Introductory Univariate Extreme Value Analysis for Precipitation Data

Miles Moran

May 25, 2023

1 Context

Climate change literature has, for decades, detailed the accelerating increase of climate and weather extremes and their "widespread, pervasive impacts to ecosystems, people, settlements, and infrastructure" (IPCC 2022). As extreme precipitation events or "EPEs" (i.e. those at-or-above the 99th percentile compared to historical benchmarks) increase in both frequency and intensity, anticipating these events will be vital in creating resilient local communities. Of primary interest are (a) describing historical trends in EPEs (both spatially and temporally) and (b) making probabilistic forecasts.

Fortunately, there is a wealth of data suitable for these two purposes: the NOAA maintains the *Global Historical Climatology Network daily* (GHCNd), a database containing daily rain- and snow-fall totals from weather stations across the United States. This raw precipitation data is, however, difficult to work with for several reasons:

1. The data exhibit complex spatial and temporal dependency

2. Large periods of data are missing with minimal information about collection method, and the temporal span of the data is not consistent from station to station

3. The distribution of daily precipitation totals is both extremely zero-inflated and heavy-tailed; so, standard exponential families don't describe them well:

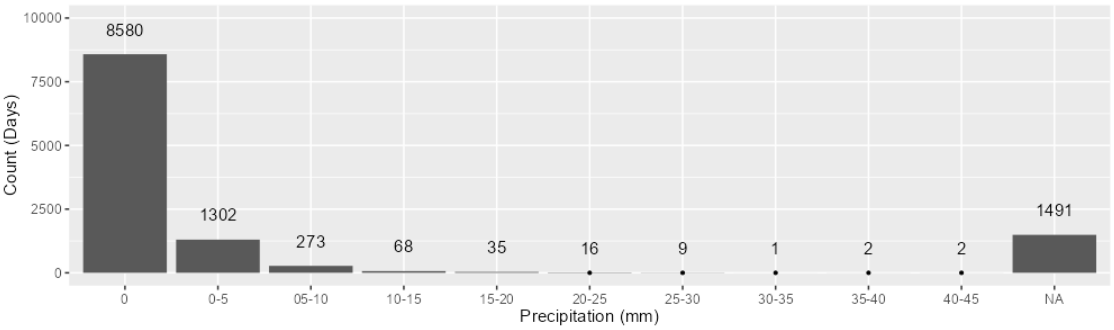

Figure 1: Density of Daily Precipitation Totals for Adel, Oregon (1979-2017)

4. The spatial dependence between observations exhibits behavior which makes spatial

> smoothing/interpolation problematic in that it "[diminishes] variability and extreme values" (Risser 2019).

So, when describing trends in *extreme* precipitation events, climatologists cannot rely on simple methods like linear regression or quantile regression, or even extensions of them like geographically-weighted regression. This niche problem motivates a new branch of statistical methods known as *extreme value analysis*, in which 'poorly-behaved' series like daily precipitation are summarized into 'well-behaved' series which capture information about low-probability events but are modeled much easier.

2 Extreme Value Analysis

2.1 Introduction

The simplest of these 'well-behaved' summaries is the *Block-Maxima Series* ("BMS"), i.e. the series containing the maximum one-day precipitation value observed within each pre-specified 'block' of time (e.g. monthly or yearly). If these blocks are large enough, then the maxima should be pairwise-independent and should *also* be robust to minor temporal incompleteness of the original data (partially addressing issues (1) and (2) above). What makes the block-maxima series shine, though, is its distributional properties.

Extreme Value Analysis 'works' because of asymptotic theorems about the tail-behavior of a sequence of random variables. The foremost of these theorems is the *Fisher–Tippett–Gnedenko Theorem* (colloquially, the "First Extreme Value Theorem" or "EVT"), which describes the limiting distribution of the sample maximum. If, for example, the sequence under investigation is used to form a block-maxima series, then the EVT allows us to make probabilistic statements about how *frequently* we expect a monthly- or yearly-maximum to exceed some threshold (or "return level").

