



Regional pooling in extreme event attribution studies: an approach based on multiple statistical testing

Leandra Zanger¹ · Axel Bücher^{1,2} · Frank Kreienkamp³ · Philip Lorenz³ · Jordis S. Tradosky^{3,4}

Received: 31 December 2022 / Revised: 4 October 2023 / Accepted: 20 November 2023 /

Published online: 16 December 2023

© The Author(s) 2023

Abstract

Statistical methods are proposed to select homogeneous regions when analyzing spatial block maxima data, such as in extreme event attribution studies. Here, homogeneity refers to the fact that marginal model parameters are the same at different locations from the region. The methods are based on classical hypothesis testing using Wald-type test statistics, with critical values obtained from suitable parametric bootstrap procedures and corrected for multiplicity. A large-scale Monte Carlo simulation study finds that the methods are able to accurately identify homogeneous locations, and that pooling the selected locations improves the accuracy of subsequent statistical analyses. The approach is illustrated with a case study on precipitation extremes in Western Europe. The methods are implemented in an R package that allows for easy application in future extreme event attribution studies.

Keywords Extreme event attribution · Extreme value statistics · Homogeneity tests · Multiple comparison problem · Parametric bootstrap · Max-stable processes

1 Introduction

Extreme event attribution studies on precipitation extremes are typically motivated by the occurrence of an extreme event which causes major impacts such as damages to infrastructure and agriculture, or even fatalities, see, for instance, van der Wiel et al. (2017), van Oldenborgh et al. (2017), Otto et al. (2018) and

✉ Axel Bücher
axel.buecher@rub.de

¹ Heinrich-Heine Universität Düsseldorf, Mathematisches Institut, Düsseldorf, Germany

² Ruhr-Universität Bochum, Fakultät für Mathematik, Bochum, Germany

³ Deutscher Wetterdienst, Regionales Klimabüro Potsdam, Potsdam, Germany

⁴ Norwegian Meteorological Institute, Oslo, Norway

Kreienkamp et al. (2021). A key task for attributing the event to anthropogenic climate change consists of a statistical analysis of available observational data products at the location or region of interest (Philip et al. 2020). Typically, the observed time period is short, often less than 100 years, which ultimately leads to large statistical uncertainties. One possibility to reduce those uncertainties is to incorporate observations from nearby locations/regions, given that their meteorological characteristics are sufficiently similar and governed by the same underlying processes as those from the region affected by an extreme event. The selection of surrounding areas for which these criteria are met can be based on expert knowledge of the meteorological characteristics and dynamics, for instance provided by experts from the national meteorological and hydrological service of the affected country, like the Deutsche Wetterdienst in Germany. The expert knowledge-based suggestion may next be assessed statistically, which, to the best of our knowledge, has been done based on ad hoc methods in the past. In this paper, we propose profound statistical methods that can complement the expert's knowledge and which are based on statistically evaluating observational data from the past. Once regions with sufficiently similar characteristics of the analysed variable, e.g., the yearly maximum of daily rainfall, have been identified, the time series of all identified regions can be combined, thereby extending the available time series for the benefit of a more efficient statistical analysis.

The building blocks for the new approach are classical Wald-type tests statistics (Lehmann and Romano 2021) for testing the null hypothesis that the time-varying distribution functions at multiple locations of interest are the same. Spatial correlation is accounted for by using appropriate consistent estimates of the spatial estimation covariance matrix. Unlike in the classical text-book case, and motivated by the fact that standard likelihood-based inference for extreme value distributions requires unreasonably large sample sizes for sufficient finite-sample accuracy, we employ a parametric bootstrap device to approximate the distribution of the test statistics under the null hypothesis. This approach is motivated by results in Lilienthal et al. (2022) for respective stationary extreme value models. Based on suitable decompositions of a global null hypothesis, we then propose to test for carefully selected sub-hypotheses, possibly after correcting the individual tests' level for multiple comparisons. The results from the last-named paper are hence both generalized to non-stationary models and extended to a method that allows for pooling locations into a homogeneous region.

The new methods are illustrated by a large-scale Monte Carlo simulation study and by an application to the severe flooding event in Western Europe during July 2021 for which spatial pooling was applied in an attribution study following the event (Kreienkamp et al. 2021; Tradowsky et al. 2023). For the benefit of researchers who would like to use this spatial pooling approach, an implementation of the method in the statistical programming environment R (R Core Team 2022) is publicly available as an R package called `findpoolreg` on GitHub (Zanger 2022).

Attribution analysis of precipitation extremes is especially challenging due to short observational time series as well as their often limited spatial extent, which further complicates the detection of a trend and estimation of return periods based on the limited time series (see Tradowsky et al. 2023, for a discussion on this). Therefore, we will

in the following present the suggested approach for a heavy rainfall event, however, the method could equally be applied to other variables.

The remaining parts of this paper are organized as follows. Section 2 explains the mathematical concept of the proposed methods, starting with a detailed description of the underlying model assumptions and a strategy for the detection of a possible pooling region in Section 2.1. In Sections 2.2 and 2.3, mathematical details on the applied estimators and test statistics are given. Next, the ideas of the bootstrap procedures that allow us to draw samples from the distribution under the null hypothesis are explained (Section 2.4). Section 2.5 goes into detail about the detection strategy of possible pooling regions and the treatment of the related multiple testing problem. Next, Section 3 gives the results of the simulation study that was performed in order to evaluate the performance of the proposed methods. These results serve as a basis for the case study conducted in Section 4. Section 5 then discusses several extensions of the proposed methods: we provide a method for estimating *region-wise* return periods and extensions to different model assumptions that suit e.g. other variables such as temperature. Last but not least, we come to a conclusion in Section 6. Some mathematical details and further illustrations on the simulation study and the case study are postponed to a supplement.

2 Assessing spatial homogeneities for precipitation extremes

2.1 A homogeneous model for precipitation extremes

The observational data of interest consists of annual or seasonal maximal precipitation amounts (over some fixed time duration, e.g., a day) collected over various years and at various locations (in practice, each *location* may correspond to a spatial *region*; we separate these two terms from the outset to avoid misunderstandings: subsequently, a *region* shall be a set of locations). More precisely, we denote by $m_d^{(t)}$ the observed maximal precipitation amount in season t and at location d , with $t = 1, \dots, n$ and $d = 1, \dots, D$. The location of primary interest shall be the one with index $d = 1$. Note that the choice of $d = 1$ is made for illustrative purposes only and can be replaced by any index $d \in \{1, \dots, D\}$.

In view of the stochastic nature, we assume that $m_d^{(t)}$ is an observed value of some random variable $M_d^{(t)}$. Since $M_d^{(t)}$ is generated by a maxima operation, standard extreme value theory (Coles 2001) suggests to assume that $M_d^{(t)}$ follows the generalized extreme value (GEV) distribution, i.e.,

$$M_d^{(t)} \sim \text{GEV}(\mu_d(t), \sigma_d(t), \gamma_d(t))$$

for some $\mu_d(t), \sigma_d(t) > 0, \gamma_d(t) \in \mathbb{R}$, where the $\text{GEV}(\mu, \sigma, \gamma)$ distribution with location parameter $\mu > 0$, scale parameter $\sigma > 0$ and shape parameter $\gamma \in \mathbb{R}$ is defined by its cumulative distribution function

$$G_{(\mu, \sigma, \gamma)}(x) = \exp \left\{ - \left(1 + \gamma \frac{x - \mu}{\sigma} \right)^{-1/\gamma} \right\} \quad (1)$$

for x such that $1 + \gamma \frac{x-\mu}{\sigma} > 0$. Due to climate change, the temporal dynamics at location d , which are primarily driven by the function $t \mapsto (\mu_d(t), \sigma_d(t), \gamma_d(t))$, are typically non-constant. Any proxy for climate change qualifies as a suitable temporal covariate, and a standard assumption in extreme event attribution studies, motivated by the Clausius-Clapeyron relation, postulates that

$$\mu_d(t) = \mu_d \exp\left(\frac{\alpha_d \text{GMST}'(t)}{\mu_d}\right), \quad \sigma_d(t) = \sigma_d \exp\left(\frac{\alpha_d \text{GMST}'(t)}{\mu_d}\right), \quad \gamma_d(t) = \gamma_d \quad (2)$$

for certain parameters $\alpha_d, \gamma_d \in \mathbb{R}$, $\mu_d, \sigma_d > 0$. Here, $\text{GMST}'(t)$ denotes the smoothed global mean surface temperature anomaly, see Philip et al. (2020). Note that (2) implies

$$\text{GEV}(\mu_d(t), \sigma_d(t), \gamma_d(t)) = \exp\left(\frac{\alpha_d \text{GMST}'(t)}{\mu_d}\right) \text{GEV}(\mu_d, \sigma_d, \gamma_d),$$

hence the model may be identified as a temporal scaling model. It is further assumed that any temporal dependence at location d is completely due to $\text{GMST}'(t)$, which we treat as deterministic and which implies that $M_d^{(1)}, \dots, M_d^{(n)}$ are stochastically independent, for each $d = 1, \dots, D$. For the moment, the spatial dependence will be left unspecified.

Recall that the location of interest is the one with $d = 1$, which is characterised by the four parameters $\mu_1, \sigma_1, \gamma_1, \alpha_1$. As described before, estimating those parameters based on the observations from location $d = 1$ only may be unpleasantly inaccurate, which is why one commonly assumes that the D locations have been carefully selected by experts to meet the following space-time homogeneity assumption:

$$H_0^{\text{ED}} : \quad \exists \mathfrak{g} \in \Theta \forall d \in \{1, \dots, D\} : \quad \mathfrak{g}_d = \mathfrak{g}, \quad (3)$$

where $\Theta := (0, \infty)^2 \times \mathbb{R}^2$ and $\mathfrak{g} = (\mu, \sigma, \gamma, \alpha)^\top$, $\mathfrak{g}_d = (\mu_d, \sigma_d, \gamma_d, \alpha_d)^\top$, and where the upper index ED stands for ‘equal distribution’, since, in short, Eq. (3) states that the location-wise GEV parameters coincide for the D locations.

In the subsequent sections, we aim at testing the validity of the expert’s hypothesis H_0^{ED} . Here, it is not only of interest to test the hypothesis for the whole set $\{1, \dots, D\}$, but also to find a (maximal) subset $A \subset \{1, \dots, D\}$ with $1 \in A$ and $|A| = k \geq 2$ on which the space-time homogeneity assumption holds. Here, for an arbitrary index set A , the latter assumption may be expressed through

$$H_0^{\text{ED}}(A) : \quad \exists \mathfrak{g}_A \in \Theta \forall d \in A : \quad \mathfrak{g}_d = \mathfrak{g}_A, \quad (4)$$

with Θ as in Eq. (3) and $\mathfrak{g}_A = (\mu_A, \sigma_A, \gamma_A, \alpha_A)^\top$, meaning that the location-wise GEV parameters coincide for all locations with index in the set A , making the respective locations a possible pooling region.

Now, a maximal subset A for which Eq. (4) holds may be determined with the following strategy: Since we are interested in finding all locations that ‘match’ the location of primary interest with index $d = 1$, we test for each pair $A_d = \{1, d\}$, $d = 2, \dots, D$, whether the null hypothesis $H_0^{\text{ED}}(A_d)$ holds. This will

provide us with a set of p -values based on which we can decide which locations to reject and which not to reject. Those locations that are not rejected can then be assumed to be sufficiently homogeneous and are thus included in the suggestion of a pooling region of maximal extent. For further details on this strategy and the impact of the induced multiple testing problem, see Section 2.5.

