



# Bayesian hierarchical modelling of intensity-duration-frequency curves using a climate model large ensemble

Alexander Lee Rischmuller<sup>1</sup>, Benjamin Poschlod<sup>1</sup>, and Jana Sillmann<sup>1,2</sup>

<sup>1</sup>Research Unit Sustainability and Climate Risk, Center for Earth System Research and Sustainability (CEN),  
Universität Hamburg, 20144 Hamburg, Germany

<sup>2</sup>Center for International Climate Research (CICERO), 0349 Oslo, Norway

**Correspondence:** Alexander Lee Rischmuller (alexander.rischmueller@uni-hamburg.de)

Received: 13 May 2025 – Revised: 20 November 2025 – Accepted: 21 November 2025 – Published: 5 January 2026

**Abstract.** Accurate modelling of extreme precipitation is vital for predicting future risks and informing adaptation strategies. Here, we compare and evaluate six different extreme value statistical models for hourly to 48 h extreme precipitation in southern Germany, with a primary focus on duration-dependent Generalized Extreme Value (dGEV) distributions. To assess model performance, particularly in capturing tail behavior, we utilize the 50-member single model initial-condition large ensemble of the Canadian Regional Climate Model version 5 for the period 1980–2019. The large sample size of 2000 simulated years enables a robust sampling of extreme quantiles. Using a sub-sampling strategy with 30 to 100 years, we compare the efficacy of Bayesian methodology, in particular Bayesian hierarchical models, against frequentist models (L-moments and Maximum Likelihood Estimation – MLE) in representing the tail risk of 100-year return levels based on limited sample sizes. Hierarchical models allow us to give special emphasis on the dimensionality of the GEV shape parameter, a critical factor for tail behavior. Our findings reveal that a shape parameter varying over durations but fixed across space is beneficial for the prediction of the 100-year return level. The resulting Intensity-Duration-Frequency (IDF) curve shows the highest accuracy and smallest confidence intervals proving its robustness. Compared to the standard GEV estimated by L-moments, our proposed model can reduce the relative error of the 100-year return level from 18.1 % to 8.8 % based on a 30-year sample size. Furthermore, our analysis reveals fundamental limitations of the Anderson-Darling test for extreme value model selection, demonstrating its poor correlation with predictive skill for upper quantiles – a critical finding for climate risk applications.

## 1 Introduction

Extreme precipitation can severely impact various sectors such as infrastructure design, water resource management and agriculture (Yang et al., 2020; Mattingly et al., 2017; Schwarzk et al., 2015). Thus, different rainfall durations can trigger a variety of impacts. Short intense duration precipitation is commonly associated with convection and can lead to urban flooding, flash floods (often steep catchments), Hortonian overland flow, and riverine floods in small catchments (Haslinger et al., 2025; Fereshtehpour and Najafi, 2025; Ruiz-Villanueva et al., 2012; Fowler et al., 2021). Long persistent precipitation is mainly associated with synoptic-

scale meteorological systems, such as frontal or orographic systems, leading to riverine floods in larger catchments, often caused by saturation overland flow (Tarasova et al., 2019). This impact will increase in the future with increasing frequency and intensity of extreme rainfall events in a warmer climate (Fischer et al., 2021; Seneviratne et al., 2022; Poschlod and Ludwig, 2021). Computing the magnitude and probability of extreme events through statistical models, can help us assess risks of extremes from climate-related disasters (Ghil et al., 2011; Naveau et al., 2020).

Extreme rainfall is often modeled using the Extreme Value Theory (EVT), which focuses on the tails of distributions (Fisher and Tippett, 1928). Sampling maximum precipitation

values per temporal block, these maxima are assumed to follow the Generalized Extreme Value (GEV) distribution, a key distribution of the EVT, which is widely applied in climatology and hydrology (Hamdi et al., 2021). The GEV features the location, scale and shape parameters (Coles et al., 2001). For suitably chosen parameter sets, the occurrence probability of a given extreme precipitation intensity can be assessed, which is often expressed as return period of a corresponding return level. The flexibility of the GEV, whose shape parameter captures diverse tail behaviors, is accompanied by a high uncertainty for the shape parameter (Bücher et al., 2021), where small errors may distort return level estimates at high return periods (De Paola et al., 2018). This issue is further exacerbated by small sample sizes, which is typically the case for both sub-daily and daily rainfall durations (Lewis et al., 2019). However, accurate evaluation of the shape parameter is of high importance in extreme rainfall models (Papalexiou and Koutsoyiannis, 2013; Koutsoyiannis and Papalexiou, 2017). Hence, there is the need for methodologies providing a robust GEV parameter estimation.

A significant practical application of extreme value theory (EVT), and a crucial tool for engineering and flood risk management, is the intensity-duration-frequency (IDF) curve (Cannon and Innocenti, 2019). IDF curves graphically illustrate the frequency of rainfall intensities (return levels) at different durations. This information is very important in infrastructure design, as both short intense or long persistent types of extreme rainfall require different types of engineering solutions (Martel et al., 2021). For example, stormwater retention basins manage short intense rainfall (Pumo et al., 2023). On the other hand, drainage systems, permeable pavements, and retention ponds address long persistent rainfall (Vijayaraghavan et al., 2021). The main purpose of IDF curves is for design of dual drainage systems: the minor system (e.g. storm sewers) typically of a 10-year return period, and the major system manifesting return levels up to 100 years e.g. for roadways or detention ponds. In contrast, green infrastructure, such as permeable pavements or green roof is designed for more frequent events, such as a 2 year return period, which are often not based on IDF curves. In order for IDF curves to be consistent over durations the physical constraints of the maxima between different durations cannot be violated (Nadarajah et al., 1998). In a traditional setup, modeling GEV separately across durations may violate this constraint. To meet this logical consistency requirement, one could use the duration-dependent Generalized Extreme Value (dGEV) distribution—a modified GEV featuring duration-dependent parameters (Koutsoyiannis et al., 1998). These duration-dependent parameters incorporate curve-fitting parameters needed to fit IDF curves. The dGEV is fit with extreme data, typically block maxima, at different discrete durations.

In our study, we aim to develop a new methodology to robustly estimate occurrence probabilities for extreme events consistently across different rainfall durations and to prove

its benefits over existing state-of-the-art methodologies. In this regard, we compare four distinct parameter estimation methods: the Maximum Likelihood Estimation (MLE), L-moments, Bayesian, (including hierarchical) inference. The MLE is a widely used method in a range of different statistical models (including EVT), which determines the parameter vector that maximizes the likelihood function of observed data (Prescott and Walden, 1980). L-moments, a robust alternative to traditional moments, are less sensitive to outliers and particularly useful for heavy-tailed distributions (Hosking et al., 1985). Bayesian inference, integrates prior knowledge and uncertainty into model estimation, which has the potential for more informed predictions (van de Schoot et al., 2021). With Bayesian hierarchical models (BHM) we are able to model dependencies among different dimensions, such as space or durations (Veenman et al., 2024). Cooley et al. (2007) modeled extreme precipitation using a BHM with a generalized Pareto distribution (GPD), characterized by geographical and climatological covariates within a spatial Gaussian process. In addition, Cooley and Sain (2010), apply a BHM on regional climate model simulations. The BHM sensibly pools information from neighboring grid cells, which enabled Cooley and Sain (2010) to implement a spatially varying shape parameter.

Jalbert et al. (2022) established Bayesian hierarchical modeling of the dGEV for spatial interpolation of precipitation extremes in Canada using station data and a regional climate model driven by reanalysis. In our study, we put the focus on the comparison of parameter estimation methods and the different BHM setups addressing the dimensionality, duration-dependent variability, and spatial variability of the shape parameter ( $\xi$ ). BHM setups exist separately over durations (Räty et al., 2022), whereas our emphasis is on the dGEV and IDF curves.

A major disadvantage of many studies is the limitation due to small sample sizes (Marra et al., 2018), especially for observations, but also in many high-resolution climate model simulations, which typically provide only decadal to 30-year simulations (Ban et al., 2020; Lucas-Picher et al., 2021; Poschlod and Daloz, 2024). This makes it difficult to reliably estimate changes in rare events like climate extremes. Long model runs or large ensembles are a way to overcome this limitation because they provide a “ground truth” for assessing the appropriate size of smaller samples while also offering extensive, gap-free time series (Stein, 2020). Using millennial-scale global climate model simulations, Huang et al. (2016) investigate changes of temperature extremes between pre-industrial and future climates in the contiguous United States, applying the GEV distribution. They show that using 20- and 50-year subsets of the full 1000-year simulation for GEV fitting can lead to poor estimates of rare return levels that might strongly distort climate change signals. Large ensembles of regional climate models can be employed to robustly assess extreme precipitation at finer spatial scales (Aalbers et al., 2018; Brönnimann et al., 2018;

Poschlod et al., 2021) or precipitation extremes occurring jointly with extreme wind speed (Huang et al., 2021), storm surge (van den Hurk et al., 2015) or snow melt (Poschlod et al., 2020). In this study, we employ the Canadian Regional Climate Model Large Ensemble (CRCM5-LE; (Leduc et al., 2019)) for the period 1980 to 2019, offering 50 realizations of the climate and therefore providing a homogeneous sample of 2000 years. These 50 realizations of the large ensemble differ only due to minor perturbations in the initial conditions – the subsequent variability can be interpreted as a model representation of internal climate variability (Deser et al., 2020). Large ensembles enable robust inferences about rare events and allow for accurate estimation of extremes from the full model output (Stein, 2020). Here, we mimic the typical situation of data availability of station observations and artificially select individual grid cells as localities with sub-sample sizes ranging from 30 to 100 years. With 2000 years of data available, we can empirically obtain “effective return levels” as our ground truth, assessing the ability of various extreme value models to reproduce the *true* return levels from the sub-samples. Martel et al. (2020) used three large ensembles, including the CRCM5-LE, to analyze extreme precipitation events. They also utilize bootstrapping to estimate projected confidence intervals. Kharin et al. (2007) uses multimodel ensembles to account for natural variability. Similar to our approach, a GEV is utilized in 20-year effective return levels are also used for validation. In a further paper, Kharin et al. (2013) also used large ensembles and GEVs in analyzing extremes. Similar to our work, model performances were also evaluated using large ensembles.

Another approach for data creation is a stochastic weather generator (WGs), which is a synthetic approach (Wilks and Wilby, 1999). These models use statistical distributions trained on observational or climate model data to create synthetic precipitation time series (Semenov, 2008). This statistical approach differs from a physically based model, such as our CRCM5-LE, as it does not simulate the underlying physical processes of the climate system. Moreover, the CRCM5-LE does not assume any underlying statistical distribution. A key trade-off with WGs is that their reliability is highly dependent on the length and quality of the precipitation data used for calibration, which can often be scarce (Beneyto et al., 2020). Furthermore, models like the autoregressive WG can overestimate the spatial correlation of extreme events, leading to unrealistic areal precipitation values (Ullrich et al., 2021).

The primary aim of this study is to assess how a range of different statistical models and estimation methods compare in their ability to robustly and accurately reproduce effective return levels from the CRCM5-LE across various durations and localities. We investigate how the performance of these models varies with sample sizes ranging from 30 to 100 years, mirroring the typical data limitations that observational datasets and climate simulations have. Given the challenges of estimating tail behavior with limited data, we ex-

plore strategies to optimally model the spatial and temporal dimensions of the dGEV shape parameter under data-scarce conditions.