2.2 Rigorous Treatment of the First EVT

Formally, the first EVT can be stated as follows:

Fisher–Tippett–Gnedenko Theorem ("First Extreme Value Theorem")

If $X_1, ..., X_n \overset{\text{iid}}{\sim} F(x)$ and there exists normalizing sequences $\{a_n\} \subset \mathbb{R}^+$ and $\{b_n\} \subset \mathbb{R}$ such that

$$\lim_{n\to\infty} \Pr\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right) = G(x)$$

then $G(x)$ must be the CDF of a Fréchet, Gumbel, or reflected-Weibull distribution – the three of which can be unified under the *Generalized Extreme Value* family of distributions ("GEV/GEVD").

The GEV family of distributions is characterized by location, scale, and shape parameters $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, and $\xi \in \mathbb{R}$. Generally, the behavior changes most notably with the choice of ξ (Figure 2); but, if a monthly- or yearly-maximum is modeled according to a GEVD, then researchers might be interested in knowing if *any* of these parameters vary according to space, time, or other covariate information.

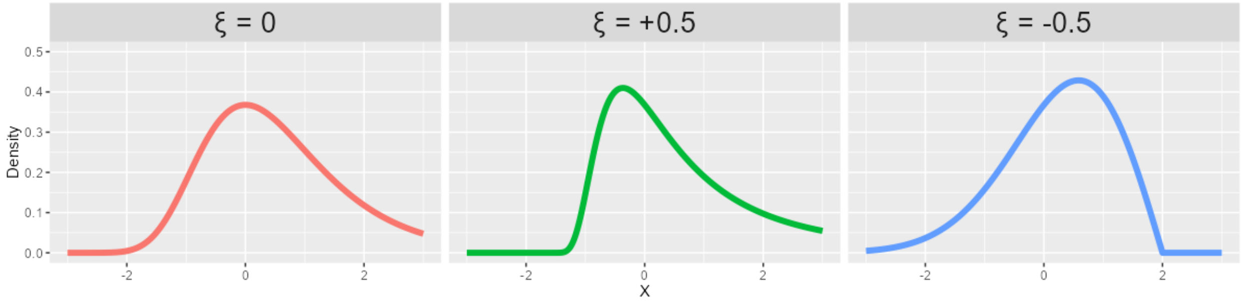

Figure 2: Three distributions, GEV($\mu = 0,\ \sigma = 1,\ \xi$), with varying shape parameter ξ.

If we can estimate these parameters for a particular block-maxima series, then, we completely define its distribution. Other quantities of interest – such as the mean or standard-deviation of the block-maxima – can be estimated and described too. Typically, the quantities most concerning to climatologists and hydrologists are *return periods* and *return levels*. A return level is the quantile that is expected to occur (or be exceeded) only once in a given time period. For instance, if a climatologist claims that 4 inches of rain in one day is a 1-in-100-year event for Portland, then 4 inches is the return level for the given return period of 100 years, and the return level can be estimated as the $q = \left(1 - \frac{1}{100}\right) = 0.99$ quantile of the GEVD fitted to the **annual** maxima series. Because the quantile function of the GEVD family has a closed form, estimating this return level z_q is straightforward:

$$\hat{z}_q = Q(q; \hat{\mu}, \hat{\sigma}, \hat{\xi}) = \begin{cases} \hat{\mu} - \hat{\sigma} \log\left(-\log\left(q\right)\right) & \text{for } \hat{\xi} = 0 \\ \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left(\left(-\log\left(q\right)\right)^{-\hat{\xi}} - 1\right) & \text{for } \hat{\xi} \neq 0 \end{cases}$$

2.3 Applying the First Extreme Value Theorem to EPE Data

With the asymptotic properties of the block-maxima series in mind, we can construct a general framework for modeling spatio-temporal extremes:

1. For each spatial location:
   (a) Summarize the daily precipitation series into the block-maxima series $Z_1, ..., Z_k$
      - If the block-size is large enough, then $Z_i \overset{\bullet}{\sim} \text{GEV}(\mu_i, \sigma_i, \xi_i)$ are independent
   (b) Estimate μ_i, σ_i, ξ_i according to some set of assumptions
      - e.g. if the original series $X_1, ..., X_n$ is assumed to be stationary, then $(\mu_i, \sigma_i, \xi_i) = (\mu, \sigma, \xi)$ for all i; so, only 3 parameters need to be estimated
2. Interpolate the GEVD parameters to a spatially-complete grid
3. For each grid cell, obtain quantities of interest (e.g. return level)

2.4 Challenges

The main complication with the framework described above is step 2. Spatial interpolation constitutes a whole field of study, and each method under that umbrella has to address the same set of questions:

- How are the GEVD parameter estimates spatially-related? Tobler's "First Law of Geography" states that "near things are more related than distant things;" but, how exactly does similarity decay with distance? How does directionality change that decay?