2.2 Coordinate-wise maximum likelihood estimation

The starting point for the subsequent test statistics are the coordinate-wise maximum likelihood estimators for the model specified in (2). Writing $c^{(t)} = \text{GMST}'(t)$ for brevity, the log-likelihood contribution of observation $(M_d^{(t)}, c^{(t)})$ is given by $\ell_{\mathfrak{g}_d}(M_d^{(t)}, c^{(t)})$, where

$$\ell_{\mathfrak{g}_d}(x, c) = \log g_{(\mu_d \exp(\alpha_d c / \mu_d), \sigma_d \exp(\alpha_d c / \mu_d), \gamma_d)}(x) \tag{5}$$

with $g_{(\gamma, \mu, \sigma)}(x) = \frac{\partial}{\partial x} G_{(\mu, \sigma, \gamma)}(x)$ the probability density function of the GEV(μ, σ, γ)-distribution. The maximum likelihood estimator for \mathfrak{g}_d at location d is then defined as

$$\hat{\mathfrak{g}}_d \in \operatorname{argmax}_{\mathfrak{g}_d \in \Theta} \sum_{t=1}^n \ell_{\mathfrak{g}_d}(M_d^{(t)}, c^{(t)}) \tag{6}$$

The arg-maximum cannot be calculated explicitly, but may be found by suitable numerical optimization routines. We denote the gradient and the Hessian matrix of $\mathfrak{g} \mapsto \ell_{\mathfrak{g}}(x, c)$ by $\dot{\ell}_{\mathfrak{g}}(x, c) \in \mathbb{R}^4$ and $\ddot{\ell}_{\mathfrak{g}}(x, c) \in \mathbb{R}^{4 \times 4}$, respectively. Under appropriate regularity assumptions, standard asymptotic expansions (van der Vaart 1998, see also Bücher and Segers 2017 for the stationary GEV family) imply that $\hat{\theta} = (\hat{\mathfrak{g}}_1^\top, \dots, \hat{\mathfrak{g}}_D^\top)^\top \in \Theta^D$ is approximately Gaussian with mean $\theta = (\mathfrak{g}_1^\top, \dots, \mathfrak{g}_D^\top)^\top$ and covariance $n^{-1} \Sigma_n$, where $\Sigma_n = (\Sigma_{n;j,k})_{j,k=1}^D \in \mathbb{R}^{4D \times 4D}$ is defined as

$$\Sigma_{n;j,k} = J_{n;j,\mathfrak{g}_j}^{-1} \left(\frac{1}{n} \sum_{t=1}^n \operatorname{Cov} [\dot{\ell}_{\mathfrak{g}_j}(M_j^{(t)}, c^{(t)}), \dot{\ell}_{\mathfrak{g}_k}(M_k^{(t)}, c^{(t)})] \right) J_{n;k,\mathfrak{g}_k}^{-1} \in \mathbb{R}^{4 \times 4} \tag{7}$$

with $J_{n;j,\mathfrak{g}} = \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\ddot{\ell}_{\mathfrak{g}}(M_j^{(t)}, c^{(t)})] \in \mathbb{R}^{4 \times 4}$. We refer to Section A.1 in the supplement for details and Section A.2 for a suitable estimator $\hat{\Sigma}_n$ for Σ_n .

2.3 Wald-type test statistics

We define test statistics which allow to test for the sub-hypotheses $H_0^{\text{ED}}(A)$ of H_0^{ED} from Eq. (4), where $A \subset \{1, \dots, D\}$. For that purpose, we propose to use classical Wald-type test statistics; see Section 14.4.2 in Lehmann and Romano (2021) for a general discussion and Lilienthal et al. (2022) for a similar approach in temporally stationary GEV models, i.e., with α_d fixed to $\alpha_d = 0$.

Write $A = \{d_1, \dots, d_k\}$ with $1 \leq d_1 < \dots < d_k \leq D$ and let $h_A : \mathbb{R}^{4D} \rightarrow \mathbb{R}^{4(k-1)}$ be defined by

$$\begin{aligned} h_A(\boldsymbol{\theta}) &= h_A(\boldsymbol{\vartheta}_1, \dots, \boldsymbol{\vartheta}_D) = (\boldsymbol{\vartheta}_{d_1}^\top - \boldsymbol{\vartheta}_{d_2}^\top, \boldsymbol{\vartheta}_{d_2}^\top - \boldsymbol{\vartheta}_{d_3}^\top, \dots, \boldsymbol{\vartheta}_{d_{k-1}}^\top - \boldsymbol{\vartheta}_{d_k}^\top)^\top \\ &= (\mu_{d_1} - \mu_{d_2}, \sigma_{d_1} - \sigma_{d_2}, \gamma_{d_1} - \gamma_{d_2}, \alpha_{d_1} - \alpha_{d_2}, \\ &\quad \dots, \gamma_{d_{k-1}} - \gamma_{d_k}, \alpha_{d_{k-1}} - \alpha_{d_k})^\top. \end{aligned}$$

We may then write $H_0^{\text{ED}}(A)$ equivalently as

$$H_0^{\text{ED}}(A) : h_A(\boldsymbol{\theta}) = 0.$$

Hence, significant deviations of $h_A(\hat{\boldsymbol{\theta}})$ from 0 with $\hat{\boldsymbol{\theta}}$ from Section 2.2 provide evidence against $H_0^{\text{ED}}(A)$. Such deviations may be measured by the Wald-type test statistic

$$T_n(A) = n(h_A(\hat{\boldsymbol{\theta}}))^\top \left(\mathbf{H}_A \hat{\boldsymbol{\Sigma}}_n \mathbf{H}_A^\top \right)^{-1} h_A(\hat{\boldsymbol{\theta}}), \quad (8)$$

where $\mathbf{H}_A = \dot{h}_A(\boldsymbol{\theta}) \in \mathbb{R}^{4(k-1) \times 4D}$ denotes the Jacobian matrix of $\boldsymbol{\theta} \mapsto h_A(\boldsymbol{\theta})$, which is a matrix with entries in $\{-1, 0, 1\}$ that does not depend on $\boldsymbol{\theta}$. It is worthwhile to mention that the test statistic does not depend on the order of the coordinates, which is due to h_A being a linear function. In view of the asymptotic normality of $\hat{\boldsymbol{\theta}}$, see Section 2.2, the asymptotic distribution of $T_n(A)$ under the null hypothesis $H_0^{\text{ED}}(A)$ is the chi-square distribution $\chi_{4(k-1)}^2$ with $4(k-1)$ degrees of freedom; see also Section 4 in Lilienthal et al. (2022). Hence, rejecting $H_0^{\text{ED}}(A)$ if $T_n(A)$ exceeds the $(1-\alpha)$ -quantile of the $\chi_{4(k-1)}^2$ -distribution provides a statistical test of asymptotic level $\alpha \in (0, 1)$. The finite-sample performance of the related test in the stationary setting was found to be quite inaccurate (see Lilienthal et al. 2022). To overcome this issue, we propose a suitable bootstrap scheme in the next section.

2.4 Parametric bootstrap devices for deriving p -values

Throughout this section, we propose two bootstrap devices that allow to simulate approximate samples from the $H_0^{\text{ED}}(A)$ -distribution of the test statistic $T_n(A)$ from Eq. (8). Based on a suitably large set of such samples, one can compute a reliable p -value for testing $H_0^{\text{ED}}(A)$, even for short sample sizes.

The first method is based on a global fit of a max-stable process model to the entire region under consideration, while the second one is based on fitting multiple pairwise models. The main difference of the two approaches is that the first one can test the hypothesis $H_0^{\text{ED}}(A)$ for arbitrary subsets $A \subset \{1, \dots, D\}$, while the second approach is restricted to testing the null hypothesis on subsets of cardinality two, i.e., it can only test whether a pair of locations is homogeneous. Depending on the question that is asked, applying the one or the other method may be advantageous.

2.4.1 Global bootstrap based on max-stable process models

The subsequent bootstrap device is a modification of the parametric bootstrap procedure described in Section 5.3 of Lilienthal et al. (2022). Fix some large number B , say $B = 200$, noting that larger numbers are typically better, but going beyond $B = 1000$ is usually not worth the extra computational effort.

The basic idea is as follows: for each $b = 1, \dots, B$, simulate artificial *bootstrap samples*

$$\mathcal{D}_b^* = \{M_{d,b}^{(t),*} : t \in \{1, \dots, n\}, d \in \{1, \dots, D\}\}$$

that have a sufficiently similar spatial dependence structure as the observed data $\mathcal{D} = \{M_d^{(t)} : t \in \{1, \dots, T\}, d \in \{1, \dots, D\}\}$ and that satisfy the null hypothesis H_0^{ED} . For each fixed $A \subset \{1, \dots, D\}$ with $k = |A| \geq 2$, the test statistics computed on all bootstrap samples, say $(T_{n,b}^*(A))_{b=1, \dots, B}$, are then compared to the observed test statistic $T_n(A)$. Since the bootstrap samples do satisfy $H_0^{\text{ED}}(A)$, the observed test statistic $T_n(A)$ should differ significantly from the bootstrapped test statistics in case $H_0^{\text{ED}}(A)$ is not satisfied on the observed data.

Here, for simulating the bootstrap samples, we assume that the spatial dependence structure of the observed data can be sufficiently captured by a max-stable process model. Max-stable processes provide a natural choice here, since they are the only processes that can arise, after proper affine transformation, as the limit of maxima of independent and identically distributed random fields $\{Y_i(x) : x \in \mathbb{R}^p\}$ (Coles 2001, Section 9.3). Parametric models for max-stable processes are usually stated for unit Fréchet (i.e., GEV(1, 1, 1)) margins. Therefore, the first steps in our algorithm below aim at transforming the margins of our observed data to be approximately unit Fréchet.

More precisely, the parametric bootstrap algorithm is defined as follows:

Algorithm 1 (Bootstrap based on max-stable processes).

1. For each $d \in \{1, \dots, D\}$, calculate $\hat{\theta}_d$ from Section 2.2.
2. For each $d \in \{1, \dots, D\}$, transform the observations to approximately i.i.d. Fréchet-distributed data, by letting

$$Y_d^{(t)} = \left\{ 1 + \hat{\gamma}_d \frac{M_d^{(t)} - \hat{\mu}_d \exp\left(\frac{\hat{\alpha}_d \text{GMST}'(t)}{\hat{\mu}_d}\right)}{\hat{\sigma}_d \exp\left(\frac{\hat{\alpha}_d \text{GMST}'(t)}{\hat{\mu}_d}\right)} \right\}_+^{1/\gamma_d} \quad (t \in \{1, \dots, n\}). \tag{9}$$

3. Fit a set of candidate max-stable process models with standard Fréchet margins to the observations $(Y_1^{(t)}, \dots, Y_D^{(t)})_{t=1, \dots, n}$ and choose the best fit according to the composite likelihood information criterion (CLIC), which is a model selection criterion that is commonly applied when fitting max-stable process models. Throughout, we chose the following three models:

- (a) Smith's model (3 parameters);
- (b) Schlather's model with a powered exponential correlation function (3 parameters including a sill effect, see Section A.3 in the supplement for details);
- (c) the Brown-Resnick process (2 parameters).

For further details on max-stable processes, the mentioned models and the CLIC, see Sections A.3 and A.4 in the supplement, Davison et al. (2012) and Davison and Gholamrezaee (2012). Respective functions are implemented in the R package `SpatialExtremes` (Ribatet 2022).

4. For $b \in \{1, \dots, B\}$ and $t \in \{1, \dots, n\}$, simulate spatial data with unit Fréchet margins from the chosen max-stable process model, denoted by

$$(Y_{1,b}^{(t),*}, Y_{2,b}^{(t),*}, \dots, Y_{D,b}^{(t),*}).$$

Note that until now we haven't used the particular hypothesis $H_0^{\text{ED}}(A)$. Subsequently, fix $A = \{d_1, \dots, d_k\}$ with $1 \leq d_1 < \dots < d_k \leq D$.

5. Assume that $H_0^{\text{ED}}(A)$ from Eq. (4) is true, and estimate the four dimensional model parameters $\boldsymbol{\vartheta}_A = (\mu_A, \sigma_A, \gamma_A, \alpha_A)^\top \in \Theta$ by (pseudo) maximum likelihood based on the pooled sample

$$(M_{d_1}^{(1)}, c^{(1)}), \dots, (M_{d_1}^{(n)}, c^{(n)}), (M_{d_2}^{(1)}, c^{(1)}), \dots, (M_{d_2}^{(n)}, c^{(n)}), \dots, (M_{d_k}^{(1)}, c^{(1)}), \dots, (M_{d_k}^{(n)}, c^{(n)}).$$

Denote the resulting parameter vector as $\hat{\boldsymbol{\vartheta}}_A = (\hat{\mu}_A, \hat{\sigma}_A, \hat{\gamma}_A, \hat{\alpha}_A)^\top$, and note that $\hat{\boldsymbol{\vartheta}}_A$ should be close to $\boldsymbol{\vartheta}_d$ for each $d \in A$, if $H_0^{\text{ED}}(A)$ is met.

