obtain  $L$  moments, the minimum sample size at 90 % truncation is 50 (45 values are discarded).

Selecting a suitable fitting method is of great importance in the context of sample size bias. For example, unlike moment-based procedures, maximum likelihood estimation (MLE) can still show an underestimation bias at small sample sizes, as shown in the Supplement. This happens in small samples ( $n < 200$ ) for distributions with bounded parameters (and the optimum of the likelihood function lying on the boundary). We refer to the Supplement for a comparison of the different methods.

The GPD quantile computation formula used in the source code of Imomco is

$$x(F) = \begin{cases} \xi + \frac{\alpha}{\kappa} (1 - (1 - F)^\kappa) & \text{if } \kappa \neq 0 \\ \xi - \alpha \times \log(1 - F) & \text{if } \kappa = 0 \end{cases},$$

with  $\xi$  = location,  $\alpha$  = scale,  $\kappa$  = shape.

### 2.5 Sample size dependency

In Sect. 2.3, we pointed out that empirical methods inherently underestimate high quantiles in small samples. In order to quantify the potential effect in the context of  $P$ – $T$  relationships, we set up the following experiment: to investigate the dependency of both quantile estimation methods on sample size, we draw random samples from a defined population. This should optimally be a large set of values following a distribution observed in nature. We therefore use a pooled dataset with all the precipitation values observed at any of the 142 stations. From this population, we draw random samples of several sizes and compute empirical and parametric quantiles from each sample. For each sample size, this is done 1000 times, resulting in a corresponding quantile distribution depending on sample size.

### 2.6 Synthetic $P$ – $T$ relationship

We apply the results of the previous Sect. 2.5 – that is, the potential small-sample effects of empirical and parametric quantile estimates – to  $P$ – $T$  scaling relationships and analyze the drop at high temperatures. To study that effect, we designed an experiment with synthetic data. Here, precipitation values are generated in a way that exhibits a stable tem-

perature scaling over all temperature ranges. The CC-scaling rate is constant, and the increase in high rainfall quantiles per degree Kelvin remains the same over all temperatures. When sampling from such synthetic data, any drop in the  $P$ – $T$  relationship must be a statistical artifact. For this purpose, we define a “temperature-dependent GPD” with parameters that depend on temperature. To achieve a realistic temperature scaling, we base the parameters on the linear regression of the fitted parameters at several dew-point temperatures.

From that synthetic GPD, 1000 random samples are generated for each temperature bin. The sample size corresponds to the average number of precipitation observations at the climate stations in each bin. From these sets of random samples, the empirical and parametric 99.9 % quantiles are calculated.

### 3 Results

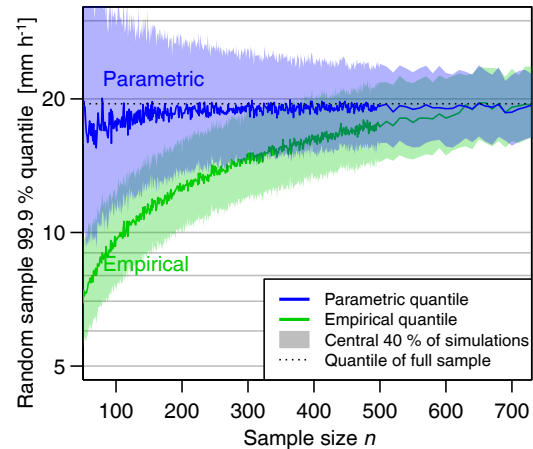
#### 3.1 Sample size dependency

The dependence on sample size, as revealed by 1000 random draws per sample size from the pooled precipitation data, is shown in Fig. 2. The 99.9 % quantile of this population ( $n = 1.16$  million) is  $19.5 \text{ mm h}^{-1}$ . It is strongly and consistently underestimated by the empirical estimator with shrinking sample size. For a sample size of 50, the median estimate is only  $7 \text{ mm h}^{-1}$ . Realistic estimates are obtained only for samples larger than about 700, around which the estimates converge to the (true) population value. The parametric estimators do not exhibit this bias – only their variance increases with smaller samples (the uncertainty range is wider). This is a typical example of the well-known bias–variance tradeoff in estimation theory.