- How do we quantify the uncertainty in our interpolation results? Certain areas are bound to have a large margin-of-error on the interpolated estimate due to the sparsity of weather stations in those areas.

- Once an interpolation method *is* decided upon, what resolution should be used for rasterization? We want the resulting interpolation to be spatially-complete; but, that necessitates partitioning a continuous space into discrete grid-cells.

For extreme-value data, the main techniques for addressing these questions are Max Stable Processes (a sort of multivariate generalization of the GEVD), Bayesian or latent models (where the spatial structure is modeled indirectly by assuming a prior distribution for the GEV parameters) and Copula-based methods (where the true random field describing the spatial-structure of the GEV parameters is transformed into a Gaussian one) (Katz 2002).

3 Literature Review: Risser et. al. (2019)

An excellent demonstration of univariate extreme value analysis can be found in the 2019 manuscript "*A probabilistic gridded product for daily precipitation extremes over the United States* by Mark Risser and colleagues. Risser et. al. model the distribution of the annual-maxima for weather stations across the continental US using a variation of the Extreme-Value Theorem, then account for spatial dependence between parameter estimates using a hierarchical Gaussian-process model. Once this model is realized, the estimates (and corresponding return levels) are interpolated to create a new gridded data product.

Risser's method is used to model EPEs for one season at a time (which makes sense given most of the US experiences distinct wet/dry seasons). Without loss of generality, the following notation and methodology describes the process for modeling annual-maxima for the Winter season: let...

$$\mathcal{S} \equiv \text{the set of weather stations under study } (\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\},\ n = 5202)$$

$$m_t \equiv \text{block-size in year } t$$

$$k \equiv \text{a day in that block } (k \in \{1, ..., m_t\})$$

$$Z_{tk}(\mathbf{s}) \equiv \text{daily precipitation (in mm) on day } k \text{ of block-year } t \text{ at station } \mathbf{s}$$

With this notation, Risser's method is outlined below.

1. For each station $\mathbf{s} \in \mathcal{S}$:
    (a) For each year $t \in \{1950, ..., 2017\}$:
        - Calculate the block-maximum $Y_t(\mathbf{s}) \equiv \max_k\{Z_{tk}(\mathbf{s})\}$

    (b) With each $Y_t(\mathbf{s}) \overset{\bullet}{\sim} \mathrm{GEV}(\mu_t(\mathbf{s}), \sigma_t(\mathbf{s}), \xi_t(\mathbf{s}))$, assume that
        - location varies over time linearly: $\mu_t(\mathbf{s}) = \mu_0(\mathbf{s}) + \mu_1(\mathbf{s}) \cdot t$
        - scale and shape do not vary over time: $\sigma_t(\mathbf{s}) \equiv \sigma(\mathbf{s})$ and $\xi_t(\mathbf{s}) \equiv \xi(\mathbf{s})$

    (c) Assuming inter-year independence, obtain MLEs $\widehat{\mu}_0(\mathbf{s}), \widehat{\mu}_1(\mathbf{s}), \widehat{\sigma}(\mathbf{s}), \widehat{\xi}(\mathbf{s})$

2. Let $\boldsymbol{\mu}_0 = \big(\mu_0(\mathbf{s}_1), ..., \mu_0(\mathbf{s}_n)\big)$; likewise for $\boldsymbol{\mu}_1$ and $\boldsymbol{\xi}$; but, to model all parameters as real-valued, define $\log(\boldsymbol{\sigma}) = \big(\log(\sigma(\mathbf{s}_1)), ..., \log(\sigma(\mathbf{s}_n))\big)$.