## 2 Data & Study Area

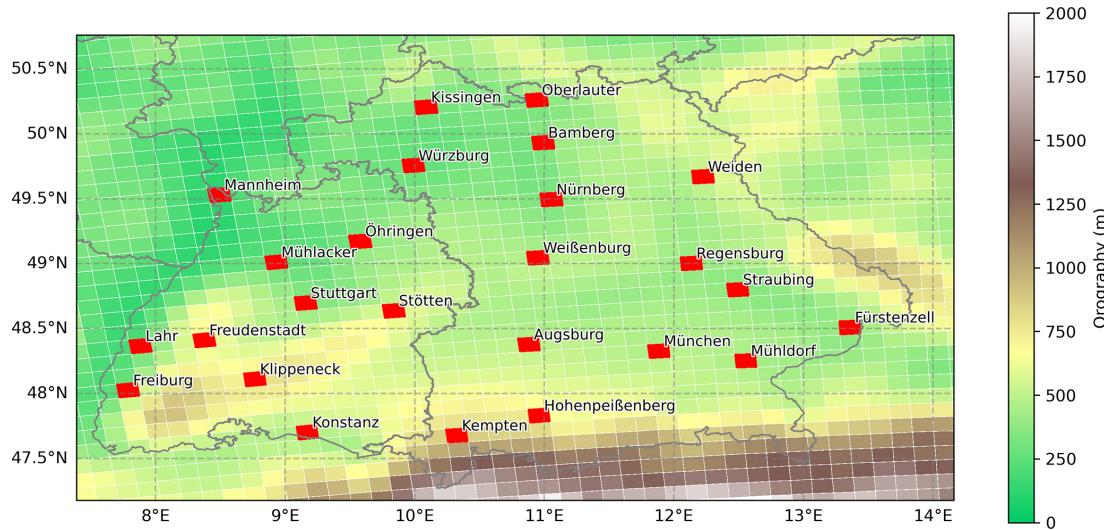
### 2.1 The study area in southern Germany

We focus on the region of southern Germany, which is characterized by heterogeneous elevation, making it an ideal testing ground for examining the robustness of statistical models (see Fig. 1). Especially during summer, southern Germany receives more heavy precipitation than other parts of Germany, due to its elevation which promotes orographic lifting and enhances convective processes (Jung and Schindler, 2019). We select 25 grid cells within our regional domain representing the typical density of observational stations providing sub-daily precipitation measurements (see Fig. 1). The localities cover an elevation range between 98 to 977 m highlighting the complex topography of the area. We extract a range of durations covering 1, 3, 6, 12, 24, and 48 h.

### 2.2 The high-resolution large ensemble CRCM5-LE

To generate a single-model initial condition large ensemble, one single global climate model is run several times driven by a particular radiative forcing scenario. Small perturbations are applied to the initial conditions of each model run leading to different weather and climate realizations. The resulting ensemble spread can be interpreted as a model representation of internal climate variability (Deser et al., 2020). Large ensembles typically feature global climate models with spatial resolutions of several hundred kilometers, which is why they cannot represent extreme precipitation on sub-daily time scales. In our study, we employ the CRCM5-LE, a 50-member large ensemble dynamically downscaled over a European domain at a spatial resolution of  $0.11^\circ$  (approximately 12.5 km) from the global Canadian Earth System Model version 2 (CanESM2; Leduc et al., 2019). The dynamical downscaling is carried out over a European domain at a spatial resolution of  $0.11^\circ$  equaling 12.5 km.

The simulations of the CRCM5-LE extend from 1950 to 2099 and are driven by the high-emission scenario RCP8.5. We extract a 40-year time period from the recent past in 1980 to 2019, where Schwalm et al. (2020) report that total cumulative CO<sub>2</sub> emissions from 2005 to 2020 are in close agreement (within 1 %) with the RCP8.5 scenario. We simplifyingly assume stationarity for this 40-year period, which is in accordance with most operational extreme value models (Van de Vyver, 2012). We employ a Mann-Kendall test (Mann, 1945; Kendall, 1938) to evaluate the stationarity assumption in the 50 members of the CRCM5-LE. On a significance level of 95 %, only 6 % of the 50 CRCM5-LE members and all six durations show a significant trend, where 5 % show a positive trend. Hence, we follow that the assumption



**Figure 1.** The study area of southern Germany with the topography of the CRCM5-LE grid, which has a resolution of  $0.11^\circ$ . The red gridcells indicate the selected grid cells of the climate simulations, which are extracted as the localities for our experiment.

of stationarity is reasonable. Despite not explicitly simulating convection, the CRCM5-LE effectively captures the intensity of sub-daily rainfall extremes over Europe (Poschlod et al., 2021), showing good agreement with observations in the study area. In a brief evaluation, we test the CRCM5-LE data against observations from weather stations (Deutscher Wetterdienst, 2022) located within the periphery of CRCM5-LE grid points. The annual maxima of the CRCM5-LE data are extracted and the Generalized Extreme Value (GEV) distribution is fitted using L-moments, with 500 bootstrap samples of 28 years (matching the modal sample size of the observational data). We apply the Anderson-Darling test at  $\alpha = 0.05$  on the GEV fits of the bootstrap parameters and the observations. For more on the Anderson-Darling test see Sect. 3.4.2 and the GEV see Sect. 3.1. Only 6 % of the fits are rejected indicating that the CRCM5-LE captures the characteristics of extreme precipitation across the durations well. Figure 2 displays the return levels of CRCM5-LE alongside observations from weather stations in Freudenstadt for all durations. The comparison demonstrates the good agreement between the two datasets (see Figs. S1–S25 in the Supplement for the other localities). This demonstrates that the model successfully captures the characteristics of extreme precipitation across various durations as well as its spatial variability between grid cells and observational stations.

It is very important to note, that we do not aim to extensively evaluate the performance of the CRCM5, for which there is ample literature (Poschlod et al., 2021; Poschlod, 2021). Instead, we use the CRCM5 as a “perfect model experiment” (Lenderink et al., 2023), assuming it as a perfect representation of the real climate system, where we know the “true” return level of the large ensemble. In this type of experiment, a climate model’s large ensemble simulations,

which sample a wide range of internal climate variability, are used as a proxy for a complete observational record. Thus, this experiment allows us to compare the performance of the different extreme value statistical models under idealized conditions. Hence, this experiment is referred to as “perfect model experiment” (Bevacqua et al., 2023).

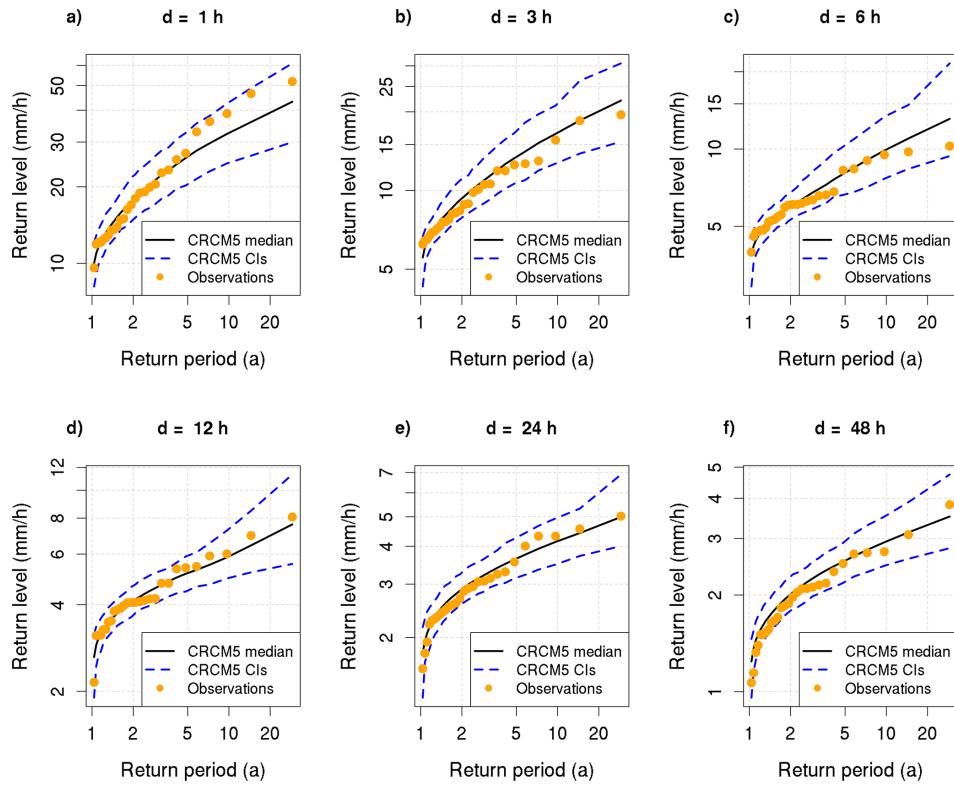
### 3 Methods

#### 3.1 The Generalized Extreme Value distribution and its duration-dependent version

According to the Fisher-Tippett-Gnedenko theorem (Fisher and Tippett, 1928; Gnedenko, 1943), for sufficiently large blocks, the distribution of block maxima asymptotically follows the Generalized Extreme Value (GEV) distribution. The cumulative distribution function (CDF) of the GEV distribution is:

$$F(x; \mu_{j,d}, \sigma_{j,d}, \xi_{j,d}) = \exp \left\{ - \left[ 1 + \xi_{j,d} \left( \frac{x - \mu_{j,d}}{\sigma_{j,d}} \right) \right]_{+}^{-1/\xi_{j,d}} \right\} \quad (1)$$

where  $\mu_{j,d}$  is location parameter,  $\sigma_{j,d}$  is the strictly positive ( $\sigma_{j,d} > 0$ ) scale parameter and  $\xi_{j,d}$  is shape parameter for durations  $d$  and localities  $j$ . The shape parameter  $\xi$  governs the tail behavior and assigns the distribution to the reversed Weibull ( $\xi_{j,d} < 0$ ), Gumbel ( $\xi_{j,d} = 0$ ), or Fréchet ( $\xi_{j,d} > 0$ ) type. The duration-dependent GEV (dGEV) combines data from all durations into one model as it assumes that  $\mu_{j,d}$  and  $\sigma_{j,d}$  follow a certain duration dependence (Koutsoyannis et al., 1998; Fauer et al., 2021):



**Figure 2.** Return-level curves for the locality of Freudenstadt for durations (a) 1 h, (b) 3 h, (c) 6 h, (d) 12 h, (e) 24 h, (f) 48 h. Return level curves of CRCM5-LE are marked in the black line and 95 % confidence intervals in blue dashed line. The confidence intervals are obtained via bootstrap resampling. The annual maxima of the observational data (German weather service – Climate Portal) are plotted via Weibull plotting positions as the orange circles.

$$\sigma_j(d) = \sigma_{0,j}(d + \theta_j)^{-\eta_j} \quad (2)$$

$$\mu_j(d) = \tilde{\mu}_j \sigma_{j,d} \quad (3)$$

with  $\sigma_j(d) > 0$ ,  $\sigma_{j,0} > 0$ ,  $\theta_j > 0$ ,  $0 > \eta_j > 0$ . Instead of modeling the durations separately, the duration dependency reduces to 4 parameters.

There are other versions of the dGEV, such as models that feature “multiscaling” (Gupta and Waymire, 1990). There,  $\eta$  is decomposed into two components  $\eta_1 + \eta_2$  rendering to  $\sigma(d) = \sigma_0 \cdot (d + \theta)^{-(\eta_1 + \eta_2)}$ . Fauer et al. (2021) found that the “multiscaling” dGEV is beneficial to cover a wide range of durations, i.e. 1 min to 5 d. However, in our study, we cover a range from 1 to 48 h, where the simple scaling approach is sufficient (see Fig. S51 in the Supplement).

### 3.2 Estimation methods

To cover the state-of-the-art procedures, we utilize four different estimation methods for fitting the statistical models. One of the estimation methods is L-moments estimation method, which are expectations of linear combinations of order statistics. They are analogous to conventional moments

but can be estimated by linear combinations of order statistics, i.e. by L-statistics (Hosking, 1990). L-moments, being less sensitive to outliers, provide more robust and reliable estimates of the distribution’s parameters, especially in small samples (Stedinger, 1993).

The Maximum Likelihood Estimation (MLE) method finds the values of the model parameters that maximize the likelihood function, which is a measure of how well the model explains the observed data. For a more comprehensive insight into the MLE see Coles et al. (2001), Martins and Stedinger (2000).

The Bayesian estimation method is a statistical method that uses prior knowledge with current data, which then updates in a posterior distribution of parameters (Zyphur and Oswald, 2015). From the posterior distribution we can obtain a measure of uncertainty. In Bayes’ theorem, prior distribution is combined with the likelihood to produce the posterior distribution (see Eq. 4).

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \quad (4)$$

where  $P(\theta|X)$  is the posterior distribution,  $P(X|\theta)$  is the likelihood function,  $P(\theta)$  is the distribution of the parameters called the “prior” and  $P(X)$  is the marginal likelihood,

which normalizes the posterior distribution. We start with a prior distribution that reflects our initial knowledge about the parameter. Prior knowledge and observed data are synthesized when the prior is updated with the data through the likelihood function resulting in the desired posterior distribution. For a more in-depth analysis of Bayes' theorem we refer to Sivia and Skilling (2006) and Gelman et al. (2013). Choosing a prior is normally subjective and can be difficult without sufficient prior knowledge about the parameters (Berger, 1985). For estimation of the posterior parameters we use the No-U-turn sampler (NUTS), NUTS is an algorithmic enhancement of the Hamiltonian Monte Carlo (HMC) which uses the gradient information to sample from the target probability distribution (Hoffman and Gelman, 2014). In contrast, the commonly used the Metropolis-Hastings algorithm utilizes gradient-free random walk which is effectively "blind" to the distribution's geometry (Chib and Greenberg, 1995). For simplicity, we will refer to the samples generated by NUTS as "MCMC" samples.