6. Transform the margins of the bootstrap samples to the ones of a GEV-model satisfying $H_0^{\text{ED}}(A)$, by letting

$$M_{d,b}^{(t),*} = \hat{\mu}_A \exp\left(\frac{\hat{\alpha}_A \text{GMST}'(t)}{\hat{\mu}_A}\right) + \hat{\sigma}_A \exp\left(\frac{\hat{\alpha}_A \text{GMST}'(t)}{\hat{\mu}_A}\right) \frac{(Y_{d,b}^{(t),*})^{\hat{\gamma}_A} - 1}{\hat{\gamma}_A} \quad (10)$$

for $t \in \{1, \dots, n\}$, $d \in A$ and $b \in \{1, \dots, B\}$. For each resulting bootstrap sample $\mathcal{D}_b^*(A) = \{M_{d,b}^{(t),*} : t \in \{1, \dots, n\}, d \in A\}$, compute the value $t_{n,b}^*(A)$ of the test statistic $T_n(A)$ from Eq. (8). Note that $T_n(A)$ only depends on the coordinates with $d \in A$.

7. Compute the value $t_n(A)$ of the test statistic $T_n(A)$ from Eq. (8) on the observed sample.
8. Compute the bootstrapped p -value by

$$p(A) = \frac{1}{B+1} \sum_{b=1}^B \mathbf{1}(t_n(A) \leq t_{n,b}^*(A)).$$

In a classical test situation, one may now reject $H_0^{\text{ED}}(A)$ for a fixed set A at significance level $\alpha \in (0, 1)$ if $p(A) \leq \alpha$. In the current pooling situation, we would

need to apply the test to multiple pooling regions A , which hence constitutes a multiple testing problem where standard approaches yield inflated levels. We discuss possible remedies in Section 2.5.

2.4.2 Pairwise bootstrap based on bivariate extreme value distributions

Recall that the location of primary interest is the one with index $d = 1$.

As stated in Section 2.1, it is of interest to test for all bivariate hypotheses $H_0^{\text{ED}}(\{1, d\})$ with $d = 2, \dots, D$. For that purpose, we may apply a modification of the bootstrap procedure from the previous section that makes use of bivariate extreme value models only. By doing so, we decrease the model risk implied by imposing a possibly restrictive global max stable process model.

The modification only affects step (3) and (4) from Algorithm 1. More precisely, for testing the hypothesis $H_0^{\text{ED}}(A_d)$ with $A_d = \{1, d\}$ for some fixed value $d = 2, \dots, D$, we make the following modifications:

Algorithm 2 (Pairwise bootstrap based on bivariate extreme value distributions). Perform step (1) and (2) from Algorithm 1 with the set $\{1, \dots, D\}$ replaced by A_d .

(3a) Fit a set of bivariate extreme value distributions to the bivariate sample $(Y_1^{(t)}, Y_d^{(t)})_{t=1, \dots, n}$, assuming the marginal distributions to be unit Fréchet. Choose the best fit according to the Akaike information criterion (AIC), a model selection criterion that rewards a good fit of a model and penalises the model's complexity at the same time (Akaike 1973). Possible models are:

- (a) the Hüsler-Reiss model (1 parameter);
- (b) the logistic model (1 parameter);
- (c) the asymmetric logistic model (3 parameters).

Note that all models are implemented in Stephenson (2002).

(4a) For $b \in \{1, \dots, B\}$ and $t \in \{1, \dots, n\}$, simulate bivariate data with unit Fréchet margins from the chosen bivariate extreme value model, denoted by $(Y_{1,b}^{(t),*}, Y_{d,b}^{(t),*})$.

Perform Steps (5)-(8) from Algorithm 1 with $A = A_d$.

Note that Algorithm 2 is computationally more expensive than Algorithm 1 since model selection and fitting of dependence models and its subsequent simulation must be performed separately for each hypothesis $H_0^{\text{ED}}(A_d)$ of interest.

2.5 Combining test statistics

As already addressed at the end of Section 2.1, it is not only of interest to test the global hypothesis H_0^{ED} , since a possible rejection of H_0^{ED} gives no indication about which locations deviate from the one of primary interest. Instead, one might want to

test hypotheses on several subsets and then pool those subsets for which no signal of heterogeneity was found. In this subsection, we provide the mathematical framework of testing sub-hypotheses and discuss how to deal with the induced multiple testing problem.

Mathematically, we propose to regard H_0^{ED} as a global hypothesis that is built up from elementary hypotheses of smaller dimension. A particularly useful decomposition is based on pairwise elementary hypotheses: recalling the notation $H_0^{\text{ED}}(A)$ from Eq. (4), we clearly have

$$H_0^{\text{ED}} = \bigcap_{d=2}^D H_0^{\text{ED}}(\{1, d\}), \quad (11)$$

i.e., H_0^{ED} holds globally when it holds locally for all pairs $\{1, d\}$ with $d \in \{2, \dots, D\}$. We may now either apply Algorithms 1 or 2 to obtain a p-value, say $p_d^{\text{raw}} = p(\{1, d\})$, for testing $H_0^{\text{ED}}(\{1, d\})$, for any $d \in \{2, \dots, D\}$. Each p-value may be interpreted as a signal for heterogeneity between locations 1 and d , with smaller values indicating stronger heterogeneity. The obtained raw list of p-values may hence be regarded as an exploratory tool for identifying possible heterogeneities.

Since we are now dealing with a multiple testing problem, it might be advisable to adjust for multiple comparison in order to control error rates. This can be done by interpreting the raw list based on classical statistical testing routines, in which p-values are compared with suitable critical values to declare a hypothesis significant. Several methods appear to be meaningful, and we discuss three of them in the following. For this, let $\alpha \in (0, 1)$ denote a significance level, e.g., $\alpha = 0.1$.

IM (Ignore multiplicity) reject homogeneity for all pairs $\{1, d\}$ for which $p_d^{\text{raw}} \leq \alpha$. In doing so, we do not have any control over false rejections. In particular, in case D is large, false rejections of some null hypotheses will be very likely. On the other hand, the procedure will have decent power properties, and will likely detect most alternatives. Hence, in a subsequent analysis based on the pooled sample of homogeneous locations, we can expect estimators to exhibit comparably little bias and large variance.

Holm (Control the family-wise error rate) apply Holm's stepdown procedure (Holm, 1979). For that purpose, sort the p-values $p_j = p_{1+j}^{\text{raw}} = p(\{1, 1+j\})$ with $j = 1, \dots, D-1$; denote them by $p_{(1)} \leq \dots \leq p_{(D-1)}$. Starting from $j = 1$, determine the smallest index j such that

$$p_{(j)} > \alpha_j := \alpha / (D - j).$$

If $j = 1$, then reject no hypotheses. If no such index exists, then reject all hypotheses. Otherwise, if $j \in \{2, \dots, D-1\}$, reject the hypotheses that belong to the p-values $p_{(1)}, \dots, p_{(j-1)}$.

The procedure can be equivalently expressed by adjusted p-values. Recursively defining $\tilde{p}_{(1)} = \min\{1, (D-1)p_{(1)}\}$ and

$$\tilde{p}_{(j)} = \min \{ 1, \max \{ \tilde{p}_{(j-1)}, (D - j)p_{(j)} \} \}$$

for $j = 2, \dots, D - 1$, we simply reject those hypotheses that belong to the adjusted p -values with $\tilde{p}_{(j)} \leq \alpha$.

Holm’s stepdown procedure is known to asymptotically control the family-wise error rate (FWER) at level α , i.e.,

$$\text{FWER} := \Pr (\text{reject any true null hypothesis } H_0^{\text{ED}}(\{1, d\})) \leq \alpha,$$

see Theorem 9.1.2 in Lehmann and Romano (2021).

In general, controlling the family-wise error rate will result in comparably little power, i.e., we might falsely identify some pairs of locations as homogeneous. Hence, in a subsequent analysis based on the pooled sample of homogeneous locations, we can expect estimators to exhibit comparably large bias and little variance.

BH (Control the false discovery rate) apply the Benjamini Hochberg stepup procedure (Benjamini and Hochberg 1995). For that purpose, sort the p -values $p_j = p_{1+j}^{\text{raw}} = p(\{1, 1 + j\})$ with $j = 1, \dots, D - 1$; denote them by $p_{(1)} \leq \dots \leq p_{(D-1)}$. Starting from $j = D - 1$, determine the largest index j such that

$$p_{(j)} \leq \alpha_j := \frac{j\alpha}{(D - 1)}.$$

If no such index exists, then reject no hypotheses. Otherwise, if $j \in \{1, \dots, D - 1\}$, reject the hypotheses that belong to the p -values $p_{(1)}, \dots, p_{(j)}$.

Again, one can compute adjusted p -values $\tilde{p}_{(j)}$ such that the procedure is equivalent to rejecting those hypotheses for which $\tilde{p}_{(j)} \leq \alpha$. For that purpose, let $\tilde{p}_{(D-1)} = \min\{1, (D - 1)p_{(D-1)}\}$ and recursively define, for $j = D - 2, \dots, 1$,

$$\tilde{p}_{(j)} = \min \left\{ 1, \min \left\{ (D - 1) \frac{p_{(j)}}{j}, \tilde{p}_{(j+1)} \right\} \right\}.$$

Under an additional assumption on the p -values that belong to the true null hypotheses (they must exhibit some positive dependence), the BH procedure is known to asymptotically control the false discovery rate (FDR) at level α , i.e.,

$$\text{FDR} := \mathbb{E} \left[\frac{\text{Number of false rejections}}{\text{Number of all rejections}} \mathbf{1}(\text{at least one rejections}) \right] \leq \alpha,$$

see Theorem 9.3.3 in Lehmann and Romano (2021). Control of the FDR will be confirmed by the simulation experiments in Section 3.

If one were interested in guaranteed theoretical control of the FDR rate, one might alternatively apply the Benjamini Yekutieli (BY) stepup procedure, see (Benjamini and Yekutieli 2001) and Theorem 9.3.3 in Lehmann and Romano (2021). In view of the fact that the procedure is much more conservative than BH, we do not recommend its application in the current setting.

Concerning a subsequent analysis, estimators based on a pooled sample obtained from the BH procedure can be expected to exhibit bias and variance to be somewhere between the IM and Holm procedure.

Remark 1 The decomposition of H_0^{ED} into hypotheses of smaller dimensionality is not unique. For instance, we may alternatively write

$$H_0^{\text{ED}} = \bigcap_{k=1}^K H_0^{\text{ED}}(B_k), \quad (12)$$

where $\{1\} \subset B_1 \subset B_2 \cdots \subset B_K = \{1, \dots, d\}$ denotes an increasing sequence of regions with $2 \leq |B_1| < |B_2| < \cdots < |B_K| = d$ (for instance, $B_k = \{1, 2, \dots, 1+k\}$ with $k = 1, \dots, D-1$). In practice, the sequence is supposed to be derived from some expert knowledge of the region of interest; it shall represent a sequence of possible pooling regions where B_k is constructed from B_{k-1} by adding the locations which are a priori ‘most likely’ homogeneous to the locations in B_k . Note that, necessarily, $K \leq D-1$, which provides an upper bound on the number of hypotheses to be tested.

The derivation of respective testing methods is straightforward. In view of the facts that the choice of the sequence is fairly subjective and that the eventual results crucially depend on that choice, we do not pursue the method any further.

3 Simulation study

A large-scale Monte Carlo simulation study was conducted to assess the performance of the proposed bootstrap procedures in finite sample situations. We aim at answering the following questions:

- (a) Regarding the test’s *power*: What percentage of locations that are heterogeneous w.r.t. the location of primary interest can be expected to be identified correctly?
- (b) Regarding the test’s *error rates*: What percentage of locations that are homogeneous w.r.t. the location of primary interest can be expected to be wrongly identified as heterogeneous (FDR)? What is the probability of wrongly identifying at least one location that is homogeneous w.r.t. the location of interest as heterogeneous (FWER)?
- (c) Regarding the chosen pooling regions: How does return level (RL) estimation based on the pooling regions proposed by the bootstrap procedures compare to RL estimation based on the location of interest only or the whole (heterogeneous) region?