3. For each $\boldsymbol{\theta} \in \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \log(\boldsymbol{\sigma}), \boldsymbol{\xi}\}$:
    (a) Assume $\boldsymbol{\theta}$ follows a nonstationary, anisotropic Gaussian process and the relationship between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ is hierarchical:

    $$\begin{aligned} \boldsymbol{\theta} &\sim \mathrm{MVN}\big(\boldsymbol{\delta}_\theta, \boldsymbol{\Sigma}_\theta\big), && \boldsymbol{\delta}_\theta \in \mathbb{R}^n,\ \boldsymbol{\Sigma}_\theta \in \mathbb{R}^{n\times n} \\ \big(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}\big) &\sim \mathrm{MVN}\big(\boldsymbol{\theta}, \mathbf{D}_\theta\big), && \mathbf{D}_\theta \in \mathbb{R}^{n\times n} \\ \implies \widehat{\boldsymbol{\theta}} &\sim \mathrm{MVN}\big(\boldsymbol{\delta}_\theta, \boldsymbol{\Sigma}_\theta + \mathbf{D}_\theta\big) \end{aligned}$$

    where $\boldsymbol{\delta}_\theta$ is assumed linear in an elevation-based covariate.

    (b) Using a variation on Empirical-Bayes Kriging (details omitted), estimate $\boldsymbol{\delta}_\theta$, $\boldsymbol{\Sigma}_\theta$, and $\mathbf{D}_\theta$ and interpolate $\widehat{\boldsymbol{\theta}}$ (along with 20-year return values) to a regular grid

4. Repeat steps (1)-(3) with block-boostrapped resamples of the annual maximum series for each station. Finally, use the bootstrap estimates to obtain standard errors.

Note how similar Risser's method is to the general framework described before. The novelty of this approach lies primarily in steps (2)-(4) as a way of addressing the aforementioned challenges with GEV parameter interpolation.

One benefit to modeling the GEVD parameters this way is that the resulting picture of the spatial dependency is both nuanced *and* interpretable. Although the computational cost is high, the assumption of anisotropy allows Risser to visualize the magnitude and directionality of the spatial dependence using Figure 3 below. From this map, we notice the extreme north-south dependence of precipitation extremes in the Pacific North-West. In contrast, we also notice the isotropy present in the Midwest and Great-Plains regions.

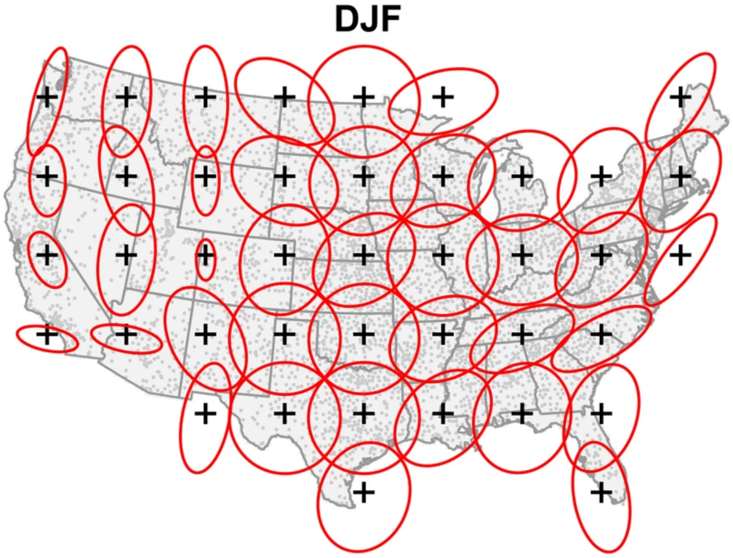

Figure 3: "Directional spatial length-scale for the location intercept $\mu_0(\mathbf{s})$ for DJF, estimated empirically from the data. The ellipses are a heuristic representation of the magnitude and direction of spatial dependence in $\mu_0(\mathbf{s})$"

Another benefit of this method is the ability account for covariate information (namely elevation) by tying it to the hyperparameters $\boldsymbol{\delta}_\theta$ and $\mathbf{D}_\theta$. For non-statistical reasons, this is information we *should* include somewhere in the model: high elevations are associated with more frequent rainfall events due to the lower temperatures.

4 Case Study: Oregon EPEs

Since Risser's strategy is so thoroughly documented, it is easy to implement on a smaller scale. After downsizing the dataset to include just weather stations in Oregon and just time points after 1-1-1979, I managed to reproduce the team's preliminary results with minimal computational issues[1].