When prior information is insufficient, Bayesian hierarchical models (BHM) may be an alternative, as they can share a common hyperprior across a dimension (or many dimensions). In this way information is borrowed and learned from the data across the sites. The shared hyperprior centers a group of parameters, such as the means of normal distributions, around a common value (Gelman et al., 2013; Congdon, 2019). The central tendency of the hyperprior allows the model to shrink estimates toward a shared mean, therefore improving accuracy and robustness, particularly when data is sparse or noisy (Congdon, 2019). During inference, as data integrates into the model, individual group means tend to cluster around an empirically identified central value due to the hierarchical structure's information sharing across groups (Gelman, 2007; McElreath, 2015). Even when the hyperprior is uninformative or improper, shrinkage and balance toward a common mean can occur. In this process the BHM balances information from individual groups with the overall population. This mechanism enables the formulation of informed prior estimates for each site, using population-level data (Gelman et al., 2013; Congdon, 2019; Gelman, 2007). In the following example the Bayesian hierarchical model is set in three different tiers, which is typical for Bayesian hierarchical models (Gelman et al., 2013). A simple depiction of the BHM is as follows:

Data-Level:  $y_j | \theta_j, \phi \sim P(y_j | \theta_j, \phi)$

Parameter Level:  $\theta_j | \phi \sim P(\theta_j | \phi)$

Hyperparameter Level:  $\phi \sim P(\phi)$

where  $P(y_j | \theta_j, \phi)$  is the likelihood,  $P(\theta_j | \phi)$  is the prior distribution and  $P(\phi)$  is the hyperprior distribution. See Congdon (2019) for a more comprehensive overview of Bayesian hierarchical models.

### 3.3 The applied EVT models

We apply six different models in order to evaluate their performance. Two frequentist and four Bayesian models make up our repertoire. As a summary, we provide an overview of the six different EVT models in Table 1.

#### 3.3.1 The frequentist models

In our first frequentist model we use a standard GEV, where the equation is expressed in Eq. (1). As this model is not duration-dependent, we model each duration separately. Additionally, since this model is not hierarchical, we also model the locations separately. This results in six GEV models for each duration and for 25 separate locations, totaling 150 GEV models. We use the R package `extRemes` (Gilleland and Katz, 2016) to apply the L-moments estimation method for this as described in Sect. 3.2. As a second frequentist model we apply the dGEV, which incorporates duration dependence as outlined in Eqs. (2) and (3). This combines all the durations into one dGEV model for each locality equaling 25 dGEV models in total. As implemented in the R package `IDF` (Fauer et al., 2022), we use the MLE to estimate this model and refer to it as dGEV-MLE. We generate confidence intervals for both frequentist models using a nonparametric bootstrap with a sample size of 500 (Efron and Tibshirani, 1994), resampling the annual maxima with replacement.

#### 3.3.2 The duration-dependent Bayesian GEV

In the Bayesian duration-dependent generalized extreme value distribution (dGEV-Bayesian) model, we employ distinct priors for the parameters of each location. These priors do not share a hyperprior, thus the model is not hierarchical. Each parameter has a prior normal distribution, with the mean of these priors based on estimates for the corresponding dGEV-MLE parameters.

$$\tilde{\mu} \sim N(\tilde{\mu}_{(MLE)}, 10) \quad (5)$$

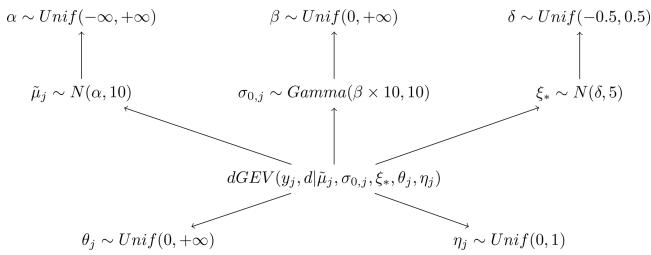
$$\sigma_0 \sim N(\sigma_{0,(MLE)}, 10) \quad T[0,] \quad (6)$$

$$\xi \sim N(\xi_{(MLE)}, 5) \quad T[-0.5, 0.5] \quad (7)$$

$$\theta \sim N(\theta_{(MLE)}, 10) \quad T[0,] \quad (8)$$

$$\eta \sim N(\eta_{(MLE)}, 5) \quad T[0, 1] \quad (9)$$

As  $\sigma_0$ ,  $\theta$  and  $\eta$  are limited parameters, we devise truncated normal distributions. These are normal distributions where either the upper end or lower end is truncated, denoted as  $T[\text{lower end}, \text{upper end}]$ . We limit normal distribution for  $\xi$  between  $-0.5$  and  $0.5$ , which is common practice and recommended for the shape parameter (Martins and Stedinger, 2001; Papalexiou and Koutsoyiannis, 2013). Given that  $\xi$  and  $\eta$  are restricted to a narrow range, we devise smaller variances in the normal priors. We derive confidence intervals from the posterior parameter distributions with 500 MCMC iterations.



**Figure 3.** Diagram of the dGEV-BHM. We devise three different variations of the dGEV-BHM, where the shape parameter  $\xi_* \in \{\xi_d, \xi_{j,d}, \xi_j\}$ .

### 3.3.3 Bayesian hierarchical duration-dependent generalized extreme value distribution

Here we introduce the *Bayesian hierarchical duration-dependent GEV* (in short: dGEV-BHM). In the dGEV-BHM, all stations and durations are incorporated into a single comprehensive model. This model is different from the previously mentioned models, as its parameters share hyperparameters over space and/or durations, making it hierarchical. The base of our model is that of the dGEV (see Eqs. 2 and 3) with different variations in the dimensions of the shape parameter  $\xi$ . We have three different versions of this model: dGEV-BHM- $\xi_d$ , dGEV-BHM- $\xi_j$  and dGEV-BHM- $\xi_{j,d}$ .

In all models, the rescaled location parameter and scale offset share a hyperparameter over space  $\alpha$  and  $\beta$ .  $\tilde{\mu}$  has normal prior, with a standard deviation of 10 and the mean is the hyperprior  $\alpha$ .

$$\tilde{\mu}_j \sim N(\alpha, 10)$$

where  $\alpha$  is an uninformative hyperprior. For the scale offset  $\sigma_{0,j}$  we use a Gamma distribution, with its shape parameter equal  $\beta \cdot 10$  and the inverse scale (or “rate”) parameter equal to 10. In that way, we ensure that the mean of the distribution is equal to  $\beta$ .

$$\sigma_{0,j} \sim \text{Gamma}(\beta \cdot 10, 10)$$

For the spread of the priors we performed a prior selection process to find a well-performing and stable model. We found that narrow spreads led to convergence issues, as this constrained the model too much. On the other hand, wide spreads yielded performance that was similar to the chosen parameters. A standard deviation of 10 was selected for the Normal distribution, as this value proved stable in preliminary tests. The Gamma prior was parameterized as  $\text{Gamma}(K\beta, K)$  to ensure its mean remained fixed at the hyperprior value  $\beta$ . The concentration parameter  $K$  was set to 10, as this value was stable in our tests. We chose a preliminary prior selection process rather than a full prior sensitivity analysis due to the large computational demand with many sub-samples.

Uninformative (improper) hyperpriors allow the data to drive inferences, especially when there is little prior information available (Banerjee and Fuentes, 2012; Cooley et al., 2007). It can still facilitate information sharing across groups through partial pooling and shrinkage effects (Röver and Friede, 2020). Partial pooling allows information to be partially shared between groups, resulting in parameter estimates that are a weighted average of the group-level data and the overall population estimate. Shrinkage effects pull the estimates of group-level parameters towards the overall population mean. This combination allows for data-driven inference while still benefiting from the hierarchical structure of the model, which is particularly useful when prior knowledge is limited.

The innovative aspect of this distribution is that we model the shape parameter of the dGEV distribution, used to model the block maxima of rainfall durations, in three distinct ways:

1. *Hierarchical over duration* ( $\xi_d$ ): The shape parameter varies systematically with the duration under consideration.
2. *Hierarchical over space* ( $\xi_j$ ): The shape parameter changes based on the geographical locality.
3. *Hierarchical over both space and duration* ( $\xi_{j,d}$ ): The shape parameter depends on both the duration and the locality.

For all these variations of the shape parameter, we assume a normal distribution with a narrower standard deviation of 5 and a hyperparameter denoted as  $\delta$ .

$$\xi_* \sim N(\delta, 5)$$

The hyperparameter  $\delta$  itself follows a weakly informative uniform distribution constrained between  $-0.5$  and  $0.5$ :

$$\delta \sim \text{Uniform}(-0.5, 0.5)$$

This setup facilitates partial pooling of information, either through duration, location, or both. Partial pooling allows information to be shared between groups or durations, resulting in more robust estimates of the shape parameter, particularly in cases with limited data. The weakly informative prior on  $\delta$  ensures that the model remains data-driven while still benefiting from the hierarchical framework. We derive confidence intervals from the posterior parameter distributions of the three versions of the dGEV-BHM, with 500 MCMC iterations each.

## 3.4 Goodness of fit and evaluation

### 3.4.1 Sub-sampling strategy

Our aim is to use the large sample size of the CRCM5-LE to evaluate the ability of the six different EVT models to reproduce rare rainfall return levels. Therefore, we randomly sample smaller sub-samples of  $\{30, 40, \dots, 90, 100\}$  years from

**Table 1.** Table of all six EVT models and their estimation method, generation of the confidence intervals and shape parameter dimension (space and/or duration). Shape parameter dimensions: *Global* = one shape parameter spanning across the dimension indices. *Local* = several shape parameters across dimension indices.

Model name	Estimator	Confidence intervals	Shape parameter dimension	
			Duration	Space
GEV L-moments	L-moments	Bootstrap	Local	Local
dGEV-MLE	MLE	Bootstrap	Global	Local
dGEV-Bayesian	Bayesian (MCMC)	Posterior distributions	Global	Local
dGEV-BHM- $\xi_d$	Bayesian (MCMC)	Posterior distributions	Local	Global
dGEV-BHM- $\xi_{j,d}$	Bayesian (MCMC)	Posterior distributions	Local	Local
dGEV-BHM- $\xi_j$	Bayesian (MCMC)	Posterior distributions	Global	Local

the 2000-year CRCM5-LE population. Each sub-sample is randomly drawn 100 times. We then fit the six EVT models to the sub-samples and compare the resulting EVT model outputs to the effective return level of the full 2000-year CRCM5-LE.

### 3.4.2 Anderson Darling

The Anderson Darling (AD) test is a method which analyzes how well empirical data fits a theoretical distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and is more sensitive to discrepancies in the tails (Stephens, 1974), making it suitable to assess extreme value distributions. The mathematical formulation for the AD test is:

$$A^2 = -n - S \quad (10)$$

$$S = \sum_i [(2i - 1)/n] \cdot [\ln F(X_i) + \ln(1 - F(X_{n-i+1}))] \quad (11)$$

where  $n$  is the sample size,  $i$  is the sample index (ordered lowest to highest) and  $F(X)$  is the cumulative distribution. Thus,  $S$  is the discrepancy between the empirical cumulative distribution and the theoretical cumulative distribution, whereby more weight is applied to the differences at the tails of the distribution. The  $p$ -value quantifies the evidence against the hypothesis that the data follows the specified distribution and is commonly obtained from tabulated critical values. However, in this study, we compute the  $p$ -value from the Anderson-Darling test statistic itself using piecewise exponential approximations as implemented in the R package `gnFit` (Saeb, 2018).

We apply the AD test at the significance  $\alpha = 0.05$ . The Anderson-Darling (AD) test is a method commonly applied to evaluate the goodness of fit for the GEV distribution, especially when fitted to annual maxima of precipitation. However, the AD test can only evaluate the quality of the fit between the theoretical GEV distribution and the empirical distribution of the block maxima samples. Although the test's  $p$ -value may indicate a good fit for a given dataset, it does not guarantee that the model will produce accurate extreme,

unseen quantiles. The test also cannot assess how well a limited empirical sample represents the full population of possible extreme rainfall events governed by internal climate variability (Poschlod, 2021). The true performance of an extreme value model lies in its ability to accurately estimate rare return levels that are critical for risk assessment.