The data was generated in such a way that the temporal spatial dynamics from the case study in Section 4 are mimicked. To achieve this, we started by fitting the scale-GEV model from Eq. (2) to annual block-maxima of observations from 1950–2021 at 16 spatial locations in Western Europe (i.e., $n = 72$ and $D = 16$) that

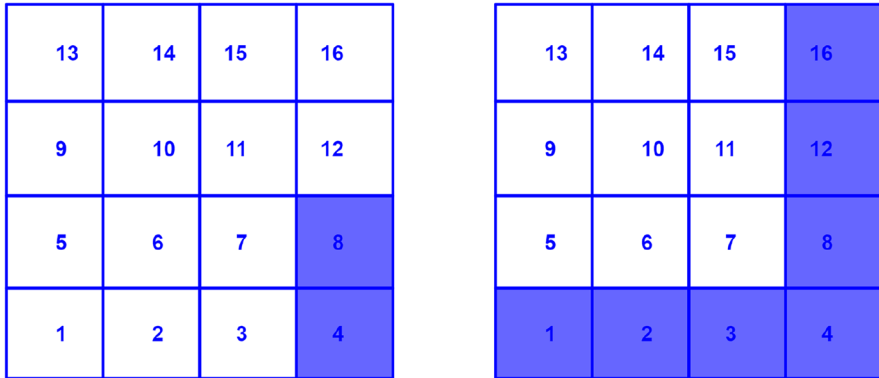


Fig. 1 Illustration of the grid used for the simulation. The regions contained in A_{dev} are shaded in blue, with $|A_{dev}| = 2$ shown in the left plot and $|A_{dev}| = 7$ shown on the right. The region of interest is the one labelled 10

are arranged in a 4×4 grid; see Fig. 1 and the additional explanations in Section 4. The locations correspond to the center points of the grid cells; the distance between the center points of two neighbouring grid cells is approximately 140 km. The location-wise GEV parameter estimates $\hat{\theta}_d$ exhibit the following approximate ranges over $d \in \{1, \dots, 16\}$: $\hat{\mu}_d \in (18.1, 30.8)$ with a mean of 20.85, $\hat{\sigma}_d \in (4.185, 7.92)$ with a mean of 5.3, $\hat{\gamma}_d \in (-0.13, 0.36)$ with a mean of 0.08 and $\hat{\alpha}_d \in (-2.3, 5.08)$ with a mean of 1.5. Fitting the scale-GEV model to the full pooled sample of size $n \cdot D = 1152$, we obtained parameter estimates that were close to the means over the location-wise parameter estimates, with 20.37, 5.8, 0.1, 1.5 for location, scale, shape and trend parameter, respectively. Next, we transformed the margins to (approximate) unit Fréchet by applying the transformation from Eq. (9), such that we can fit several max-stable process models to the transformed data. The best fit was Smith’s model with approximate dependence parameters $\sigma_{11} = 0.4, \sigma_{12} = 0.2, \sigma_{22} = 0.9$; see Davison et al. (2012) for details on the model.

Based on these model fits, we chose to generate data with the following specifications: first, the sample size was fixed to $n = 75$ and the regional 4×4 grid was chosen as described before, i.e., $d = 16$. The grid cell/location labelled ‘10’ is chosen as the one of primary interest. Further, the dependence structure is fixed to that of Smith’s model with (approximately) those parameters that gave the best fit on the observed data, i.e. $\sigma_{11} = 0.4, \sigma_{12} = 0.2, \sigma_{22} = 0.9$ (results based on samples from an anisotropic version of the Brown-Resnick model can be found in Section B.3 in the supplement). For simulating data, we first simulate from this max-stable process model (Ribatet 2022) and then transform the margins to scale-GEV distributions, either in a homogeneous or in a heterogeneous manner. Here, the globally homogeneous model is defined by fixing the marginal scale-GEV parameters to approximately the mean values of the location-wise GEV parameters obtained for the real observations, i.e.,

$$\mu_d = 20, \sigma_d = 5.5, \gamma_d = 0.1, \alpha_d = 1.5 \tag{13}$$

for each $d \in \{1, \dots, 16\}$.

Starting from this homogeneous model, we consider two different heterogeneous scenarios. In the first scenario, we fix $\boldsymbol{\theta}_d = (\mu_d, \sigma_d, \gamma_d, \alpha_d)^\top$ as in Eq. (13) for all $d \in A_{\text{hom}} = \{1, \dots, 16\} \setminus \{4, 8\}$, while

$$\begin{aligned} \mu_d &= 20 + c_\mu, & c_\mu &\in \{-3, -1.5, 0, 1.5, 3\}, \\ \sigma_d &= 5.5 \cdot c_\sigma, & c_\sigma &\in \{0.7, 0.85, 1, 1.15, 1.3\}, \\ \gamma_d &= 0.1 + c_\gamma, & c_\gamma &\in \{-0.1, 0, 0.1\}, \\ \alpha_d &= 1.5 + c_\alpha, & c_\alpha &\in \{-1, 0, 1\}, \end{aligned} \quad (14)$$

for $d \in A_{\text{dev}} = \{4, 8\}$ with $(c_\mu, c_\sigma, c_\gamma, c_\alpha) \neq (0, 0, 0, 0)$. Note that this defines $5 \cdot 5 \cdot 3 \cdot 3 - 1 = 224$ different heterogeneous models. In the second scenario, we consider the same construction with $A_{\text{hom}} = \{5, 6, 7, 9, 10, 11, 13, 14, 15\}$ and $A_{\text{dev}} = \{1, 2, 3, 4, 8, 12, 16\}$. An illustration of the grid cells and their partition into homogeneous and non-homogeneous areas can be found in Fig. 1. Overall, we obtain 448 different heterogeneous models and one homogeneous model.

For each of the 449 models, we now apply the following three bootstrap procedures, each carried out with $B = 300$ bootstrap replications (recall that the grid cell of interest is the one labelled with 10):

- (B1) The bootstrap procedure from Algorithm 1 with $A = \{1, \dots, 16\}$.
- (B2) The bootstrap procedure from Algorithm 1 for all sets $A_d = \{10, d\}$ with $d \in \{1, \dots, 16\} \setminus \{10\}$.
- (B3) The bootstrap procedure from Algorithm 2 for all sets $A_d = \{10, d\}$ with $d \in \{1, \dots, 16\} \setminus \{10\}$.

Note that the second and third method both yield 15 raw p -values. Each procedure was applied to 500 simulated samples from all models under consideration.

Regarding (B1), we compute the percentage of rejections among the 500 replications, which represents the empirical type I error of the test under the homogeneous model and the empirical power under the heterogeneous models. The results can be found in Fig. 2. The null hypothesis is met in the central square only, and we observe that the nominal level of $\alpha = 0.1$ is perfectly matched. All non-central squares correspond to different alternatives, and we observe decent power properties in both scenarios. Note that a rejection only implies that the entire region $\{1, \dots, 16\}$ is not homogeneous; there is no information on possible smaller subgroups that are homogeneous to the location of interest.

Regarding (B2) and (B3), rejection decisions were obtained for each hypothesis $H_0^{\text{ED}}(\{10, d\})$ by one of the three methods from Section 2.5. The empirical family-wise error rate is then the percentage of cases (over 500 replications) for which at least one null hypothesis was rejected. Likewise, for the false discovery rate, we calculate, for each replication, the number of false rejections and divide that by the total number of rejections (when the number of total rejections is 0, this ratio is set to 0). The empirical false discovery rate is obtained by taking the mean over all 500 replications. Similarly, for assessing the power properties, we calculate the empirical

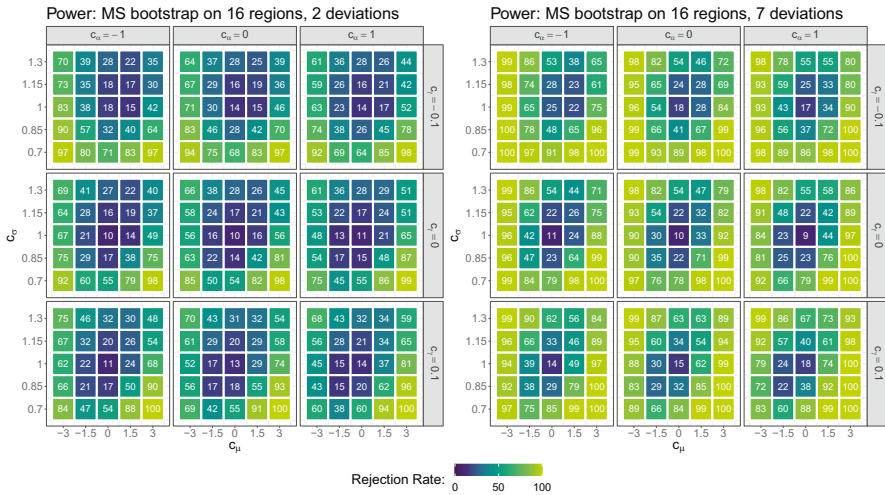


Fig. 2 Rejection rates in % obtained for (B1), in the setting where either 2 (left plot) or 7 (right plot) regions deviate from the others. Each coloured square contains the rejection rate for one of the 225 different models, with the central square with $c_\mu = c_\sigma = 0$ corresponding to the null hypothesis. The x - and y -axis and the facets determine the values of the scale-GEV parameter vector of the deviating locations through Eq. (14)

proportion of correct rejections (i.e., among the 2 or 7 locations that deviate, the proportion of detected heterogeneous locations) over all 500 replications.

Results for the false discovery and family-wise error rate are given in Table 1. We find that the p -value combination methods from Section 2.5 are sufficiently accurate: the BH method controls the false discovery rate, while Holm’s method controls the family-wise error rate. This holds exactly for procedures (B3), where the maximal FDR (FWER) of the BH (Holm) method is at 9.4% (8.7%), and approximately for (B2), where the maximal FDR (FWER) is at 12.2% (12.6%). Further, we see that the IM procedure neither controls the FWER nor the FDR.

Table 1 False Discovery Rate (FDR) and family-wise Error Rate (FWER) for the three p -value combination methods from Section 2.5 and the two bootstrap methods (B2) and (B3). The stated values are the minimum, maximum and mean across the 224 alternative models from each scenario

Method	min FDR		max FDR		mean FDR		min FWER		max FWER		mean FWER	
	(B2)	(B3)	(B2)	(B3)	(B2)	(B3)	(B2)	(B3)	(B2)	(B3)	(B2)	(B3)
<i>Scenario 1: $A_{dev} = 2$</i>												
BH	7.3	5.6	12.2	9.4	9.4	7.5	9.1	6.9	21.2	19.0	14.4	11.8
Holm	3.0	2.3	11.7	8.3	7.1	5.1	6.9	5.0	12.6	8.7	9.5	6.6
IM	25.8	25.3	61.7	60.1	37.7	37.2	53.4	53.4	64.5	62.8	59.1	58.3
<i>Scenario 2: $A_{dev} = 7$</i>												
BH	3.6	2.4	12.1	8.9	5.6	4.9	6.2	4.2	32.0	29.8	18.6	16.8
Holm	1.0	0.9	11.3	7.9	3.4	2.6	3.8	2.4	11.3	7.9	7.1	5.1
IM	7.9	7.8	61.5	60.1	16.0	15.8	40.9	40.9	61.5	60.1	47.3	46.4