#### 3.2 $P$ – $T$ relationship: empirical vs. parametric quantiles

The procedure of obtaining parametric (using the GPD) and empirical quantiles was applied per temperature bin to the datasets of each of the 142 stations. The empirical precipitation quantiles per bin are presented in the left panel of Fig. 3. The shape of the  $P$ – $T$  relationships is consistent with the behavior of  $P$ – $T$  relationships shown in Fig. 1 of the introductory section. The empirical quantile estimates start decreasing between 15 and  $20^\circ\text{C}$ . Some stations show the empirical quantile drop more distinctly than others. The figure also shows the average across stations, where the drop becomes particularly clear. Compared to the red line depicting the CC scaling of 7 to  $6\% \text{ K}^{-1}$ , the precipitation increase follows a super-CC scaling with a rise that is steeper than the CC rate. This is in accordance with previous findings, e.g., by Berg and Haerter (2013).

The parametric estimates are displayed in the right panel. At temperature ranges where empirical quantiles decrease, parametric quantiles keep increasing. This difference is less pronounced for smaller quantiles (see Supplement Sect. S4).



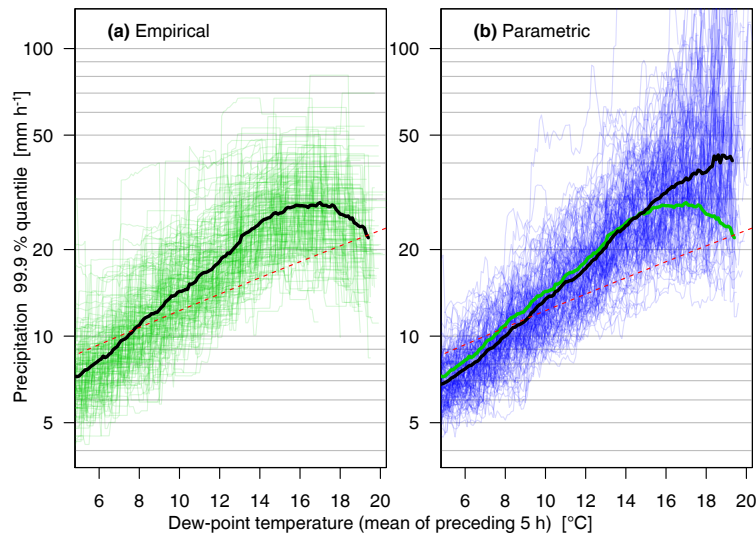
**Figure 2.** Median of the empirical and parametric 99.9 % quantile estimates depending on the size of samples drawn from all the precipitation intensity values along with their uncertainty bands. The horizontal dashed line marks the empirical quantile of the complete dataset ( $n = 1.16$  million). For  $n > 500$ , we used a step size of 10 (instead of 1) for the sample size, so the curve appears smoother there.

#### 3.3 Synthetic $P$ – $T$ relationship

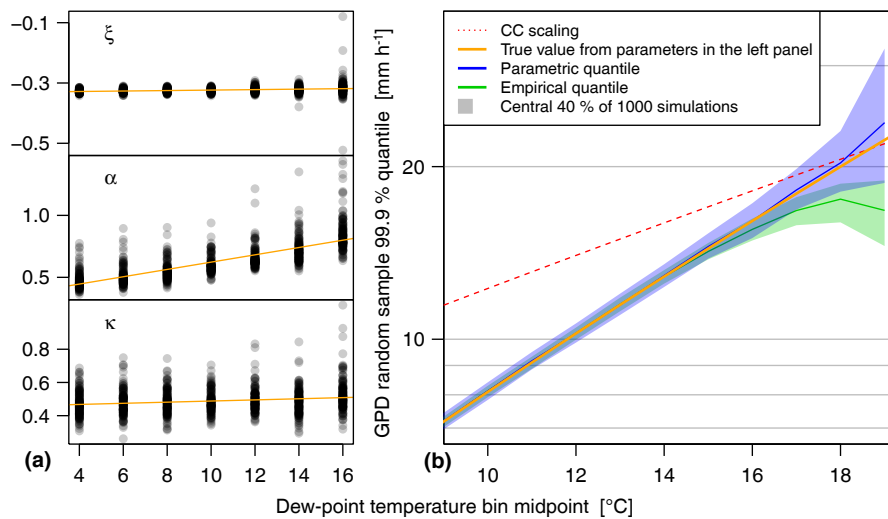
The synthetic  $P$ – $T$  relationship that continuously rises with temperature (see Sect. 2.6) is defined with the parameters shown in the left panels of Fig. 4, where each dot represents one of the stations. The right panel shows the median of the 99.9 % quantile estimates from random samples with the original sample sizes. Even though the distribution continues to increase with temperature, empirical quantiles from random samples stagnate or drop around  $18^\circ\text{C}$  where sample size decreases quickly. Parametric quantiles obtained by distribution fitting do not drop and follow the theoretical quantile from the distribution function.