For this reason, while distributional tests like the AD test are useful for initially checking the fit, we also evaluate if they are sufficient as the sole metric for validating extreme value models, particularly when the goal is to evaluate tail risk. Our aim is therefore not to validate the models with the AD test, but to investigate if it is reliable in estimating extreme quantiles, which we will assess using our tail risk metrics, introduced in Sect. 3.4.5. To do this, we apply the AD test on the 100 sub-samples of 30 to 100 years generated from the CRCM5-LE population (mentioned in Sect. 3.4.1) and assess the respective GEV and dGEV parameter estimates of each duration and location. In total 720 000 hypothesis tests (8 sub-sample ranges  $\times$  100 sub-samples  $\times$  25 localities  $\times$  6 durations) are performed, and from this, the proportion of tests fulfilling  $p > 0.05$  is determined.

### 3.4.3 Akaike information criterion

The Akaike information criterion (AIC) is a statistical measure for model selection that balances goodness-of-fit with its complexity (Sakamoto et al., 1988). It aims to achieve “parsimony” in a model by selecting the model that balances complexity and likelihood. To effectively use the AIC, it must be applied to a set of candidate models that have been fitted to the same data. The model with the lowest AIC is then chosen as the preferred model (Anderson et al., 1998). The formula for the AIC is:

$$AIC = 2k - 2\log(\hat{L}) \quad (12)$$

where  $k$  represents the number of estimated parameters, serving as a measure of the model's complexity. This term acts a penalty to help prevent overfitting, which is a common issue where a model fits the training data too closely but fails to generalize to new data.  $\hat{L}$  is the maximum value of the likelihood function for the model, which quantifies how well the

model fits the data. Hence, the AIC thereby relies on a single maximum likelihood estimate, whereas the four Bayesian models in our study provide a full posterior distribution of parameters via MCMC instead of a single point estimate. Therefore, as a workaround, we use the posterior means of the Bayesian models for the AIC calculation. In turn, the widely applicable information criterion (WAIC) utilizes the entire posterior distribution of Bayesian models. However, it is only applicable to Bayesian models. As our model setup features four Bayesian models and two frequentist models, we provide the AIC as a consistent measure across all six models and include the WAIC in the supplement.

#### 3.4.4 Effective return levels

An important foundation of this study is the use of effective return levels, which we directly infer from the 2000-year large ensemble. We derive effective return levels, such as the 100-year return level represented by the 20th highest index ( $n_{(20)}$ ) and the 10-year return level represented by the 200th highest index ( $n_{(200)}$ ). To obtain a robust estimate of the effective return level with a measure of uncertainty, we employ a bootstrapping procedure, resampling the entire dataset 500 times with replacement. By computing the return levels from the sub-samples via the six EVT models (see Table 1) and comparing them to the effective return levels from the full CRCM5-LE sample, we can determine how accurately each statistical model replicates the ensemble data. This procedure is paramount for validating the models' reliability and robustness.

#### 3.4.5 Error metrics: relative error and return level concordance

We assess the accuracy of an EVT model, by calculating the relative error (RE) between the modeled return level and the effective return level. Therefore, for a given return period, we determine the percentage difference between the model return level and the effective return level. For each EVT model, we compute the RE for 10- and 100-year return levels for sub-samples sizes ranging between 30 and 100 years. Furthermore, we assess the 80 % confidence intervals derived via bootstrapping for the frequentist models and the posterior parameter distribution for the Bayesian models on each of the 100 sub-samples. Thereby, we assess the uncertainty ranges indicating the robustness of the EVT models. The choice of 80 % confidence intervals helps visualize the RE better in the figures.

The return level concordance (RLC) describes the agreement between the EVT model's computed return levels from the 100 sub-samples and the confidence intervals derived from the large ensemble climate data. The RLC quantifies the fraction of the EVT model 100-year return levels that are within the 95 % confidence interval of the effective 100-year return level (see Fig. 4).

#### 3.5 Intensity-duration-frequency (IDF) curves

Lastly, we construct IDF curves from the parameter estimates of dGEV models. The rainfall return level at a given duration  $d$  and return period  $T$  is:

$$x(T) = \mu(d) + \frac{\sigma(d)}{\xi} \times \left\{ \left[ -\log(1 - 1/T) \right]^{-\xi} - 1 \right\} \quad (13)$$

where  $\xi$  is the shape parameter, and  $\sigma(d)$  and  $\mu(d)$  are the scale and location parameters respectively given at specific durations. The scale offset  $\sigma_0$  and rescaled location parameter  $\tilde{\mu}$  are both incorporated into the duration-dependent parameters (Koutsoyiannis et al., 1998; Fauer et al., 2021):

$$\sigma(d) = \sigma_0 \times (d + \theta)^{-\eta} \quad (14)$$

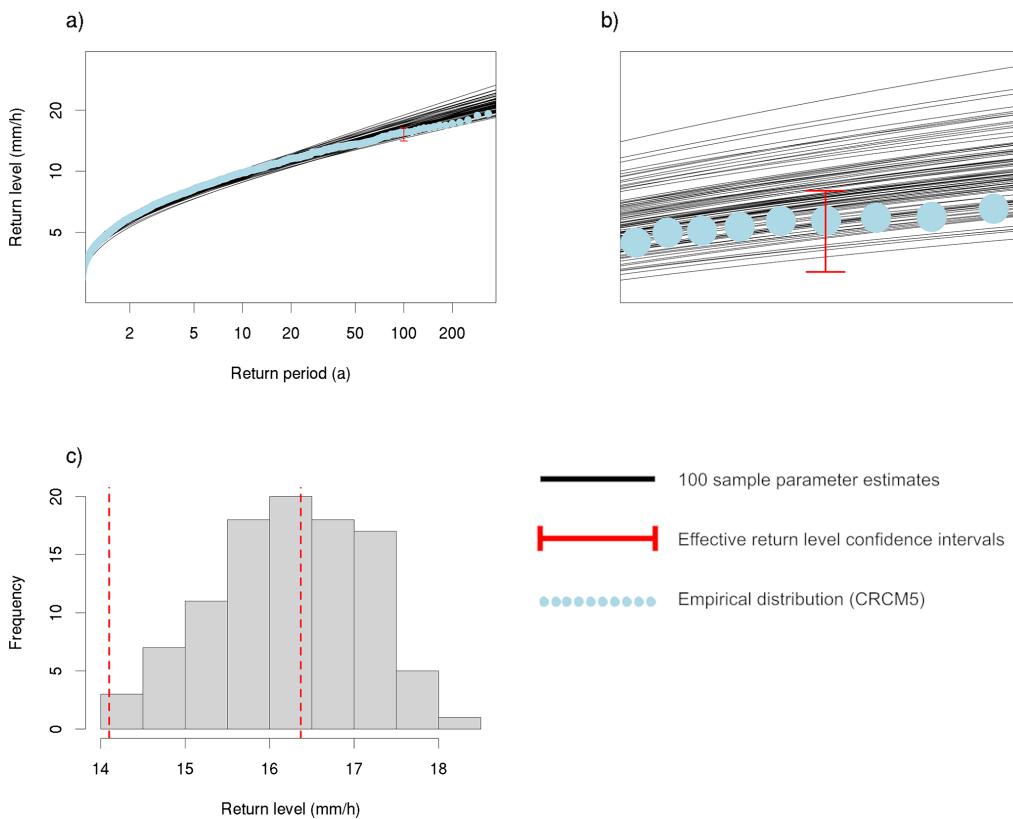
$$\mu(d) = \tilde{\mu} \times \sigma(d) \quad (15)$$

In the Bayesian hierarchical models both the scale offset  $\sigma_0$  and rescaled location parameter  $\tilde{\mu}$  share hyperparameters across space. In the two models dGEV-BHM- $\xi_d$  and dGEV-BHM- $\xi_{j,d}$  the shape parameter  $\xi_d$  is modeled along the duration  $d$ , and therefore requires interpolation to construct the resultant IDF curve between the discrete durations. To construct a continuous IDF curve between the durations, an initial approximation is generated by linearly interpolating the mean of the shape parameter  $\xi$  between the discrete durations. This preliminary curve is then modified using an adjustment step. The required adjustment – defined as the difference between the approximate curve and the actual quantiles calculated using the full posterior distribution of  $\xi$  at the discrete durations – is then interpolated and applied to the initial curve.

### 4 Results

#### 4.1 Anderson-Darling test

First, we present the results of the AD test, as the goodness-of-fit evaluation would be the typical next step after the fitting of EVT models. Figure 5 illustrates the fraction of passed AD tests for all 25 localities, 6 durations, and 100 sub-samples across the 8 sub-sample sizes between 30 and 100 years. Given the significance level of  $\alpha = 0.05$ , a 95 % fraction of passed AD tests would be expected if all data samples came from the tested theoretical distribution specified by the EVT model fits. We see that the overall test results decrease with increasing sample sizes, as the AD test gains more power to reject the null hypothesis detecting deviations between the sample and the theoretical distribution (Shin et al., 2012). The AD test diagnoses the highest rejection rates for the dGEV-BHM- $\xi_d$ . The dGEV-BHM- $\xi_j$ , dGEV-Bayesian, and dGEV-MLE reach 95 % acceptance for 30-year samples dropping to 93 % for 100-year samples. The dGEV-BHM- $\xi_{j,d}$  and GEV-Lmoments show the lowest rejection rates. We note that this performance can be linked to



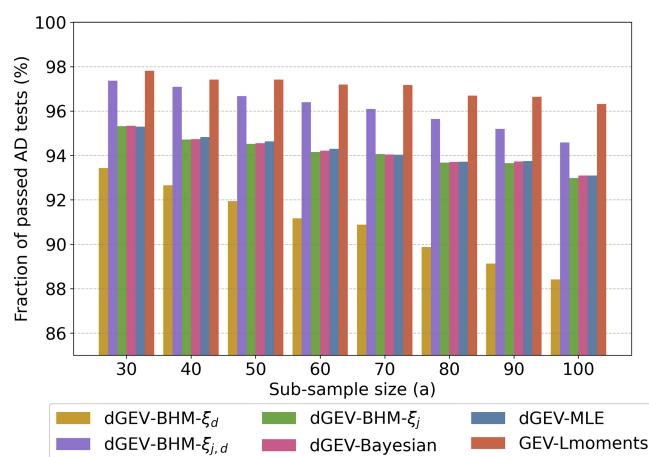
**Figure 4.** Conceptual representation of the 100-year return level concordance (RLC) for one locality (Augsburg) and duration (6 h) and sub-sample of 40 years. (a) The return levels based on a EVT model (black lines) are compared to the confidence interval (red bar) of the effective 100-year return level, which is derived from the full 2000-year CRCM5-LE (light blue dots). (b) Zoom-in to the RLC at the 100-year return level. In this example the RLC would amount to 55 %. (c) Histogram of the RLC, the red dashed lines are the confidence intervals of the effective 100-year return level.

the flexibility of the EVT models. The GEV-Lmoments estimates location, scale, and shape parameters separately for each locality and duration providing the highest flexibility to fit the GEV to the samples. With the duration dependence, the dGEV limits the flexibility. The dGEV-BHM- $\xi_{j,d}$  is the most flexible version allowing the shape parameter to vary over space and duration (see Table 1). In turn, the dGEV-BHM- $\xi_d$  fixes the shape parameter over space resulting in the highest rejection rate of the AD test.

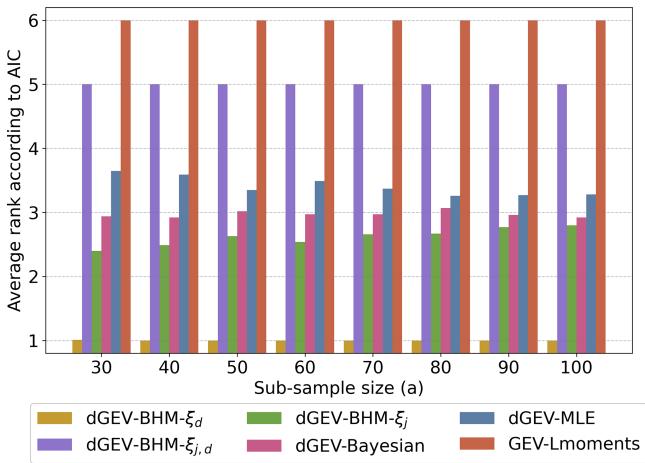
#### 4.2 Akaike information criterion

The average ranks according to the AIC values are illustrated in Fig. 6. The dGEV-BHM- $\xi_d$  model consistently ranks first (the lowest average rank) across all sub-sample sizes, indicating it is the best model choice according to the Akaike Information Criterion (AIC).