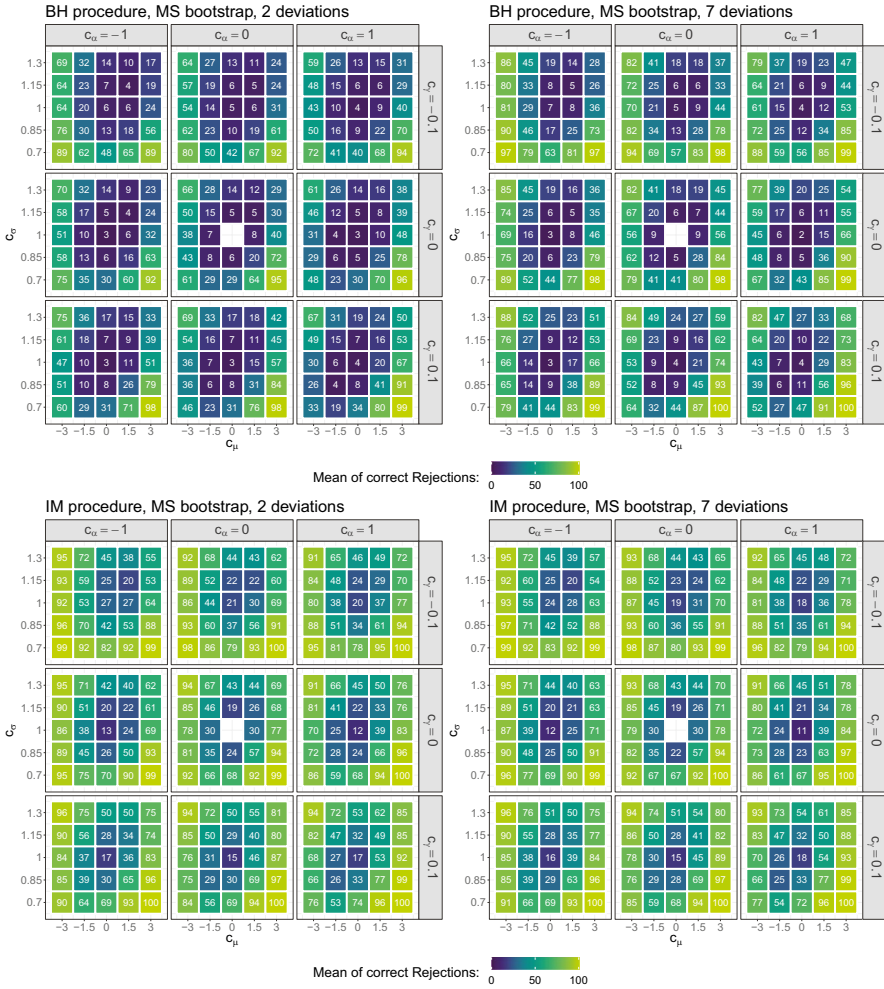


Fig. 3 Proportion of correct rejections in % obtained with the BH procedure (upper row) and the IM procedure (lower row) at a level of 0.1, in the setting where two stations deviate from the rest (left column) or 7 locations deviate from the rest (right column), with the bootstrap procedure based on max-stable processes. The axis and facets are as described in Fig. 2

The power properties for procedure (B2) combined with the BH method are shown in Fig. 3. We see that the procedure is able to detect some of the deviations of the null hypothesis, with more correct rejections the stronger the deviation is. The method is particularly powerful when the location and scale parameters deviate into opposite directions, i.e. when $c_\mu > 0$ and $c_\sigma < 1$ or $c_\mu < 0$ and $c_\sigma > 1$. There is no obvious pattern regarding the deviations of the shape and trend parameter. Further, we analogously show the power properties of the IM method with bootstrap (B2) in Fig. 3. As expected, this method has more power against all alternatives under

consideration. However, this comes at the cost of more false discoveries, as can be seen in Table 1.

The results for bootstrap scheme (B3) were very similar and are therefore not shown here, but can be found in Section B of the supplementary material. Likewise, we omit the results for the more conservative Holm procedure, which exhibits, as expected, less power against all alternatives. Further, we repeated the simulation study with an increased location-wise sample size of $n = 100$. As one would expect, the tests have more power in this case. Finally, additional results under misspecification of the max-stable dependence models in the bootstrap are presented in Section B.3 in the supplement; they reveal that the methods are rather robust with respect to misspecification.

The results presented so far show that the proposed pooling methods work ‘as intended’, since the theoretical test characteristics are well approximated in finite sample situations, and since we observe decent power properties. In practical applications however, spatial pooling of locations is usually the starting point for subsequent analyses. For instance, one may be interested in estimating return levels at the location of interest based on the data from all locations that were identified as homogeneous. Moreover, the analysis of alternative data sets like climate model data may be based on the homogeneous locations identified within the analysis of observations.

This suggests that the methods should be examined with regard to their quality in subsequent analyses. For that purpose, we consider, as an example, the problem of return level estimation at the location of interest. The state-of-the-art method would consist of GEV fitting at the location of interest only, which results in (asymptotically) unbiased estimators that suffer from large variance. Basing the estimator on pooled regions will decrease the variance, but at the same time increase its bias if some heterogeneous locations have been wrongly identified as homogeneous.

In particular, pooling based on a conservative testing approach like the BH procedure leads to the acceptance of many locations and thus to a large pooling area and low estimation variance. Most likely, some of the chosen locations will be violating the null hypothesis though, which yields a rather large estimation bias. For pooling based on a more liberal rejection approach like the IM procedure, the estimation bias and variance behave exactly opposite: since the null hypotheses are more likely to be rejected, the resulting pooling sample is smaller (i.e., larger estimation variance) but ‘more accurate’ (i.e., smaller estimation bias).

For our comparison, we consider fitting the scale-GEV model based on pooled locations that have been obtained from one of the following eight methods

$$m \in \{\text{LOI, full, MS IM, MS Holm, MS BH, biv. IM, biv. Holm, biv. BH}\}.$$

Here, LOI refers to considering the location of interest only (no pooling), full refers to pooling all available locations, and the last six methods encode pooling based on any combination of the proposed p-value combination methods and bootstrap approaches.

For each method, we compute the maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\gamma}, \hat{\alpha})^T$ of the scale-GEV model parameters and transform this to an estimate of the T -year return level (RL) in the reference climate of year t by

$$\widehat{\text{RL}}_t(T) = G_{(\hat{\mu}(t), \hat{\sigma}(t), \hat{\gamma})}^{-1}(1 - 1/T),$$

where $\hat{\mu}(t) = \hat{\mu} \exp(\hat{\alpha} \text{GMST}'(t)/\hat{\mu})$ and $\hat{\sigma}(t) = \hat{\sigma} \exp(\hat{\alpha} \text{GMST}'(t)/\hat{\mu})$ and where G is the cumulative distribution function of the GEV-distribution, see Eq. (1). Now, in our simulation study, we know that the true value of the target RL is given by $\text{RL}_t(T) = G_{(\mu_0(t), \sigma_0(t), \gamma_0)}^{-1}(1 - 1/T)$ with

$$\mu_0(t) = 20 \exp\left(\frac{1.5 \text{GMST}'(t)}{20}\right), \quad \sigma_0(t) = 5.5 \exp\left(\frac{1.5 \text{GMST}'(t)}{20}\right), \quad \gamma_0 = 0.1.$$

From the 500 replications we can therefore compute the empirical Mean Squared Error (MSE) of method m as

$$\text{MSE}(m) = \frac{1}{500} \sum_{j=1}^{500} \left(\widehat{\text{RL}}_t^{(m,j)}(T) - \text{RL}_t(T) \right)^2,$$

where $\widehat{\text{RL}}_t^{(m,j)}(T)$ denotes the estimated RL obtained in the j -th replication with method m . Note that we have suppressed the MSE's dependence on T and t from the notation.

In Fig. 4 we compare MSEs of the 100-year RL with reference climate as in year 2021, which is given by $\text{RL}_{2021}(100) = 55.87$, by plotting the difference $\text{MSE}(m_1) - \text{MSE}(m_2)$ with $m_1 \in \{\text{MS BH}, \text{MS IM}\}$ and $m_2 \in \{\text{full}, \text{ROI}\}$ as obtained for the setting where $|A_{\text{dev}}| = 7$. The plots reveal that both the MS BH and the MS IM method are superior to the LOI fit for almost all scenarios. Comparing the two methods to the full fit reveals that there are certain scenarios for which the full fit performs substantially worse, mostly when the shape and scale parameter deviate towards the same direction for the alternatives. For those scenarios where the full fit outperforms the two methods, the discrepancy is not very large, with the BH method performing slightly better than the IM method.

A comparison between MS BH and MS IM is shown in Fig. 5 for $|A_{\text{dev}}| \in \{2, 7\}$. The results reveal that the BH method slightly outperforms the IM method in the case $|A_{\text{dev}}| = 2$ for almost all alternative scenarios. In case $|A_{\text{dev}}| = 7$, the results are quite mixed, with the IM method becoming clearly superior when the shape, scale and location parameters deviate jointly to the top. In all other scenarios, the differences are only moderate, sometimes favoring one method and sometimes the other. Corresponding results for the bootstrap methods based on bivariate extreme value distributions are very similar and therefore not shown. Further, the results were found to be robust against the choices of $t = 2021$ and $T = 100$ that were made here for the return level estimation.

Overall, the results suggest the following practical recommendation: if the full sample is expected to be quite homogeneous a priori (for instance, because it was built based on expert knowledge), then estimation based on BH-based pooling is preferable over the other options (LOI, the full and the IM-based fit). If one expects to have a larger number of heterogeneous locations, it is advisable to apply the IM procedure (or any other liberal procedure), which likely rejects most of the heterogeneous locations and hence reduces the bias. In general, the

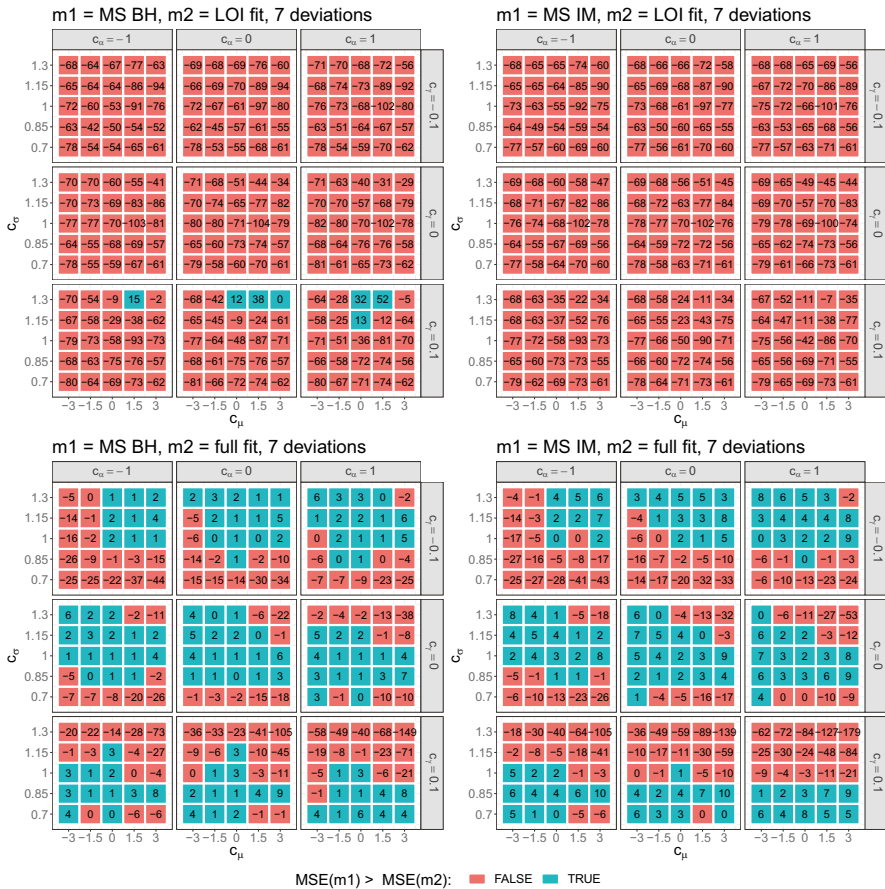


Fig. 4 Comparison of MSEs of $RL_{2021}(100)$ obtained for different choices of the method m , in the setting where $|A_{dev}| = 7$. Shown are the differences $MSE(m_1) - MSE(m_2)$ with m_1 and m_2 as indicated in the plot title. Negative values (red) therefore indicate a lower MSE for the method mentioned first, and vice versa for positive values. The axis and facets are as described in Fig. 2

liberal behavior of IM-based pooling suggests its use when it is of highest practical interest to obtain a pooled region that is as homogeneous as possible (as a trade-off, one has to accept that the region is probably much smaller than the initial full region).