### 4 Discussion and conclusions

Precipitation quantile estimates rise with temperature until they reach a turning point, beyond which they decrease. For this drop in the CC-scaling relation towards higher temperatures, a number of explanations have been suggested. In this study we offer the alternative view that the drop can be understood, at least in some cases, as a statistical artifact of small samples. At higher temperatures, fewer precipitation observations are available because (1) wet events are less frequent at high temperatures and (2) precipitation events at higher temperatures are generally convective in nature and very localized in space; they are thus often missed by the observing network, resulting in smaller sample sizes compared to large-scale precipitation at lower temperatures. A rather simple argument shows that empirical quantile estimators have an underestimation bias for return periods exceeding the sample



**Figure 3.** The 99.9 % precipitation intensity per temperature bin with empirical and parametric quantile estimate (a and b respectively). Each line represents one of the 142 stations, with the black line as the average across stations. The red line denotes CC scaling as in Fig. 1. The green line in (b) repeats the average from (a) for comparison.



**Figure 4.** (a) Parameters of a temperature-dependent GPD:  $\xi$  (location),  $\alpha$  (scale), and  $\kappa$  (shape). The orange lines show a linear regression as per Sect. 2.6. (b) Corresponding 99.9 % distribution quantile (orange) and median of the 99.9 % quantile estimates generated from samples in 1000 random draws along with their variance bands.

size, and we verified this behavior in a set of Monte Carlo experiments. It turned out that the underestimation of high quantiles, such as those relevant for the upper portion of the CC-scaling relationship, can be substantial. We have shown that when empirical estimators are appropriately replaced by parametric ones, the high-temperature drop in CC scaling disappears. The method of parametric estimation is crucial, nevertheless, as similar small-sample biases are known, e.g., from using MLE estimators (see above and more examples in the Supplement). The most robust estimates were obtained from moment-based methods. Past CC-scaling studies

that have relied on empirical or ML-based quantile estimators are likely affected by the small-sample artifacts for high temperatures that we have described here. For those, we find it necessary to revisit the corresponding estimation step using other, e.g., moment-based, procedures. This may be especially interesting for quantiles beyond the 99.9 % level.

To exclude potential physical effects related to precipitation as much as possible, we have repeated the analysis with synthetic data and obtained essentially the same results. Furthermore, we have used dew-point temperature instead of air temperature in order to rule out that the drop in the  $P$ – $T$  rela-

tionship is caused by a lack of moisture supply. It should be noted, though, that the use of dew-point temperatures only accounts for moisture that is already stored in the local atmosphere. It does not account for large-scale moisture convergence which becomes more important with longer precipitation duration intervals. This is evidence that the drop in empirical quantile estimates is precipitation independent; it is less a physical phenomenon but rather a statistical artifact caused by small samples, and it can largely be overcome by employing parametric estimators. Still, alternative physical explanations considering physical processes should not lightly be discarded. Some were summarized briefly in Sect. 1. It might also, for example, be hypothesized that near-surface temperature is not an adequate proxy for air temperature at the height where precipitation-forming patterns unfold on very warm days.

Parametric quantiles from fitted distributions provide a means to retrieve less biased estimates of extreme quantiles. The price to be paid is the larger uncertainty of those estimates. This should be quantified by confidence intervals or application to several datasets to avoid singular non-representative results. The parametric method requires significantly fewer data points in a sample than empirical quantiles need to converge to the actual (unknown) value. In the combination of small sample sizes and very high quantiles, the use of parametric quantiles is recommended.

*Code and data availability.* The datasets are freely available through the DWD Climate Data Center. The complete analysis code and more graphical results are available at <https://doi.org/10.5281/zenodo.892004>.

**The Supplement related to this article is available online at <https://doi.org/10.5194/nhess-17-1623-2017-supplement>.**

*Author contributions.* BB conducted the analysis and wrote the manuscript. GB and MH came up with the original idea and provided guidance and review.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* We wish to thank DWD for preparing and providing the datasets as well as William Asquith for reviewing our manuscript before submission. We are indebted to the reviewers for their many suggestions that led to the published version of this article.