The second best model is dGEV-BHM- $\xi_j$ , which has the second-lowest average rank across all sub-sample sizes. Following that, dGEV-Bayesian and dGEV-MLE are the third and fourth best, respectively. The dGEV-BHM- $\xi_{j,d}$  and



**Figure 5.** Fraction of passed Anderson-Darling tests for the six EVT models and 100 sub-samples at sizes reaching from 30 to 100 years. The fractions are averaged over the 25 localities and 6 durations.



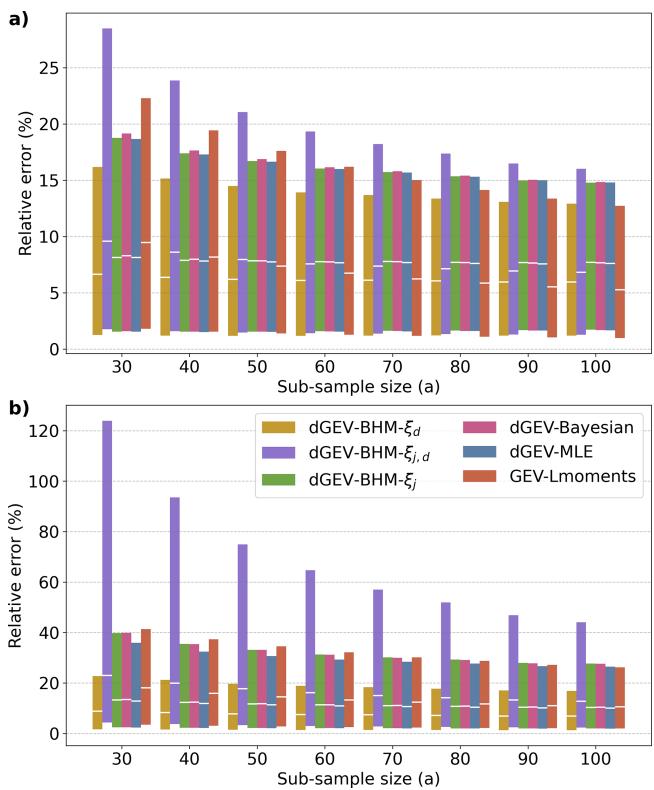
**Figure 6.** Average ranks according to the AIC values for all models and all their sub-sample sizes.

GEV-Lmoments models consistently rank as fifth and sixth, respectively.

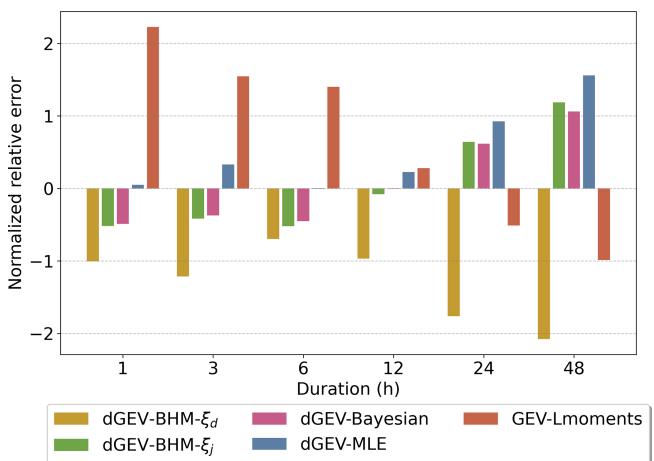
#### 4.3 Relative error

The relative error is presented in Fig. 7 for 10-year and 100-year return levels. The performance of EVT models reveals distinct patterns across varying sub-sample sizes. For the 10-year RE (Fig. 7a)], dGEV-BHM- $\xi_d$  consistently demonstrates superior performance across sub-sample sizes up to 70 years, with GEV-Lmoments showing the lowest RE for sample sizes of 80 to 100 years. The dGEV-BHM- $\xi_j$ , dGEV-Bayesian, and dGEV-MLE score very similarly and show low sensitivity to the sample size. Conversely, the dGEV-BHM- $\xi_{j,d}$  model exhibits the poorest performance for low sample sizes up to 50 years, but improves for larger samples. In contrast, for the 100-year RE (Fig. 7b), dGEV-BHM- $\xi_d$  consistently emerges as the most accurate model across all sub-sample sizes. The dGEV-BHM- $\xi_{j,d}$  and GEV-Lmoments are identified as the least accurate. Again, dGEV-BHM- $\xi_j$ , dGEV-Bayesian, and dGEV-MLE show an equivalent accuracy. The robustness of the six EVT models behaves similar for the 10-year and 100-year return levels. The error ranges of the dGEV-BHM- $\xi_{j,d}$  are by far the largest showing high sensitivity to the sample size. In turn, the dGEV-BHM- $\xi_d$  is diagnosed the most robust model consistently across the sub-sample sizes.

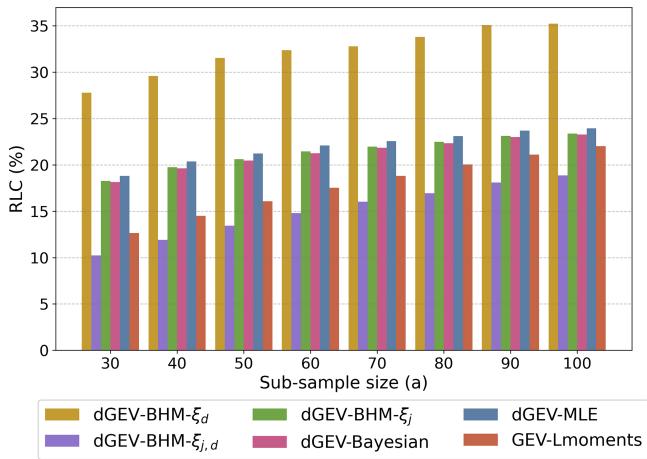
In Fig. 8 we present the location average 100-year RE for sub-sample size  $n = 30$  at different durations. We normalize the RE for each model across five models for each duration. We intentionally leave the dGEV-BHM- $\xi_{j,d}$  out, as the results of RE are so high that they distort the standardized results. Across all durations, the dGEV-BHM- $\xi_d$  consistently shows the lowest RE. GEV-Lmoments shows high errors for the short durations of 6 h and below, but comparably low REs for 24 and 48 h. In contrast, dGEV-BHM- $\xi_j$ ,



**Figure 7.** The relative error (RE) of the (a) 10-year return levels and (b) 100-year return levels averaged across all localities and durations. The bars indicate the range of the inner 80 % of the 100 sub-samples and 500 bootstrap or MCMC iterations, where the white marker represents the median.



**Figure 8.** Average 100-year relative error (RE) across durations for sub-sample size of 30 years. Note that the the dGEV-BHM- $\xi_{j,d}$  is not shown as the high REs distort the standardized results.



**Figure 9.** Return level concordance (RLC) for 100-year return levels, for the six EVT models and sub-sample sizes ranging from 30 to 100 years. The RLC is averaged over the localities and durations.

dGEV-Bayesian, and dGEV-MLE show the tendency to perform better for shorter than for longer durations.

#### 4.4 Return level concordance

Figure 9 features the RLC for the 100-year return level. For all six EVT models, the RLC increases with increasing sample size. The dGEV-BHM- $\xi_d$  can achieve the highest RLC between 27 % and 35 % outperforming the other five EVT models by around 10 %, proving its robustness. The dGEV-BHM- $\xi_j$ , dGEV-Bayesian, and dGEV-MLE perform similarly in a range from 17 % to 24 %. In contrast, both the dGEV-BHM- $\xi_{j,d}$  and GEV-Lmoments consistently exhibit the lowest RLC between 10 % and 22 %.

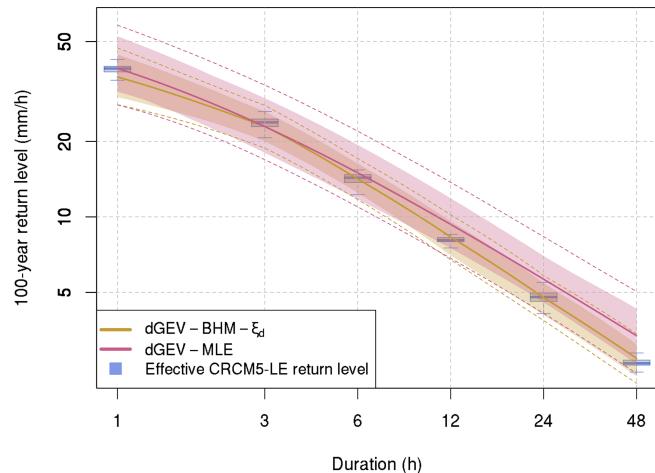
#### 4.5 Intensity-duration-frequency (IDF) curves

Figure 10 illustrates IDF curves for the best performing model dGEV-BHM- $\xi_d$  and the state-of-the-art dGEV-MLE for one example locality in Würzburg (see Figs. S26–S50 in the Supplement for the IDF curves of the other localities). The IDF curves are constructed on the sub-sample size of  $n = 30$  reflecting typical data availability for sub-daily rainfall durations in Germany. In this case, the dGEV-BHM- $\xi_d$  can reproduce the 100-year return levels across all durations, while the dGEV-MLE overestimates for the durations of 6 h and above. The 95 % confidence intervals of the dGEV-BHM- $\xi_d$  are narrower than for the dGEV-MLE indicating its robustness even under data scarce conditions.

### 5 Discussion

#### 5.1 Goodness-of-fit and information criteria

Our experiment reveals that the rejection rates of the Anderson-Darling test align with the flexibility of the EVT



**Figure 10.** IDF curves of the dGEV-BHM- $\xi_d$  and dGEV-MLE for the locality Würzburg and the 100-year return period. IDF curves are constructed based on a sub-sample size of 30 years. The dashed lines depict the 95 % confidence intervals for all 100 sub-samples and combined MCMC/Bootstrap iterations, whilst the polygons represent the 95 % confidence intervals for the 100 sub-sample means. The boxplots represent the ranges of the 100-year return level derived from resampling the 2000-year CRCM5-LE with replacement.

models, where more flexible models are rewarded by lower rejection rates. Conversely, the dGEV-BHM- $\xi_d$  showing the highest rejection rates performs better than the other models in terms of the RE and RLC. Hence, we argue that the AD test on small samples of 30 to 100 years is not well-suited to assess the quality of how well an EVT model fits the full sample of 2000 years. We attribute this observed pattern to a lack of robustness to internal climate variability. Specifically, if a small sample, due to random sampling, does not accurately represent the full 2000-year sample, flexible EVT models are prone to overfitting it. This is rewarded by a successful AD test on the small sample, but diagnosed by a potential higher RE and lower RLC when considering the full 2000-year sample. This is a clear manifestation of the bias-variance trade-off: more flexible models have lower bias but higher variance, leading to poor generalization. While a practitioner with a limited observational sample may only see a good AD test result and interpret it as model success, our study demonstrates the inherent limitation of this goodness-of-fit testing when evaluating the true performance of extreme quantiles. Hence, this is driven by the variability of the annual maxima of precipitation within the 2000-year sample. The source of this variability is the chaotic nature of the climate system, or in other words, internal climate variability, which is a huge source of uncertainty in assessments of extreme precipitation (Aalbers et al., 2018; Bhatia and Ganguly, 2019; Martel et al., 2018; Poschlod et al., 2021; Poschlod and Ludwig, 2021). We emphasize that one of the 100 sub-samples could be a time series of observations, while the EVT models are

supposed to estimate the occurrence probabilities of extreme precipitation within the boundaries of internal climate variability. Our findings do not question the usefulness of the AD test, but show that GEV models that pass the AD test are not necessarily suitable for predicting very rare probabilities of occurrence.

The assessment of the Bayesian EVT models via AIC reveals results contrast with the evaluation of the AD test. In the AD test, the dGEV-BHM- $\xi_d$  has the highest rejection rates, whereas AIC ranks are consistently highest across all sub-sample sizes. The AIC aligns well with return level metrics (RE and RLC), yet it penalizes the frequentist models more, especially those with more parameters  $k$ . The models dGEV-BHM- $\xi_j$ , dGEV-Bayesian and dGEV-MLE all have the same number of parameters. Therefore, the observed differences in their averaged AIC ranks are due to the differences in the models' likelihood. We see that the Bayesian models, especially the hierarchical ones, increase the likelihood. However, the models' performances in terms of Relative Error (RE) and Return Level Curve (RLC) are similar. This suggests that a model's likelihood itself is not a good indicator of its predictive skill for tail risk.