4 Severe flooding in Western Europe during July 2021 revisited

We illustrate the new pooling methods in a case study by revisiting the extreme event attribution study for the heavy precipitation event that led to severe flooding in Western Europe during July 2021, see Kreienkamp et al. (2021) and Tradowsky et al. (2023). In that study, observational data were pooled together based on expert

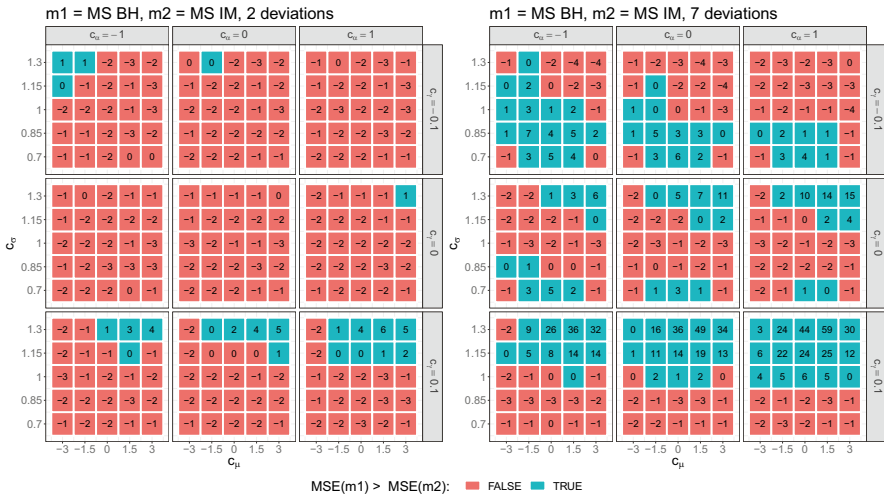


Fig. 5 Comparison of MSEs of $RL_{2021}(100)$ in the setting where $|A_{dev}| = 2$ (left) and $|A_{dev}| = 7$ (right). Shown are the differences $MSE(MS\ BH) - MSE(MS\ IM)$. Negative values (red) therefore indicate a lower MSE for the BH method, while positive values (blue) indicate a lower MSE for the IM method. The axis and facets are as described in Fig. 2

knowledge and on ad hoc tests, with the ultimate goal of assessing the influence of human-made climate change on the likelihood and severity of similar events in Western and Central Europe.

The full region under investigation in Kreienkamp et al. (2021) and Tradowsky et al. (2023) consists of sixteen ($2.0^\circ \times 1.25^\circ$) (i.e. about $140\text{ km} \times 140\text{ km}$) grid cells reaching from the northern Alps to the Netherlands, see Fig. 5 in

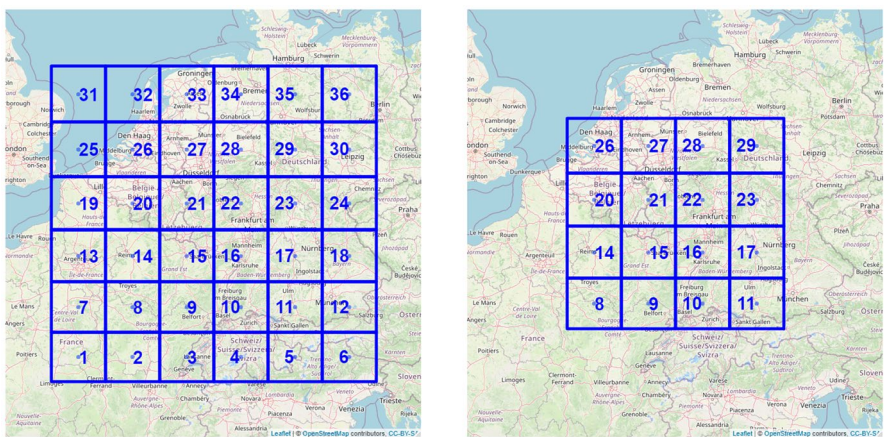


Fig. 6 Regions analysed within this case study and the respective numbering used here. The data consists of April-September block maxima of tile-wise averaged daily precipitation sums (RX1day) from 1950–2021

Kreienkamp et al. (2021) or the right-hand side of Fig. 6. Two of the 16 locations were rejected in that study due to expert knowledge and too large deviations in fitted GEV-parameters (grid cells 17 and 11 of Fig. 6). Among other things, our illustrative application of the methods explained above will reveal that grid cell 11 has been rightfully dismissed, while grid cell 17 might have been considered homogeneous. Further, there is no clear evidence that any other grid cell that has been declared homogeneous should rather be considered non-homogeneous.

For illustrative purposes, we apply our methods to two different initial areas:

- (A) An area consisting of 6×6 grid cells covering a large part of Western/ Central Europe, as shown in Fig. 6 on the left.
- (B) The original 4×4 grid cells from Kreienkamp et al. (2021) as shown in Fig. 6 on the right.

Note that homogeneity for the 20 grid cells at the boundary of the larger area in (A) has been dismissed based on expert knowledge in Kreienkamp et al. (2021); the larger area is included here for illustrative purposes only.

The data used throughout the study consists of April–September block-maxima of tile-wise averaged 1-day accumulated precipitation amounts of the E-OBS data set (Cornes et al. (2018), Version 23.1e). In both cases, the grid cell with label 21 is the one of primary interest, since it is the one containing the target location of the study, i.e., the region that accumulated the highest precipitation sum and experienced the largest impacts during the flooding of July 2021. The time series are shown in Fig. C.1 in the supplementary material. There, we also plot values of $\hat{\mu}(t) = \hat{\mu} \exp(\hat{\alpha} \text{GMST}'(t)/\hat{\mu})$ obtained from different data sets: once from data of location 21 only, once from data of the respective location only, and once from the pooled data of the respective pair $(21, d)$ for $d \in \{1, \dots, 36\} \setminus \{21\}$.

We apply the two proposed bootstrap procedures to areas (A) and (B). Note that the raw p -values obtained with the bootstrap based on bivariate extreme value distributions should be very similar (or even identical when using the same seed for random number generation) for those grid cells that appear in both areas, while they may differ to a greater extent for the MS bootstrap. This is because the p -value for a given pair obtained with the bivariate bootstrap procedure only depends on the observations of the pair, while the respective p -value obtained with the MS bootstrap also depends on the spatial model that was fitted to the whole area. However, even if the raw p -value of a given pair obtained for setting (B) coincides with the raw p -value obtained for setting (A), the adjustment for multiple testing can lead to slightly different rejection decisions of the pair at a given level α . The bootstrap procedures are applied with $B = 2000$ bootstrap replications.

We start by discussing the results of the application to the larger grid in (A). Recall that, for a given significance level α , one rejects the null hypothesis for all grid cells whose p -value is smaller than α . To visualise the results, we therefore shade the grid cells according to the magnitude of their (adjusted) p -value. Here, we divide the colour scale into three groups: $[0, 0.05]$, $(0.05, 0.1]$ and $(0.1, 1]$, with a dark red tone assigned to the first group, a brighter red tone for Group

2 and an almost transparent shade for Group 3. This allows us to see the test decisions for significance levels of $\alpha \in \{0.05, 0.1\}$: when the significance level is chosen as $\alpha = 0.1$, all tiles with a reddish shade are rejected, while when working with a level of $\alpha = 0.05$ only tiles shaded in the dark shade are rejected.

Results for both the conservative BH procedure and the liberal IM procedure are shown in Fig. 7. For completeness, results on Holm's method, which is even more conservative than BH, as well as the BH and IM p -values themselves can be found in the supplementary material, Tables C.2 and C.3. One can see that, for a given rejection method (i.e. BH or IM), the MS and bivariate procedures mostly agree on the rejection decisions that would be made at a level of 10% (compare the rows of Fig. 7 to see this). The same holds when working with a significance level of 5%.

Further, as expected, the IM method rejects more hypotheses than the BH method. However, according to the results of the simulation study, it is quite likely that at least one of these rejections is a false discovery.

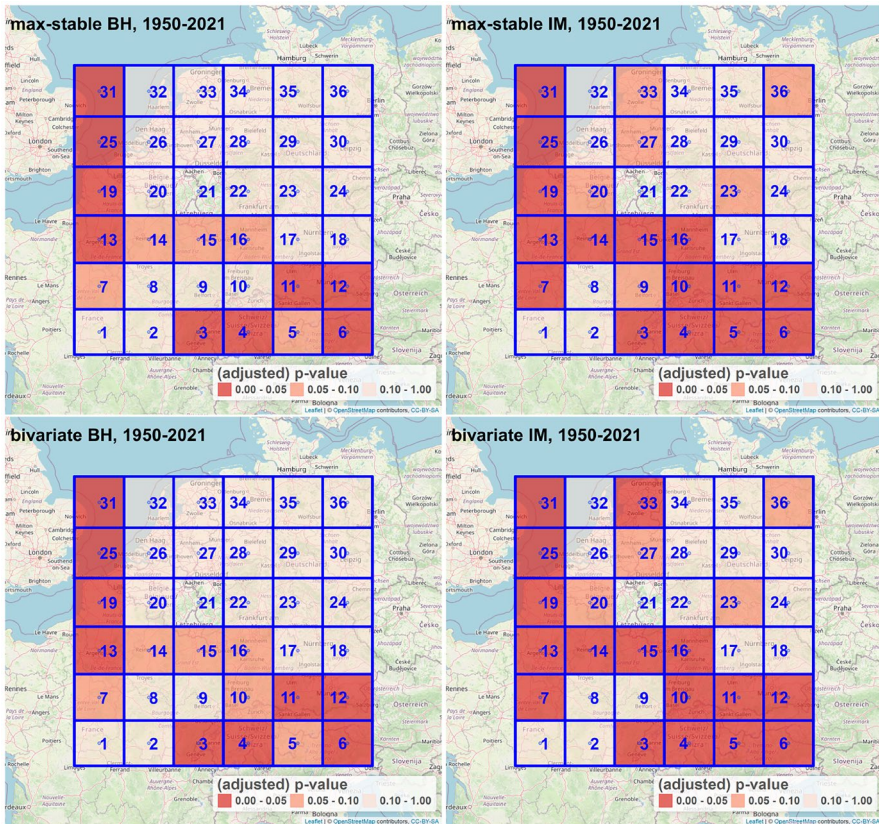


Fig. 7 (Adjusted) p -values obtained with the BH (left) and the IM (right) method on the 6 × 6 grid, with the bootstrap based on max-stable processes (top row) and the bootstrap based on bivariate extreme value distributions (bottom row)

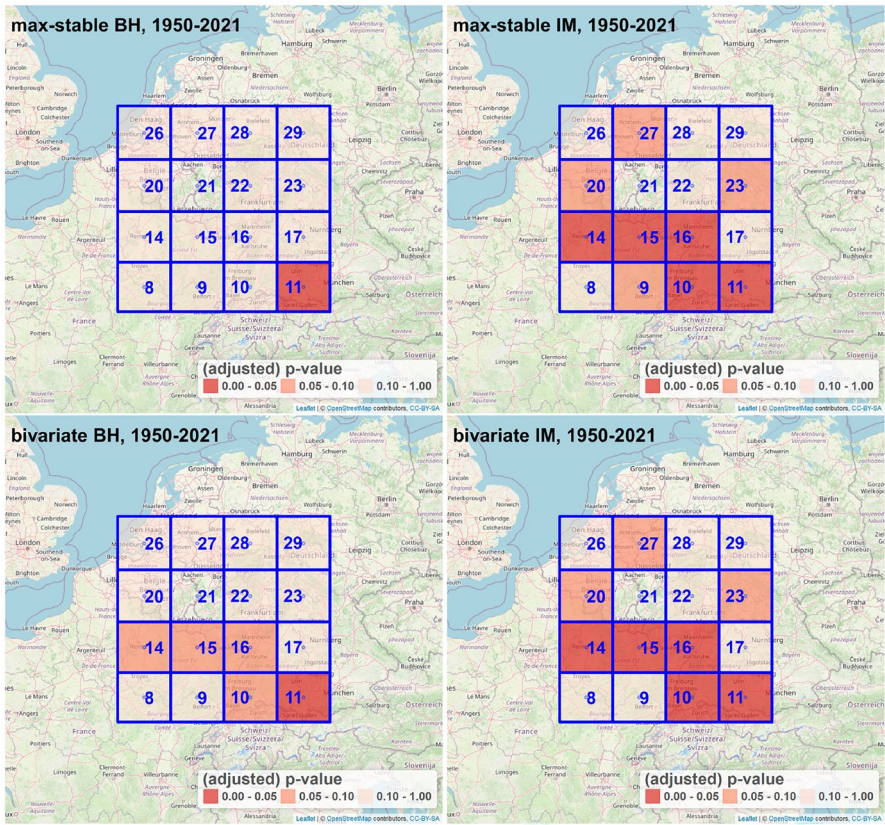


Fig. 8 Adjusted p -values obtained with the BH (left) and the IM (right) method on the 4×4 grid, obtained with the bootstrap based on max-stable processes (top row) and the bootstrap based on bivariate extreme value distributions (bottom row)

Analogous results for the 4×4 grid in (B) are shown in Fig. 8. As discussed above, except for the MS BH method, the results are consistent with the results obtained for the 6×6 grid in the sense that for those locations which are contained in both grids, the locations with p -values of critical magnitude ($< 10\%$) coincide (compare the plot in the upper right corner of Fig. 8 to the plot in the upper right corner of Fig. 7 to see this for the MS IM method, and similar for the other methods). For the MS BH method, grid cells 10, 14, 15, and 16 are not significant anymore at a level of 10 %, but we recorded an adjusted p -value of 0.106 for those four grid cells, so this is a rather tight decision. The p -values obtained for the 4×4 grid can be found in Table C.1 in the supplementary material.