Figure 4 depicts the maximum-likelihood estimates for the GEVD parameters at each weather station studied. Large-scale spatial patterns are similar to the results obtained by Risser et. al., such as the stark difference in location-intercept estimates (i.e. $\hat{\mu}_0(\mathbf{s})$) for locations East of the Cascades vs. West of the Cascades. Compare this to Risser's interpolated result in Figure 5. The spatial pattern for location-trend $\hat{\mu}_1(\mathbf{s})$ is noisier; but, it appears that locations West of the Cascades are generally experiencing an increasing linear trend in their (Winter) annual-maxima series, whereas locations East of the Cascades are noticing minimal change over time.

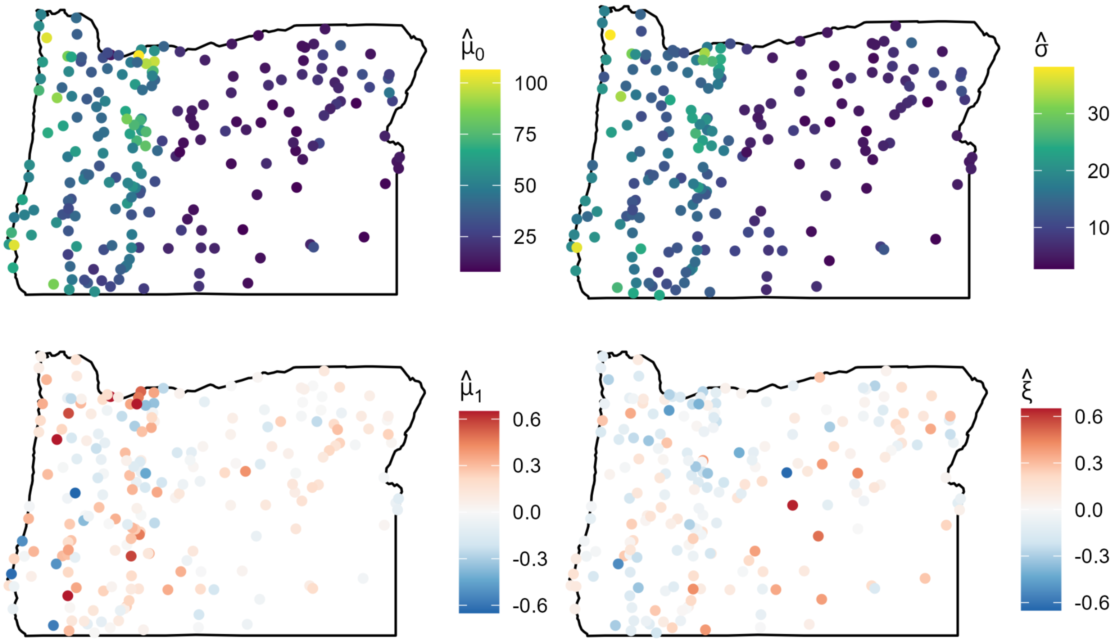

Figure 4: Estimated GEVD Parameters for the Winter Annual-Maxima Distribution Across 208 Weather Stations in Oregon, Obtained by Partially Re-creating Risser's Method

[1] As in Risser (2019), I leveraged the 'climextRemes' package for fitting the GEVD parameters

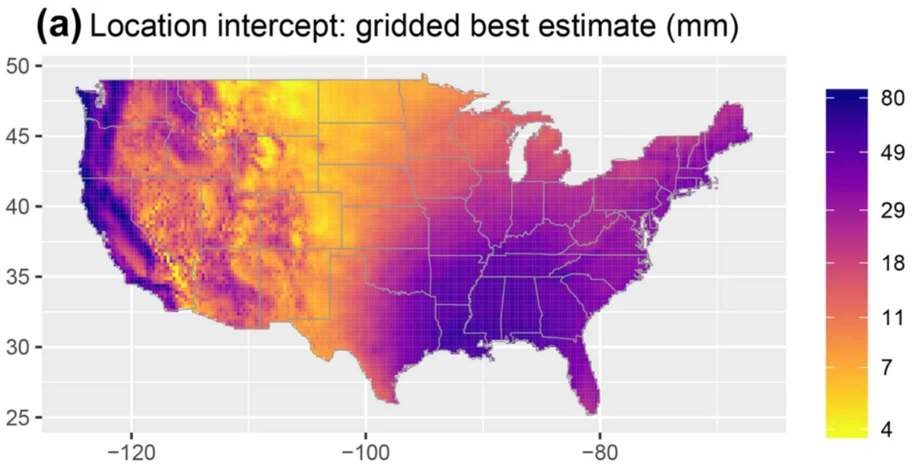

Figure 5: “Gridded best estimates of the location intercept $\mu_0(\mathbf{s})$ over [the continental USA] for [Winter]”

Another way of visualizing these results is to plot the fitted GEVD density functions for select locations of interest. Between 1980 and 2010, Figure 6 illustrates the noticeable increase in maxima for stations in Portland and Medford and the less-noticeable decrease in maxima for stations in Baker City and Redmond (both of which are East of the Cascades).