## 5.2 Accuracy and robustness of the six EVT models

Based on the evaluation of the RE and RLC, the dGEV-BHM- $\xi_d$  is the most accurate and robust model. We can attribute this to its efficient and robust representation of the shape parameter  $\xi$ , which is kept constant across all localities. The model therefore reduces complexity by minimizing overfitting, and improves generalization. By assuming a constant shape parameter over space, the dGEV-BHM- $\xi_d$  effectively addresses the estimation uncertainty of the shape parameter to potential outlier sub-samples, which is a similar finding to that of Poschlod (2021), Poschlod and Koh (2024), Shehu et al. (2023) over the study area. This also confirms Ragulina and Reitan (2017), showing that small regions often exhibit similar shape parameters. Further, the choice of the dGEV-BHM- $\xi_d$  as best model indicates that a shape parameter flexible over durations from 1 to 48 h is beneficial in our experiment. There, heavier tails are diagnosed for short durations ( $\xi_{1\text{h}} = 0.15$  and  $\xi_{3\text{h}} = 0.2$ ) as suggested by Papalexiou et al. (2018), while the shape parameter for longer durations decreases to 0.0973 for  $\xi_{24\text{h}}$  and 0.0644 for  $\xi_{48\text{h}}$ . However, in different study areas or climate models, the shape parameter might not significantly differ across durations (Alaya et al., 2020; Overeem et al., 2008). In contrast, the dGEV-BHM- $\xi_{j,d}$  model, which allows the shape parameter to vary across localities and durations, shows the highest RE and lowest RLC. Our experiment with 25 localities and sample sizes of 100 years or less are not sufficient to robustly fit the most complex model dGEV-BHM- $\xi_{j,d}$ . This might however be different for a larger amount of data, i.e. larger sample sizes and more localities. The dGEV-BHM- $\xi_j$ , dGEV-Bayesian, and dGEV-MLE achieve a comparable performance in RE

and RLC. The additional complexity of the Bayesian estimation or Bayesian hierarchical framework does not add value over the dGEV-MLE. The GEV-Lmoments is the simplest approach and shows a high sensitivity to low sample sizes. While it is well suited to assess moderate extremes with sufficient sample sizes (see Fig. 7a; Poschlod, 2021), we do not recommend the application for the assessment of rare extremes.

The study finds that while more complex models such as our dGEV-BHMs show superior accuracy and robustness in estimating rare precipitation events, their practical application involves a trade-off with usability. Many practitioners in the field, particularly those dealing with limited data, currently rely on simpler models such as the standard GEV-Lmoments or Gumbel distributions. Our proposed dGEV-BHMs offers a robust alternative for developing IDF curves under small sample sizes. However, its implementation requires a more computationally intensive Bayesian framework compared to traditional methods like Maximum Likelihood Estimation (MLE) or L-moments. A key benefit of our approach is its ability to reduce the amount of parameters, which improves generalization and minimizes overfitting, especially with small sample sizes. This makes the model a more reliable tool for risk assessment, even if it requires a higher initial investment in computational resources.

## 5.3 Limitations and possible improvements

To compute all Bayesian models (including hierarchical) we use the “No-U-turn-sampler” (NUTS) provided by Stan and R package `rstan` (Stan Development Team, 2025), which is a form of Hamiltonian Monte Carlo simulation. This is computationally more demanding than many other methods, such as maximum likelihood estimation or L-moments. The EVT models in this study are stationary and did not account for non-stationarity, which is very important for incorporating climate change effects on extreme rainfall. For the chosen 40-year time period of 1980–2019 signals of intensifying extreme rainfall can already be diagnosed in the large ensemble (Lang and Poschlod, 2024; Poschlod, 2022). However, most operational EVT models based on observational data are stationary (Van de Vyver, 2012; Shehu et al., 2023). Hence, with our focus on the consistency across durations and the Bayesian hierarchical setup of the dGEV, we simplifyingly assume stationarity. For longer time periods with stronger climate change signals or for climate projections, an implementation of non-stationarity is beneficial using temperature-related co-variates (Agilan and Umamahesh, 2017). Other temporal co-variates relating to large-scale atmospheric patterns such as the North Atlantic Oscillation (NAO) may further add value as the NAO influences extreme precipitation in Europe (Fauer and Rust, 2023; Steirou et al., 2019; Yiou and Nogaj, 2004). Shehu et al. (2023) propose to estimate dGEV parameters locally for rain gauges in Germany and then apply kriging with external drift in a next step for regionalization.

Our proposed framework can potentially assist as a robust alternative in the estimation of the dGEV parameters. Fauer et al. (2021) add additional parameters to the dGEV framework, in the form of “flattening”, which acts to flatten the IDF curve and “multiscaling”, which adds another parameter to the duration exponents. This increases the flexibility of the dGEV to extend from even sub-hourly to multi-day durations. As the CRCM5 output is stored at hourly duration, our study is restricted to durations above one hour. Hence, this modification of the dGEV is not considered here. Furthermore, Ulrich et al. (2021) maximize data utilization by using monthly, rather than annual maxima. While this approach increases the amount of data used, it may include values that are not truly extreme. Another approach for spatial modeling is Regional Frequency Analysis (RFA) (Hosking and Wallis, 1997). RFA involves grouping sites within a region based on hydrological, meteorological, and geographical characteristics, and then analyzing extreme precipitation statistics collectively rather than individually. While we did not include a formal RFA model in this study due to the already comprehensive and complex set of models, we acknowledge that it represents a valuable alternative for spatial modeling of extreme events. The implementation of a dGEV-RFA could be valuable, especially given its focus on robust estimation for data-scarce regions (Kim et al., 2020).

The assumption of similar extreme value characteristics across localities in the dGEV-BHM- $\xi_d$  model restricts its use to smaller regions, rendering it less effective for large-scale applications. However, different dGEV-BHM- $\xi_d$  models could be set up for clusters with similar characteristics of extreme precipitation. Furthermore, one could allow  $\xi$  to vary over space following a co-variate (Dyrrdal et al., 2015; Lehmann et al., 2016). In a study similar to ours, Jalbert et al. (2022) also employed a Bayesian hierarchical dGEV model with a spatially constant shape parameter for each duration. While there are many commonalities between our study and theirs, key distinctions must also be highlighted. Both studies utilize the CRCM5 climate model. The most immediate difference is the geographical focus: Jalbert et al. (2022) is centered on Canada, whereas our study is conducted over Southern Germany. Their approach leverages the spatial coverage of the CRCM5 to interpolate extremes across grid cells, effectively pooling information from observational stations to improve estimates at ungauged sites. Conversely, our study uses the sample size of the CRCM5’s large ensemble, which provides a framework for evaluating the tail risk prediction skill of different models. The leave-one-out cross-validation (LOOCV) employed by Jalbert et al. (2022) is a suitable validation strategy, given the spatial nature of their study. Using Cramér–von Mises test used in the LOOCV is appropriate for assessing the overall goodness-of-fit of a distribution. On the other hand, as our study is more concerned with the accuracy of high quantile estimates, the AD test is more applicable due to its heightened sensitivity to discrepancies in the tails of a distribution.

The findings by Jalbert et al. (2022) and recent work by Poschlod and Koh (2024) show that the combination of high-resolution climate models and observations is beneficial for the representation of extreme rainfall at high spatial detail. Both studies demonstrate that adding spatial covariates from climate models improves the performance of spatial GEV models over models that use only topographical information, suggesting that this approach could be beneficial in other frameworks like dGEV as well. Lehmann et al. (2016) set up a Bayesian hierarchical dGEV model based on rain gauges adding information from a regional climate model as climatological co-variate over Australia. This allows for a robust regionalization and an investigation of climate change impacts on extreme rainfall levels by calculating IDF curves under future climate. Dyrrdal et al. (2015) use a Bayesian hierarchical model which was not duration-dependent for observations in Norway. Both Dyrrdal et al. (2015) and Jalbert et al. (2022) incorporated Gaussian processes (GPs) in their model design. This assumes the parameter values at different locations are jointly Gaussian, and the correlation between them decreases as the distance increases. Whilst this kind of spatial modeling is a powerful tool, we chose to focus on point scale analysis to directly compare and thoroughly investigate different estimation methods, particularly for small samples. In this context, we also explored incorporating physical co-variates (altitude and mean annual precipitation) into our BHMs. However, these did not improve the model performance and, in some CRCM5-LE sub-samples, lead to convergence issues. Given these challenges and the specific focus on the research question, we decided not to proceed with a full spatial model. Nevertheless, we agree that GPs represent a promising avenue for future work on the regionalization of our BHM, especially for larger datasets with more spatial information.

Even though both our CRCM5 evaluation (Fig. 2) and Poschlod et al. (2021) have shown that the CRCM5 performs well for sub-daily and daily extreme precipitation despite the parameterization of convection, Poschlod (2021) has shown in a direct comparison of a convection-permitting model (CPM) with the CRCM5 that the higher resolution and explicit simulation of convective processes add value to the representation of extreme precipitation, especially over complex topography. Generally, CPMs are a promising tool in climate modeling, offering improved, high-resolution simulations of deep convective processes, which are parameterized in lower-resolution models (Fosser et al., 2020; Lucas-Picher et al., 2021; Meredith et al., 2015). This increased detail requires a higher computational cost, making large ensembles of CPMs currently unaffordable for decadal to centennial simulation periods (Prein et al., 2015), while Bassett et al. (2020) pioneered a CPM large ensemble over London for a 4-month period. Consequently, there is a fundamental trade-off between the high resolution and accuracy of CPMs and the long time scales and large sample sizes that are necessary for robust climate projections and uncertainty analy-

sis, which are typically addressed by large ensembles of less detailed models.

## 6 Conclusions

With the large ensemble as *perfect model experiment* featuring the sub-sampling strategy, we are able to assess the predictive power of EVT models. The 30 to 100-year sample sizes and the 25 localities thereby mimic the typical data availability conditions for sub-daily rainfall. Within these boundaries, the dGEV-BHM- $\xi_d$ , a Bayesian hierarchical model of the duration-dependent dGEV with a shape parameter flexible over durations, but fixed over space outperforms existing state-of-the-art GEV and dGEV models. With the focus of the analysis on rare events (100-year return periods), we can recommend this approach as accurate and robust methodology to derive IDF curves under data-scarce conditions. This can be beneficial for the generation of observation-based rainfall return levels (Shehu et al., 2023) or the analysis of extreme precipitation in high-resolution climate simulations, which also typically cover limited time periods (Poschlod, 2021). Furthermore, with the help of the large ensemble and the sub-sampling strategy, we show that traditional goodness-of-fit measures, such as Anderson-Darling test, are not well suited to choose EVT models on small sample sizes with regard to rare event probabilities. Here, the AIC proves as a useful tool in assessing the performance of EVT models. In the current state, the dGEV-BHM- $\xi_d$  model is characterized by its simplicity, as only sub-daily precipitation data from several stations with similar characteristics are required. Next steps may involve the implementation of non-stationarity and regionalization with the help of co-variates, where the CRCM5-LE can also act as a testbed providing future climate projections and full spatial coverage.

**Code and data availability.** The code used in this study is available from the corresponding author upon reasonable request. The CRCM5 precipitation data are openly available from [https://climex-data.srv.lrz.de/Public/CanESM2\\_driven\\_50\\_members/pr/](https://climex-data.srv.lrz.de/Public/CanESM2_driven_50_members/pr/) (last access: 5 March 2025). The observations by the German Weather Service are openly available from [https://opendata.dwd.de/climate\\_environment/CDC/observations\\_germany/climate/hourly/precipitation/historical/](https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/precipitation/historical/) (last access: 6 April 2025).

**Supplement.** The supplement related to this article is available online at <https://doi.org/10.5194/ascmo-12-1-2026-supplement>.

**Author contributions.** All three authors conceptualized the study. AR wrote the software code, performed the formal analysis, investigation, and visualization under supervision of BP and JS. AR and BP developed the methodology and are responsible for data

curation. AR and BP prepared the original manuscript. JS acquired funding and contributed to the review and editing of the draft.

**Competing interests.** The contact author has declared that none of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

**Acknowledgements.** This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2037 “CLICCS – Climate, Climatic Change, and Society” – project no. 390683824, a contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg. The authors utilized Gemini from Alphabet and GPT-4 from OpenAI to enhance language and readability. The authors thank Petra Friederichs for fruitful discussions on the manuscript.

**Financial support.** This research has been supported by the Deutsche Forschungsgemeinschaft (grant no. EXC 2037, project no. 390683824).