Let us now move on to the interpretation: considering the larger grid first, some grid cells for which the characteristics of extreme precipitation are different (according to expert opinion) from the grid cell of the target location are detected as heterogeneous. These rejected grid cells are located along the coast and in the mountainous

terrain. Comparing the results with Kreienkamp et al. (2021) and Tradowsky et al. (2023), we observe that grid cell 11 has been rejected in their study based on expert knowledge. For grid cell 17, however, we do not detect any statistical evidence that the probabilistic behavior of extreme precipitation is different from the grid cell of the target location, even when applying the liberal IM procedure. We would like to stress though that non-rejection of a null hypothesis does not provide any evidence of the null hypothesis, even when ignoring the multiple testing issue. Hence, expert knowledge that leads to rejection should, in general, outweigh any statistical non-rejection. This particularly applies to the eastern (continental) grid cells in the larger 6×6 -grid, which can be influenced by heavy precipitation caused by different synoptic situations compared to the target region.

Moreover, as the results for locations 10, 14, 15, and 16 showed some discrepancy across the different testing procedures, we suggest that the final decision on the exclusion or inclusion of these locations in a spatial pooling approach should be based on expert knowledge of the meteorological characteristics, and the willingness to trade possible bias for variance (with a possibly larger bias when including the locations – note that statistical evidence against homogeneity in the bivariate extreme value distribution-based bootstrap is only weak, and wrongly declaring the regions as homogeneous is possibly not too harmful). The same holds for locations 9, 20, 23 and 27 for which only the IM method yielded p -values between 5% and 10%. Again, these rather small p -values could be subject to false discoveries though, and since the heterogeneity signal is also not too strong, there is no clear evidence that these need to be excluded from pooling.

For a last evaluation of results from pairwise tests, we estimated the 100-year RLs in the reference climate of the year 2021, i.e. with reference value $\text{GMST}'(2021) = 0.925^\circ\text{C}$, on five different data sets obtained from the 4×4 grid. Here, we use the data sets consisting of data from

- the location of interest only
- pooling those grid cells suggested by the results of the case study (i.e., all cells but 11, or all cells but 10, 11, 14, 15, 16) or expert opinion (i.e., all cells but 11, 17)
- pooling all grid cells of the 4×4 grid.

The results can be found in Table 2 and reveal that excluding cell 11 has a clear effect on the estimated RL. Ex- or including grid cell 17 once 11 is excluded does not have a large effect, while excluding cells 10, 14, 15 and 16 additionally to cell 11 has a moderate effect.

Table 2 Estimated parameters and estimate of $\text{RL}_{2021}(100)$ obtained when pooling the grid cells given in the first column

Pooling cells	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\alpha}$	$\widehat{\text{RL}}_{2021}(100)$
all	20.37	5.80	0.1039	1.50	58.43
all but 11	20.01	5.44	0.0676	1.45	52.74
all but 11, 17	20.01	5.40	0.0760	1.29	52.82
all but 10, 11, 14, 15, 16	19.90	5.41	0.0484	1.79	51.93
only 21	21.92	6.08	0.0634	-0.00	54.37

Finally, we would like to mention that similar results were obtained when applying the BH test procedures to all triples containing the pair of grid cells (20, 21), i.e., the extended target region considered in the study of Kreienkamp et al. (2021) and Tradowsky et al. (2023), consisting of those regions in Germany and Belgium affected worst by the July 2021 flooding.

5 Extensions

In this section, we discuss how to estimate region-wise return levels under homogeneity assumptions (Section 5.1). We also propose two possible extensions of the pooling approach from the previous sections to other hypotheses (Section 5.2) or other underlying model assumptions (Section 5.3).

5.1 Estimation of regional return levels and return periods

As pointed out in Kreienkamp et al. (2021) and Tradowsky et al. (2023) among others, an estimated return period (RP) of T years for a given event and in a fixed reference climate (e.g., the preindustrial climate), obtained based on a fit of the GEV distribution to a pooled homogeneous sample, has the following interpretation: for each fixed location/tile within the region, one can expect one event of the same or larger magnitude within T (imaginary) years of observing the reference climate. We refer to this quantity as the local return period. Obviously, one would expect more than one event of similar magnitude happening at *at least one of the locations* of the pooling region. Likewise, for a given T , one would expect a higher T -year return level for the whole region. The latter corresponds to the value that is expected to be exceeded only once in T years at *at least one of the locations*.

Mathematically, using the notation from Section 2.1, the exceedance probability of value r at at least one among $D \geq 2$ locations in the reference climate corresponding to year t is given by

$$p_t(r) = P(\exists j \in \{1, \dots, D\} : M_j^{(t)} \geq r) = P\left(\max_{d=1, \dots, D} M_d^{(t)} \geq r\right),$$

such that the return period for event r of the region is $RP_t^{\text{reg}}(r) = \frac{1}{p_t(r)}$. Further, the T -year return level of the region in the climate corresponding to year t is the minimal value $RL_t^{\text{reg}}(T)$ for which

$$P\left(\max_{d=1, \dots, D} M_d^{(t)} \geq RL_t^{\text{reg}}(T)\right) \leq \frac{1}{T}$$

holds. Both quantities could be computed (exactly) if one had access to the distribution of $\max_{d=1, \dots, D} M_d^{(t)}$. For example, if the random variables $M_d^{(t)}$, $d = 1, \dots, D$ were independent, $p_t(r)$ could be further simplified to

$$p_t(r) = 1 - P\left(\max_{d=1, \dots, D} M_d^{(t)} \leq r\right) = 1 - (G_{(\mu(t), \sigma(t), \gamma)}(r))^D,$$

where G is the distribution function of the GEV distribution and where $\mu(t), \sigma(t)$ and γ denote the parameters at reference climate of year t from Eq. (2) under the homogeneity assumption from Eq. (3).

The locations are, however, usually not independent in applications. In the following, we propose a simulation-based estimation method that involves max-stable process models to account for the spatial dependence. As before, the R package `SpatialExtremes` (Ribatet 2022) allows for fitting and simulating max-stable process models.

Algorithm 3 (Simulation-based estimation of the regionwise RL and RP)

1. Fit the scale-GEV parameters to the pooled homogeneous sample, resulting in the parameter vector $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\gamma}, \hat{\alpha})^\top$.
2. Transform the margins of the pooled data to approximately unit Fréchet by applying transformation from Eq. (9) with the parameter estimate from Step 1. Then fit several max-stable process models to the obtained data and choose the best fit according to the information criterion CLIC.
3. Replicate for $b = 1, \dots, B$ the following steps:
 - (i) Generate one random observation $(y_{1,b}^{(t),*}, \dots, y_{D,b}^{(t),*})$ from the chosen max-stable process model.
 - (ii) Transform the margins to GEV margins, by applying the transformation in (10) with parameters as estimated in Step 1, resulting in the observation $(m_{1,b}^{(t),*}, \dots, m_{D,b}^{(t),*})$.
 - (iii) Compute the maximum $m_{\max,b}^{(t),*} = \max_{d=1,\dots,D} m_{d,b}^{(t),*}$.
4. The regionwise T -year return level $\text{RL}_{t,\text{reg}}(T)$ and the return period $\text{RP}_{t,\text{reg}}(r)$ of an event with value r can now be estimated from the empirical cumulative distribution function \hat{F}_t^* of the sample $(m_{\max,b}^{(t),*})_{b=1,\dots,B}$ through

$$\widehat{\text{RL}}_t^{\text{reg}}(T) = (\hat{F}_t^*)^{-1}(1 - 1/T), \quad \widehat{\text{RP}}_t^{\text{reg}}(r) = \frac{1}{1 - \hat{F}_t^*(r)}.$$

Especially, when we have estimated the local 100-year RL, we can get an estimate of the return time this event has for the whole region. Likewise, when we have an estimate of the local return period of an event with value r , we can estimate what the event value for that return period would be for the whole region.

We illustrate the estimators for the pooled data sets from Section 4. The estimates are based on $B = 100\,000$ simulation replications and are shown in Table 3. We see that the local 100-year return levels have substantially shorter region-wise return periods. In the region with 15 tiles (only cell 11 excluded), the estimated local 100-year RL at reference climate of 2021 can be expected to be exceeded once in approximately 19 years in at least one of the 15 tiles. We find a similar region-wise return period for the pooling region consisting of 14 tiles. In the pooling region consisting of 11 tiles, the local 100-year return level becomes a region-wise 33-year

Table 3 Estimated local (second column) and regional (fourth column) 100-year RLs for reference climate 2021, for three possible choices of pooling regions as indicated by the first column. Column 3 shows the regional return periods of the local 100-year events

Pooling cells	RL ₂₀₂₁ (100)	RP ^{reg} ₂₀₂₁ (RL ₂₀₂₁ (100))	RL ^{reg} ₂₀₂₁ (100)
all but 11	52.74	18.90	66.40
all but 11, 17	52.82	18.32	67.08
all but 10, 11, 14, 15, 16	51.93	32.76	60.93

event. This comparably larger value arises from the smaller region that is considered: the smaller the region, the less likely it is that one of the locations exceeds a high threshold. Further, as expected, we find that the region-wise 100-year return levels at reference climate of 2021 are larger than their local counterparts. For the regions consisting of 15 and 14 tiles, this increase is approximately 26%, while it is approximately 17.3% for the region consisting of 11 tiles.

5.2 A homogeneous scaling model with location-wise scaling factor

In this section, we maintain the temporal dynamics from the scale-GEV model from Eq. (2). However, instead of testing for the homogeneity assumption from Eq. (3), we additionally allow for a location-wise scaling factor under the null hypothesis. Such an approach can be useful when it is known that observations from different locations occur on different scales, but, apart from that, show a common probabilistic behaviour. In fact, a stationary version of the following model is commonly used in hydrology, where it is known as the Index Flood approach (Dalrymple 1960).