Edited by: Uwe Ulbrich

Reviewed by: Reik Donner and two anonymous referees

## References

- Asquith, W. H.: Distributional analysis with L-moment statistics using the R environment for statistical computing, CreateSpace Independent Publishing Platform, <http://scholar.google.com/scholar?cluster=4144393830145643403&hl=en&oi=scholar> (last access: 15 September 2017), 2011.
- Asquith, W. H.: Imomco: L-moments, Censored L-moments, Trimmed L-moments, L-comoments, and Many Distributions, <https://cran.r-project.org/package=lmomco> (last access: 15 September 2017), 2016.
- Berg, P. and Haerter, J. O.: Unexpected increase in precipitation intensity with temperature. A result of mixing of precipitation types?, *Atmos. Res.*, 119, 56–61, <https://doi.org/10.1016/j.atmosres.2011.05.012>, 2013.
- Berg, P., Haerter, J. O., Thejll, P., Piani, C., Hagemann, S., and Christensen, J. H.: Seasonal characteristics of the relationship between daily precipitation intensity and surface temperature, *J. Geophys. Res.-Atmos.*, 114, D18102, <https://doi.org/10.1029/2009JD012008>, 2009.
- Berg, P., Moseley, C., and Haerter, J. O.: Strong increase in convective precipitation in response to higher temperatures, *Nat. Geosci.*, 6, 181–185, <https://doi.org/10.1038/ngeo1731>, 2013.
- Brandsma, T. and Buishand, T. A.: Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature, *J. Hydrol.*, 198, 98–123, [https://doi.org/10.1016/S0022-1694\(96\)03326-4](https://doi.org/10.1016/S0022-1694(96)03326-4), 1997.
- Buck, A. L.: New Equations for Computing Vapor Pressure and Enhancement Factor, *J. Appl. Meteorol.*, 20, 1527–1532, [https://doi.org/10.1175/1520-0450\(1981\)020<1527:NEFCVP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1981)020<1527:NEFCVP>2.0.CO;2), 1981.
- DWD: hourly precipitation records, [ftp://ftp-cdc.dwd.de/pub/CDC/observations\\_germany/climate/hourly/precipitation/historical/](ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/hourly/precipitation/historical/) (last access: 15 September 2017), 2016.
- Haerter, J. O. and Berg, P.: Unexpected rise in extreme precipitation caused by a shift in rain type?, *Nat. Geosci.*, 2, 372–373, <https://doi.org/10.1038/ngeo523>, 2009.
- Hardwick Jones, R., Westra, S., and Sharma, A.: Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity, *Geophys. Res. Lett.*, 37, L22805, <https://doi.org/10.1029/2010GL045081>, 2010.
- Hosking, J. R. M.: L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics, *J. Roy. Stat. B*, 52, 105–124, 1990.
- Hyndman, R. J. and Fan, Y.: Sample Quantiles in Statistical Packages, *The American Statistician*, 50, 361, <https://doi.org/10.2307/2684934>, 1996.
- Klein Tank, A. M. G. and Koennen, G. P.: The dependence of daily precipitation on temperature, Proceedings of the 18th annual climate diagnostics workshop, Boulder, Colorado, 207–211, 1993.
- Lenderink, G. and Meijgaard, E. v.: Increase in hourly precipitation extremes beyond expectations from temperature, *Nat. Geosci.*, 1, 511–514, <https://doi.org/10.1038/ngeo262>, 2008.
- Lenderink, G., Mok, H. Y., Lee, T. C., and van Oldenborgh, G. J.: Scaling and trends of hourly precipitation extremes in two different climate zones – Hong Kong and the Netherlands, *Hydrol. Earth Syst. Sci.*, 15, 3033–3041, <https://doi.org/10.5194/hess-15-3033-2011>, 2011.
- Panthou, G., Mailhot, A., Laurence, E., and Talbot, G.: Relationship between Surface Temperature and Extreme Rainfalls: A Multi-

- Time-Scale and Event-Based Analysis, *J. Hydrometeorol.*, 15, 1999–2011, <https://doi.org/10.1175/JHM-D-14-0020.1>, 2014.
- Utsumi, N., Seto, S., Kanae, S., Maeda, E. E., and Oki, T.: Does higher surface temperature intensify extreme precipitation?, *Geophys. Res. Lett.*, 38, L16708, <https://doi.org/10.1029/2011GL048426>, 2011.
- Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, *Rev. Geophys.*, 52, 2014RG000464, <https://doi.org/10.1002/2014RG000464>, 2014.