**Review statement.** This paper was edited by Mark Risser and reviewed by Whitney Huang and Jean-Luc Martel.

## References

- Aalbers, E. E., Lenderink, G., van Meijgaard, E., and van den Hurk, B. J.: Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability?, *Climate Dynamics*, 50, 4745–4766, 2018.
- Agilan, V. and Umamahesh, N.: What are the best covariates for developing non-stationary rainfall intensity-duration-frequency relationship?, *Advances in Water Resources*, 101, 11–22, 2017.
- Alaya, M. B., Zwiers, F., and Zhang, X.: An evaluation of block-maximum-based estimation of very long return period precipitation extremes with a large ensemble climate simulation, *Journal of Climate*, 33, 6957–6970, 2020.
- Anderson, D., Burnham, K., and White, G.: Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies, *Journal of Applied Statistics*, 25, 263–282, 1998.
- Ban, N., Rajczak, J., Schmidli, J., and Schär, C.: Analysis of Alpine precipitation extremes using generalized extreme value theory in convection-resolving climate simulations, *Climate Dynamics*, 55, 61–75, 2020.

Banerjee, S. and Fuentes, M.: Bayesian modeling for large spatial datasets, *WIREs Computational Statistics*, 4, 59–66, <https://doi.org/10.1002/wics.187>, 2012.

Bassett, R., Young, P., Blair, G., Samreen, F., and Simm, W.: A large ensemble approach to quantifying internal model variability within the WRF numerical model, *Journal of Geophysical Research: Atmospheres*, 125, e2019JD031286, <https://doi.org/10.1029/2019JD031286>, 2020.

Beneyto, C., Aranda, J. Á., Benito, G., and Francés, F.: New approach to estimate extreme flooding using continuous synthetic simulation supported by regional precipitation and non-systematic flood data, *Water*, 12, 3174, <https://doi.org/10.3390/w12113174>, 2020.

Berger, J. O.: Statistical decision theory and Bayesian analysis, Springer-Verlag, ISBN 9781475742862, 1985.

Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., Yiou, P., and Zscheischler, J.: Advancing research on compound weather and climate events via large ensemble model simulations, *Nature Communications*, 14, 2145, <https://doi.org/10.1038/s41467-023-37847-5>, 2023.

Bhatia, U. and Ganguly, A. R.: Precipitation extremes and depth-duration-frequency under internal climate variability, *Scientific reports*, 9, 9112, <https://doi.org/10.1038/s41598-019-45673-3>, 2019.

Brönnimann, S., Rajczak, J., Fischer, E. M., Raible, C. C., Rohrer, M., and Schär, C.: Changing seasonality of moderate and extreme precipitation events in the Alps, *Nat. Hazards Earth Syst. Sci.*, 18, 2047–2056, <https://doi.org/10.5194/nhess-18-2047-2018>, 2018.

Bücher, A., Lilienthal, J., Kinsvater, P., and Fried, R.: Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis, *Extremes*, 24, 325–348, 2021.

Cannon, A. J. and Innocenti, S.: Projected intensification of sub-daily and daily rainfall extremes in convection-permitting climate model simulations over North America: implications for future intensity-duration-frequency curves, *Nat. Hazards Earth Syst. Sci.*, 19, 421–440, <https://doi.org/10.5194/nhess-19-421-2019>, 2019.

Chib, S. and Greenberg, E.: Understanding the metropolis-hastings algorithm, *The American Statistician*, 49, 327–335, 1995.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P.: An introduction to statistical modeling of extreme values, vol. 208, Springer, ISBN 9781849968744, 2001.

Congdon, P. D.: Bayesian hierarchical models: with applications using R, Chapman and Hall/CRC, ISBN 9780429113352, 2019.

Cooley, D. and Sain, S. R.: Spatial Hierarchical Modeling of Precipitation Extremes From a Regional Climate Model, *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 381–402, <https://doi.org/10.1007/s13253-010-0023-9>, 2010.

Cooley, D., Nychka, D., and Naveau, P.: Bayesian Spatial Modeling of Extreme Precipitation Return Levels, *Journal of the American Statistical Association*, 102, 824–840, <https://doi.org/10.1198/016214506000000780>, 2007.

De Paola, F., Giugni, M., Pugliese, F., Annis, A., and Nardi, F.: GEV parameter estimation and stationary vs. non-stationary analysis of extreme rainfall in African test cities, *Hydrology*, 5, 28, <https://doi.org/10.3390/hydrology5020028>, 2018.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, *Nature Climate Change*, 10, 277–286, 2020.

Deutscher Wetterdienst: CDC – Climate Data Center, Deutscher Wetterdienst [data set], <https://cdc.dwd.de/portal/> (last access: 6 April 2025), 2022.

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., and Stordal, F.: Bayesian hierarchical modeling of extreme hourly precipitation in Norway, *Environmetrics*, 26, 89–106, <https://doi.org/10.1002/env.2301>, 2015.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, Chapman and Hall/CRC, ISBN 9780429246593, 1994.

Fauer, F. S. and Rust, H. W.: Non-stationary large-scale statistics of precipitation extremes in central Europe, *Stochastic Environmental Research and Risk Assessment*, 37, 4417–4429, 2023.

Fauer, F. S., Ulrich, J., Jurado, O. E., and Rust, H. W.: Flexible and consistent quantile estimation for intensity-duration-frequency curves, *Hydrol. Earth Syst. Sci.*, 25, 6479–6494, <https://doi.org/10.5194/hess-25-6479-2021>, 2021.

Fauer, F. S., Ulrich, J., Mack, L., Jurado, O. E., Ritschel, C., Detring, C., and Joedicke, S.: IDF: Intensity-Duration-Frequency Curves, r package version 2.1.2, CRAN [code], <https://doi.org/10.32614/CRAN.package.IDF>, 2022.

Fereshtehpour, M. and Najafi, M. R.: Urban stormwater resilience: Global insights and strategies for climate adaptation, *Urban Climate*, 59, 102290, <https://doi.org/10.1016/j.uclim.2025.102290>, 2025.

Fischer, E., Sippel, S., and Knutti, R.: Increasing probability of record-shattering climate extremes, *Nature Climate Change*, 11, 689–695, 2021.

Fisher, R. A. and Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample, in: *Mathematical proceedings of the Cambridge philosophical society*, Cambridge University Press, 24, 180–190, <https://doi.org/10.1017/S0305004100015681>, 1928.

Fosser, G., Kendon, E. J., Stephenson, D., and Tucker, S.: Convection-permitting models offer promise of more certain extreme rainfall projections, *Geophysical Research Letters*, 47, e2020GL088151, <https://doi.org/10.1029/2020GL088151>, 2020.

Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H. X., Guerreiro, S., Haerter, J. O., Kendon, E. J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., and Zhang, X.: Anthropogenic intensification of short-duration rainfall extremes, *Nature Reviews Earth & Environment*, 2, 107–122, 2021.

Gelman, A.: Data analysis using regression and multilevel/hierarchical models, Cambridge university press, 2007.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: Bayesian Data Analysis, 3rd Edn., CRC Press, ISBN 978-1-4398-4095-5, google-Books-ID: ZXL6AQAAQBAJ, 2013.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W., Shebalin, P., Vrac, M., Witt, A., and Zaliapin, I.: Extreme events: dynam-

ics, statistics and prediction, *Nonlin. Processes Geophys.*, 18, 295–350, <https://doi.org/10.5194/npg-18-295-2011>, 2011.

Gilleland, E. and Katz, R. W.: extRemes 2.0: An extreme value analysis package in R, *Journal of Statistical Software*, 72, 1–39, 2016.

Gnedenko, B.: Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of mathematics*, 44, 423–453, 1943.

Gupta, V. K. and Waymire, E.: Multiscaling properties of spatial rainfall and river flow distributions, *Journal of Geophysical Research: Atmospheres*, 95, 1999–2009, 1990.

Hamdi, Y., Haigh, I. D., Parey, S., and Wahl, T.: Preface: Advances in extreme value analysis and application to natural hazards, *Nat. Hazards Earth Syst. Sci.*, 21, 1461–1465, <https://doi.org/10.5194/nhess-21-1461-2021>, 2021.

Haslinger, K., Breinl, K., Pavlin, L., Pistotnik, G., Bertola, M., Olefs, M., Greilinger, M., Schöner, W., and Blöschl, G.: Increasing hourly heavy rainfall in Austria reflected in flood changes, *Nature*, 639, 667–672, <https://doi.org/10.1038/s41586-025-08647-2>, 2025.

Hoffman, M. D. and Gelman, A.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.*, 15, 1593–1623, 2014.

Hosking, J. R.: L-moments: analysis and estimation of distributions using linear combinations of order statistics, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 52, 105–124, 1990.

Hosking, J. R. M. and Wallis, J. R.: Regional frequency analysis, Cambridge University Press, ISBN 9780511529443, 1997.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, 27, 251–261, 1985.

Huang, W. K., Stein, M. L., McInerney, D. J., Sun, S., and Moyer, E. J.: Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions, *Adv. Stat. Clim. Meteorol. Oceanogr.*, 2, 79–103, <https://doi.org/10.5194/ascmo-2-79-2016>, 2016.

Huang, W. K., Monahan, A. H., and Zwiers, F. W.: Estimating concurrent climate extremes: A conditional approach, *Weather and Climate Extremes*, 33, 100332, <https://doi.org/10.1016/j.wace.2021.100332>, 2021.

Jalbert, J., Genest, C., and Perreault, L.: Interpolation of precipitation extremes on a large domain toward idf curve construction at unmonitored locations, *Journal of Agricultural, Biological and Environmental Statistics*, 27, 461–486, 2022.

Jung, C. and Schindler, D.: Precipitation Atlas for Germany (GePrA), *Atmosphere*, 10, 737, <https://doi.org/10.3390/atmos10120737>, 2019.

Kendall, M. G.: A new measure of rank correlation, *Biometrika*, 30, 81–93, 1938.

Kharin, V. V., Zwiers, F. W., Zhang, X., and Hegerl, G. C.: Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations, *Journal of Climate*, 20, 1419–1444, 2007.

Kharin, V. V., Zwiers, F. W., Zhang, X., and Wehner, M.: Changes in temperature and precipitation extremes in the CMIP5 ensemble, *Climatic change*, 119, 345–357, 2013.

Kim, H., Shin, J.-Y., Kim, T., Kim, S., and Heo, J.-H.: Regional frequency analysis of extreme precipitation based on a nonstationary population index flood method, *Advances in Water Resources*, 146, 103757, <https://doi.org/10.1016/j.advwatres.2020.103757>, 2020.

Koutsoyiannis, D. and Papalexiou, S. M.: Extreme rainfall: Global perspective, in: *Handbook of applied hydrology*, McGraw-Hill New York, NY, vol. 74, 1–74, ISBN 9780071835091, 2017.

Koutsoyiannis, D., Kozonis, D., and Manetas, A.: A mathematical framework for studying rainfall intensity-duration-frequency relationships, *Journal of hydrology*, 206, 118–135, 1998.

Lang, A. and Poschlod, B.: Updating catastrophe models to today's climate—An application of a large ensemble approach to extreme rainfall, *Climate Risk Management*, 44, 100594, <https://doi.org/10.1016/j.crm.2024.100594>, 2024.

Leduc, M., Mailhot, A., Frigon, A., Martel, J. L., Ludwig, R., Brietzke, G. B., Giguère, M., Brissette, F., Turcotte, R., Braun, M., and Scinocca, J.: The ClimEx project: A 50-member ensemble of climate change projections at 12-km resolution over Europe and northeastern North America with the Canadian Regional Climate Model (CRCM5), *Journal of Applied Meteorology and Climatology*, 58, 663–693, 2019.

Lehmann, E. A., Phatak, A., Stephenson, A., and Lau, R.: Spatial modelling framework for the characterisation of rainfall extremes at different durations and under climate change, *Environmetrics*, 27, 239–251, 2016.

Lenderink, G., de Vries, H., van Meijgaard, E., van der Wiel, K., and Selen, F.: A perfect model study on the reliability of the added small-scale information in regional climate change projections, *Climate Dynamics*, 60, 2563–2579, 2023.

Lewis, E., Fowler, H., Alexander, L., Dunn, R., McClean, F., Barbero, R., Guerreiro, S., Li, X.-F., and Blenkinsop, S.: GSDR: a global sub-daily rainfall dataset, *Journal of Climate*, 32, 4715–4729, 2019.

Lucas-Picher, P., Argüeso, D., Brisson, E., Tramblay, Y., Berg, P., Lemonsu, A., Kotlarski, S., and Caillaud, C.: Convection-permitting modeling with regional climate models: Latest developments and next steps, *Wiley Interdisciplinary Reviews: Climate Change*, 12, e731, <https://doi.org/10.1002/wcc.731>, 2021.

Mann, H. B.: Nonparametric tests against trend, *Econometrica: Journal of the econometric society*, 13, 245–259, <https://doi.org/10.2307/1907187>, 1945.