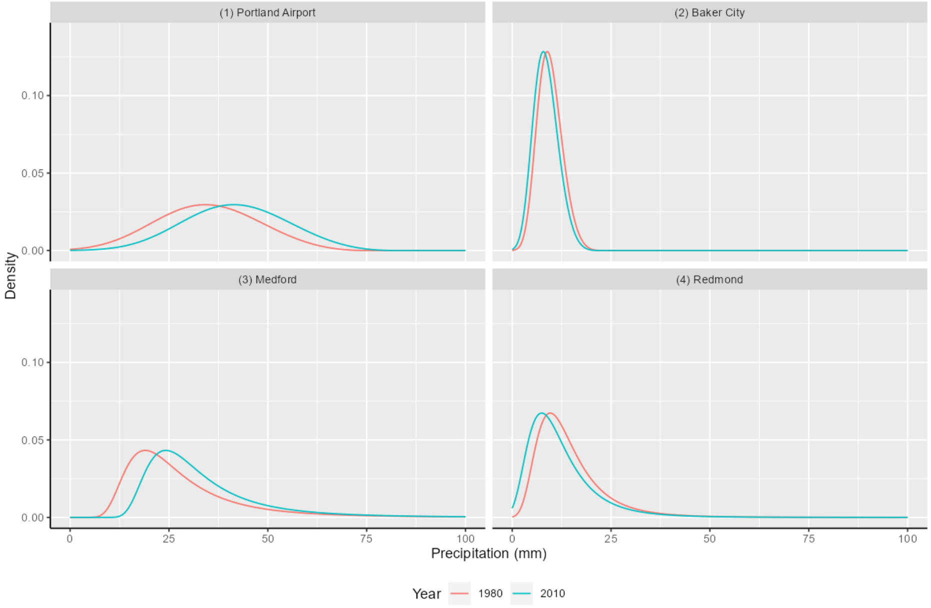

Figure 6: Estimated GEVD Densities for Select Locations, Estimated for $t \in \{1980, 2010\}$

5 Conclusion

The need to understand spatio-temporal patterns in extreme precipitation events presents researchers with a unique combination of problems; but, established methods in extreme-value analysis offer asymptotically-sound ways to address them. Univariate modeling of extreme-value time series is both easy to understand and computationally efficient; but, the ability to generalize a collection of univariate results to new locations is ultimately limited by the choice of spatial interpolation method. Risser et. al. (2019) demonstrate this procedure and propose a framework utilizing hierarchical Gaussian-processes for interpolation. This paper is an excellent starting point for understanding the univariate-and-combine paradigm; but, to demonstrate further, a basic data analysis leveraging similar method has been performed in R and documented in the Github repository below.

https://github.com/MilesMoran/MS-Project-Oregon-EPEs

6 References

Fagnant, C. Spatiotemporal Extreme Value Modeling with Environmental Applications. (2021) Diss., Rice University. https://hdl.handle.net/1911/111502.

IPCC, 2022: Summary for Policymakers [H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.)]. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 3-33, doi:10.1017/9781009325844.001.

Karlovits, G. "Extreme Value Theory." Statistical Methods in Hydrology, USACE Hydrologic Engineering Center, Davis, CA, 2023.

Katz W. R., Parlange, M. B., Naveau, P. Statistics of extremes in hydrology. Advances in Water Resources 25, 8-12, 1287-1304 (2002). https://doi.org/10.1016/S0309-1708(02)00056-8

Risser, M.D., Paciorek, C.J., Wehner, M.F. et al. A probabilistic gridded product for daily precipitation extremes over the United States. Clim Dyn 53, 2517–2538 (2019). https://doi.org/10.1007/s00382-019-04636-0