Marra, F., Nikolopoulos, E. I., Anagnostou, E. N., and Morin, E.: Metastatistical Extreme Value analysis of hourly rainfall from short records: Estimation of high quantiles and impact of measurement errors, *Advances in Water Resources*, 117, 27–39, <https://doi.org/10.1016/j.advwatres.2018.05.001>, 2018.

Martel, J.-L., Mailhot, A., Brissette, F., and Caya, D.: Role of natural climate variability in the detection of anthropogenic climate change signal for mean and extreme precipitation at local and regional scales, *Journal of Climate*, 31, 4241–4263, 2018.

Martel, J.-L., Mailhot, A., and Brissette, F.: Global and regional projected changes in 100-yr subdaily, daily, and multiday precipitation extremes estimated from three large ensembles of climate simulations, *Journal of Climate*, 33, 1089–1103, 2020.

Martel, J.-L., Brissette, F. P., Lucas-Picher, P., Troin, M., and Arseneault, R.: Climate change and rainfall intensity-duration-frequency curves: Overview of science and guidelines for adaptation, *Journal of Hydrologic Engineering*, 26, 03121001, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0002122](https://doi.org/10.1061/(ASCE)HE.1943-5584.0002122), 2021.

Martins, E. S. and Stedinger, J. R.: Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resources Research*, 36, 737–744, 2000.

Martins, E. S. and Stedinger, J. R.: Historical information in a generalized Maximum Likelihood Framework with partial duration and annual maximum series, *Water Resources Research*, 37, 2559–2567, <https://doi.org/10.1029/2000WR000009>, 2001.

Mattingly, K. S., Seymour, L., and Miller, P. W.: Estimates of Extreme Precipitation Frequency Derived from Spatially Dense Rain Gauge Observations: A Case Study of Two Urban Areas in the Colorado Front Range Region, *Annals of the American Association of Geographers*, 107, 1499–1518, 2017.

McElreath, R.: *Statistical Rethinking, A Course in R and Stan*, CRC Press, ISBN 9780367139919, 2015.

Meredith, E. P., Maraun, D., Semenov, V. A., and Park, W.: Evidence for added value of convection-permitting models for studying changes in extreme precipitation, *Journal of Geophysical Research: Atmospheres*, 120, 12500–12513, 2015.

Nadarajah, S., Anderson, C. W., and Tawn, J. A.: Ordered multivariate extremes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 473–496, 1998.

Naveau, P., Hannart, A., and Ribes, A.: Statistical methods for extreme event attribution in climate science, *Annual Review of Statistics and Its Application*, 7, 89–110, 2020.

Overeem, A., Buishand, A., and Holleman, I.: Rainfall depth-duration-frequency curves and their uncertainties, *Journal of Hydrology*, 348, 124–134, 2008.

Papalexiou, S. M. and Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily rainfall, *Water Resources Research*, 49, 187–201, 2013.

Papalexiou, S. M., AghaKouchak, A., and Foufoula-Georgiou, E.: A diagnostic framework for understanding climatology of tails of hourly precipitation extremes in the United States, *Water Resources Research*, 54, 6725–6738, 2018.

Poschlod, B.: Using high-resolution regional climate models to estimate return levels of daily extreme precipitation over Bavaria, *Nat. Hazards Earth Syst. Sci.*, 21, 3573–3598, <https://doi.org/10.5194/nhess-21-3573-2021>, 2021.

Poschlod, B.: Attributing heavy rainfall event in Berchtesgadener Land to recent climate change—Further rainfall intensification projected for the future, *Weather and Climate Extremes*, 38, 100492, <https://doi.org/10.1016/j.wace.2022.100492>, 2022.

Poschlod, B. and Daloz, A. S.: Snow depth in high-resolution regional climate model simulations over southern Germany – suitable for extremes and impact-related research?, *The Cryosphere*, 18, 1959–1981, <https://doi.org/10.5194/tc-18-1959-2024>, 2024.

Poschlod, B. and Koh, J.: Convection-permitting climate models can support observations to generate rainfall return levels, *Water Resources Research*, 60, e2023WR035159, <https://doi.org/10.1029/2023WR035159>, 2024.

Poschlod, B. and Ludwig, R.: Internal variability and temperature scaling of future sub-daily rainfall return levels over Europe, *Environmental Research Letters*, 16, 064097, <https://doi.org/10.1088/1748-9326/ac0849>, 2021.

Poschlod, B., Zscheischler, J., Sillmann, J., Wood, R. R., and Ludwig, R.: Climate change effects on hydrometeorological compound events over southern Norway, *Weather and climate extremes*, 28, 100253, <https://doi.org/10.1016/j.wace.2020.100253>, 2020.

Poschlod, B., Ludwig, R., and Sillmann, J.: Ten-year return levels of sub-daily extreme precipitation over Europe, *Earth Syst. Sci. Data*, 13, 983–1003, <https://doi.org/10.5194/essd-13-983-2021>, 2021.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F., Brisson, E., Kollet, S., Schmidli, J., van Lipzig, N. P. M., and Leung, R.: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges, *Reviews of Geophysics*, 53, 323–361, 2015.

Prescott, P. and Walden, A.: Maximum likelihood estimation of the parameters of the generalized extreme-value distribution, *Biometrika*, 67, 723–724, 1980.

Pumo, D., Francipane, A., Alongi, F., and Noto, L. V.: The potential of multilayer green roofs for stormwater management in urban area under semi-arid Mediterranean climate conditions, *Journal of Environmental Management*, 326, 116643, <https://doi.org/10.1016/j.jenvman.2022.116643>, 2023.

Ragulina, G. and Reitan, T.: Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and the Bayesian approach, *Hydrological Sciences Journal*, 62, 863–879, 2017.

Räty, O., Laine, M., Leijala, U., Särkkä, J., and Johansson, M. M.: Bayesian hierarchical modelling of sea-level extremes in the Finnish coastal region, *Nat. Hazards Earth Syst. Sci.*, 23, 2403–2418, <https://doi.org/10.5194/nhess-23-2403-2023>, 2023.

Röver, C. and Friede, T.: Dynamically borrowing strength from another study through shrinkage estimation, *Statistical Methods in Medical Research*, 29, 293–308, 2020.

Ruiz-Villanueva, V., Borga, M., Zoccatelli, D., Marchi, L., Gaume, E., and Ehret, U.: Extreme flood response to short-duration convective rainfall in South-West Germany, *Hydrol. Earth Syst. Sci.*, 16, 1543–1559, <https://doi.org/10.5194/hess-16-1543-2012>, 2012.

Saeb, A.: *gnFit: Goodness-of-Fit Tests Based on Chen-Balakrishnan*, r package version 0.2.0, CRAN [code], <https://doi.org/10.32614/CRAN.package.gnFit>, 2018.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G.: Akaike information criterion statistics, *J. Am. Stat. A.*, 83, <https://doi.org/10.1080/01621459.1988.10478680>, 1988.

Schwalm, C. R., Glendon, S., and Duffy, P. B.: RCP8. 5 tracks cumulative CO<sub>2</sub> emissions, *Proceedings of the National Academy of Sciences*, 117, 19656–19657, 2020.

Schwarzak, S., Haensel, S., and Matschullat, J.: Projected changes in extreme precipitation characteristics for Central Eastern Germany (21st century, model-based analysis), *International Journal of Climatology*, 35, 2724–2734, 2015.

Semenov, M. A.: Simulation of extreme weather events by a stochastic weather generator, *Climate Research*, 35, 203–212, 2008.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Luca, A. D., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S. M., Wehner, M., Zhou, B., and Allan, R.: Weather and climate extreme events in a changing climate, 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*, edited by: Pörtner, H.-O. Chan, J., Sorteberg, A., and Vera, C., Chap. 4, Cambridge University Press, Cambridge, 2022.

UK and New York, NY, USA, 571–766, ISBN 9781009325844, 2022.

Shehu, B., Willems, W., Stockel, H., Thiele, L.-B., and Haberlandt, U.: Regionalisation of rainfall depth–duration–frequency curves with different data types in Germany, *Hydrol. Earth Syst. Sci.*, 27, 1109–1132, <https://doi.org/10.5194/hess-27-1109-2023>, 2023.

Shin, H., Jung, Y., Jeong, C., and Heo, J.-H.: Assessment of modified Anderson–Darling test statistics for the generalized extreme value and generalized logistic distributions, *Stochastic Environmental Research and Risk Assessment*, 26, 105–114, 2012.

Sivia, D. and Skilling, J.: Data analysis: a Bayesian tutorial, OUP Oxford, ISBN 9780198568322, 2006.

Stan Development Team: RStan: the R interface to Stan, r package version 2.32.7, Stan [code], <https://mc-stan.org/> (last access: 1 February 2025), 2025.

Stedinger, J. R.: Frequency analysis of extreme events, *Handbook of hydrology*, ISBN 9780070397323, 1993.

Stein, M. L.: Some statistical issues in climate science, *Statist. Sci.*, 35, 31–41, <https://doi.org/10.1214/19-STS730>, 2020.

Steirou, E., Gerlitz, L., Apel, H., Sun, X., and Merz, B.: Climate influences on flood probabilities across Europe, *Hydrol. Earth Syst. Sci.*, 23, 1305–1322, <https://doi.org/10.5194/hess-23-1305-2019>, 2019.

Stephens, M. A.: EDF statistics for goodness of fit and some comparisons, *Journal of the American statistical Association*, 69, 730–737, 1974.

Tarasova, L., Merz, R., Kiss, A., Basso, S., Blöschl, G., Merz, B., Viglione, A., Plötner, S., Guse, B., Schumann, A., Fischer, S., Ahrens, B., Anwar, F., Bárdossy, A., Bühler, P., Haberlandt, U., Kreibich, H., Krug, A., Lun, D., Müller-Thomy, H., Pidoto, R., Primo, C., Seidel, J., Vorogushyn, S., and Wietzke, L.: Causative classification of river flood events, *Wiley Interdisciplinary Reviews: Water*, 6, e1353, <https://doi.org/10.1002/wat2.1353>, 2019.

Ullrich, S. L., Hegnauer, M., Nguyen, D. V., Merz, B., Kwadijk, J., and Vorogushyn, S.: Comparative evaluation of two types of stochastic weather generators for synthetic precipitation in the Rhine basin, *Journal of Hydrology*, 601, 126544, <https://doi.org/10.1016/j.jhydrol.2021.126544>, 2021.

Ulrich, J., Fauer, F. S., and Rust, H. W.: Modeling seasonal variations of extreme rainfall on different timescales in Germany, *Hydrol. Earth Syst. Sci.*, 25, 6133–6149, <https://doi.org/10.5194/hess-25-6133-2021>, 2021.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C.: Bayesian statistics and modelling, *Nature Reviews Methods Primers*, 1, 1, <https://doi.org/10.1038/s43586-020-00001-2>, 2021.

Van de Vyver, H.: Spatial regression models for extreme precipitation in Belgium, *Water Resources Research*, 48, <https://doi.org/10.1029/2011WR011707>, 2012.

van den Hurk, B., van Meijgaard, E., de Valk, P., van Heeringen, K.-J., and Gooijer, J.: Analysis of a compounding surge and precipitation event in the Netherlands, *Environmental Research Letters*, 10, 035001, <https://doi.org/10.1088/1748-9326/10/3/035001>, 2015.

Veenman, M., Stefan, A. M., and Haaf, J. M.: Bayesian hierarchical modeling: An introduction and reassessment, *Behavior Research Methods*, 56, 4600–4631, 2024.

Vijayaraghavan, K., Biswal, B. K., Adam, M. G., Soh, S. H., Tsen-Tieng, D. L., Davis, A. P., Chew, S. H., Tan, P. Y., Babovic, V., and Balasubramanian, R.: Bioretention systems for stormwater management: Recent advances and future prospects, *Journal of Environmental Management*, 292, 112766, <https://doi.org/10.1016/j.jenvman.2021.112766>, 2021.

Wilks, D. S. and Wilby, R. L.: The weather generation game: a review of stochastic weather models, *Progress in physical geography*, 23, 329–357, 1999.

Yang, L., Franzke, C. L., and Fu, Z.: Evaluation of the ability of regional climate models and a statistical model to represent the spatial characteristics of extreme precipitation, *International Journal of Climatology*, 40, 6612–6628, 2020.

Yiou, P. and Nogaj, M.: Extreme climatic events and weather regimes over the North Atlantic: when and where?, *Geophysical Research Letters*, 31, <https://doi.org/10.1029/2003GL019119>, 2004.

Zyphur, M. J. and Oswald, F. L.: Bayesian estimation and inference: A user's guide, *Journal of Management*, 41, 390–420, 2015.