More precisely, suppose that

$$M_{t,d} \sim c_d \exp\left(\frac{\alpha \text{GMST}'(t)}{\mu}\right) \text{GEV}(\mu, \sigma, \gamma) \quad \forall t, d, \tag{15}$$

where $c_d > 0$ is a location-specific scaling factor that we may fix to 1 at the location of primary interest (say $d = 1$, i.e., $c_1 = 1$). Writing $\mu_d = c_d \mu, \sigma_d = c_d \sigma, \alpha_d = c_d \alpha$, the model in Eq. (15) can be rewritten as

$$M_{t,d} \sim \text{GEV}(\mu_d(t), \sigma_d(t), \gamma) \quad \forall t, d,$$

where

$$\mu_d(t) = \mu_d \exp\left(\frac{\alpha_d \text{GMST}'(t)}{\mu_d}\right), \quad \sigma_d(t) = \sigma_d \exp\left(\frac{\alpha_d \text{GMST}'(t)}{\mu_d}\right). \tag{16}$$

Note that the parameters $\mu_1, \dots, \mu_D, \sigma_1, \dots, \sigma_D, \alpha_1, \dots, \alpha_D$ satisfy the relationships

$$\frac{\mu_d}{\sigma_d} \equiv \delta, \quad \frac{\alpha_d}{\mu_d} \equiv \eta, \quad \frac{\alpha_d}{\sigma_d} \equiv \kappa$$

for certain parameters δ, η, κ ; in particular, $\mu_1, \dots, \mu_D, \sigma_1, \dots, \sigma_D, \alpha_1, \dots, \alpha_D$ can be retrieved from $\mu_1, \dots, \mu_D, \delta, \eta$ (note that the constraint on α_d/σ_d is not needed, but comes as a consequence of the first two relations). Fitting this model instead of fitting the scale-GEV distribution to each location separately has the advantage of reducing the number of parameters that need to be estimated substantially ($4 + (D - 1) = D + 3$ instead of $4D$ parameters). Once the local scaling factors are identified, we can bring all observations to the same scale by dividing each location by its location-specific scaling factor.

Now one can test whether such a local scaling model holds on a subset $A = \{d_1, \dots, d_k\} \subset \{1, \dots, D\}$ with cardinality $k = |A| \geq 2$, by testing the hypothesis

$$H_0^{LS}(A) : \exists \delta_A, \eta_A, \gamma_A \forall d \in A : \frac{\mu_d}{\sigma_d} = \delta_A, \quad \frac{\alpha_d}{\mu_d} = \eta_A, \quad \gamma_d = \gamma_A, \quad (17)$$

with a Wald-type test statistic. In this case, the latter is defined as

$$T_n^{LS}(A) = n(g_A(\hat{\theta}))^\top \left(G_A(\hat{\theta}) \hat{\Sigma}_n G_A(\hat{\theta})^\top \right)^{-1} g_A(\hat{\theta}), \quad (18)$$

where $g_A : \mathbb{R}^{4D} \rightarrow \mathbb{R}^{3(k-1)}$ is given by

$$g_A(\theta) = \left(\frac{\mu_{d_1}}{\sigma_{d_1}} - \frac{\mu_{d_2}}{\sigma_{d_2}}, \gamma_{d_1} - \gamma_{d_2}, \frac{\alpha_{d_1}}{\mu_{d_1}} - \frac{\alpha_{d_2}}{\mu_{d_2}}, \dots, \gamma_{d_{k-1}} - \gamma_{d_k}, \frac{\alpha_{d_{k-1}}}{\mu_{d_{k-1}}} - \frac{\alpha_{d_k}}{\mu_{d_k}} \right)^\top,$$

with Jacobian matrix $G_A(\theta) \in \mathbb{R}^{3(k-1) \times 4D}$, since the hypothesis in Eq. (17) may be rewritten as $H_0^{LS}(A) : g_A(\theta) = 0$.

When considering this kind of modification, the bootstrap algorithms from Section 2.4, steps (5)-(7), must be adapted accordingly. In step (5), one has to estimate the parameter under the constraint of the considered null hypothesis by adapting the log-likelihood accordingly. The estimated parameters are then used during the transformation step (6). Further, the test statistic in steps (6) and (7) is replaced by $T_n^{LS}(A)$ from (18). Further details are omitted for the sake of brevity.

5.3 General homogeneous models with smooth parametrization

In this section, we consider general GEV models in which the location, scale and shape parameters are allowed to depend in a (fixed) differentiable way on some parameter vector $\vartheta \in \mathbb{R}^q$ and some temporal covariate $c^{(t)} \in \mathbb{R}^p$ with $p, q \in \mathbb{N}$. More precisely, suppose that f_μ, f_σ and f_γ are (known) real-valued functions of ϑ and c that are differentiable with respect to their first argument ϑ . We assume that, for each $d = 1, \dots, d$, there exists an unknown parameter ϑ_d such that $M_d^{(t)} \sim \text{GEV}(\mu_d(t), \sigma_d(t), \gamma_d(t))$ with

$$\mu_d(t) = f_\mu(\vartheta_d; c^{(t)}), \quad \sigma_d(t) = f_\sigma(\vartheta_d; c^{(t)}), \quad \gamma_d(t) = f_\gamma(\vartheta_d; c^{(t)}).$$

The global null hypothesis of interest within this model is assumed to be expressible as $h(\vartheta_1, \dots, \vartheta_D) = 0$ for a differentiable function $h : \mathbb{R}^{qD} \rightarrow \mathbb{R}^s$ with $s \in \mathbb{N}$.

An example is given by the linear shift model that is frequently considered when modelling temperature extremes in Extreme Event Attribution studies (see Philip et al. 2020), where

$$\mu_d(t) = \mu_d + \alpha_d \text{GMST}'(t), \quad \sigma_d(t) \equiv \sigma_d, \quad \gamma_d(t) \equiv \gamma_d.$$

A possible global null hypothesis of interest could be

$$H_0 : \exists \boldsymbol{\vartheta} \in \mathbb{R} \times (0, \infty) \times \mathbb{R}^2 \forall d \in \{1, \dots, D\} : \boldsymbol{\vartheta}_d = \boldsymbol{\vartheta},$$

where $\boldsymbol{\vartheta} = (\mu, \sigma, \gamma, \alpha)^\top$ and $\boldsymbol{\vartheta}_d = (\mu_d, \sigma_d, \gamma_d, \alpha_d)^\top$.

When considering this kind of extension, one has to adapt the maximum likelihood estimator as well as the estimator of its covariance matrix, hence steps (1)-(2) and (5)-(7) in the bootstrap algorithms are affected. Further details are omitted for the sake of brevity.

6 Conclusion

Extreme event attribution studies can build upon a GEV scaling model. Depending on the analysed variable, it may be useful to apply spatial pooling and fit the GEV distribution to a pooled sample of observations collected at sufficiently homogeneous spatial locations as it has been done in Kreienkamp et al. (2021), Tradowsky et al. (2023) and Vautard et al. (2015), among others. Here, we propose several statistical methods that enable the selection of a homogeneous pooling region from a larger initial region. The BH approach was found to be quite conservative, hence some heterogeneous locations are likely to be declared homogeneous. The IM approach is a more liberal alternative with a higher proportion of rejected locations that may contain some homogeneous ones. In subsequent analyses, the selected pooling region typically results in a classical bias-variance trade-off: the larger the pooling region, the smaller the variance. At the same time, the bias may be larger, given that some heterogeneous regions may have been declared homogeneous. In practice, the tests should always be complemented by expert knowledge on the driving meteorological/climatological background processes.

To make the statistical approach to select homogeneous pooling regions for attribution studies as described here usable for the extreme event attribution community, we have developed a software package that can be freely downloaded and used by applied researchers (Zanger 2022). The selection of spatial pooling regions for attribution studies may hence be based on a combination of expert knowledge and thorough statistical tests. The experts applying the methods can thereby decide between a conservative approach, which tends to reject more locations and a liberal approach which tends to accept more locations as being homogeneous. This decision depends on the a priori knowledge about the meteorology of the analysed area and the specific requirements of the study.

If the ultimate interest is estimation of, for example, return levels, one may, as an alternative to the classical approach based on pooling selected locations, consider penalized maximum likelihood estimators with a penalty on large heterogeneities

(Bücher et al. 2021). A detailed investigation of the resulting bias-variance trade-off would be a worthwhile topic for future research.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10687-023-00480-y>.

Acknowledgements Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf, which is gratefully acknowledged. The authors are grateful to two unknown referees, an associate editor and an editor for their constructive comments on an earlier version of this article which helped to improve the presentation substantially.

Author contributions LZ and AB wrote the main manuscript and worked out the mathematical details. LZ implemented all methods and carried out the Monte Carlo simulation study and the case study. FK, PL and JT contributed by extensive discussions, provided the data for the case study and improved the text.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the integrated project “Climate Change and Extreme Events - ClimXtreme Module B - Statistics (sub-projects B1.2, grant number: 01LP1902B and B3.3, grant number: 01LP1902L)” funded by the German Federal Ministry of Education and Research (BMBF).

Data availability All methods are implemented in the R package `findpoolreg` (Zanger 2022) that is publicly available at <https://github.com/leandrazan/findpoolreg>. The data used throughout the case study is derived from the E-OBS gridded data set, publicly available at <https://www.ecad.eu/download/ensembles/download.php>.

Declarations

Ethical approval Not applicable.

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory (Tsahkadsor, 1971), pp. 267–281. Akadémiai Kiadó, Budapest (1973)
- Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57**(1), 289–300. ISSN 0035-9246 (1995). [http://links.jstor.org/sici?sici=0035-9246\(1995\)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN](http://links.jstor.org/sici?sici=0035-9246(1995)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN)
- Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**(4), 1165–1188 (2001). ISSN 0090-5364. <https://doi.org/10.1214/aos/1013699998>
- Bücher, A., Lilienthal, J., Kinsvater, P., Fried, R.: Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis. *Extremes.* **24**(2), 325–348 (2021). ISSN 1386-1999. <https://doi.org/10.1007/s10687-020-00379-y>

- Bücher, A., Segers, J.: On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes* **20**(4), 839–872 (2017). ISSN 1386-1999. <https://doi.org/10.1007/s10687-017-0292-6>
- Coles, S.: An introduction to statistical modeling of extreme values. Springer Series in Statistics. Springer-Verlag London, Ltd., London. ISBN 1-85233-459-2 (2001). <https://doi.org/10.1007/978-1-4471-3675-0>
- Cornes, R.C., van der Schrier, G., van den Besselaar, E.J., Jones, P.D.: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.* **123**(17), 9391–9409 (2018). <https://doi.org/10.1029/2017JD028200>
- Dalrymple, T.: Flood-frequency analyses, manual of hydrology: Part 3. Technical report, USGPO (1960)
- Davison, A.C., Gholamrezaee, M.M.: Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **468**(2138), 581–608 (2012). <https://doi.org/10.1098/rspa.2011.0412>
- Davison, A.C., Padoan, S.A., Ribatet, M.: Statistical modeling of spatial extremes. *Stat. Sci.* **27**(2), 161–186 (2012). <https://doi.org/10.1214/11-STS376>
- Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**(2), 65–70. ISSN 0303-6898 (1979). <https://www.jstor.org/stable/4615733>
- Kreienkamp, F., Philip, S.Y., Tradowsky, J.S., Kew, S.F., Lorenz, P., Arrighi, J., Belleflamme, A., Bettmann, T., Caluwaerts, S., Chan, S.C., et al.: Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021 (2021). <https://www.worldweatherattribution.org/wp-content/uploads/Scientific-report-Western-Europe-floods-2021-attribution.pdf>
- Lehmann, E.L., Romano, J.P.: Testing statistical hypotheses. Springer Texts in Statistics. Springer, Cham, fourth edition (2021). ISBN 978-3-030-70577-0; 978-3-030-70578-7
- Lilienthal, J., Zanger, L.: Bücher A, Fried R.; A note on statistical tests for homogeneities in multivariate extreme value models for block maxima. *Environmetrics* **33**(7), e2746 (2022). <https://doi.org/10.1002/env.2746>
- Otto, F.E.L., van der Wiel, K., van Oldenborgh, G.J., Philip, S., Kew, S.F., Uhe, P., Cullen, H.: Climate change increases the probability of heavy rains in Northern England/Southern Scotland like those of storm Desmond—a real-time event attribution revisited. *Environ. Res. Lett.* **13**(2), 024006 (2018). <https://doi.org/10.1088/1748-9326/aa9663>
- Philip, S., Kew, S., van Oldenborgh, G.J., Otto, F., Vautard, R., van der Wiel, K., King, A., Lott, F., Arrighi, J., Singh, R., et al.: A protocol for probabilistic extreme event attribution analyses. *Adv. Stat. Climatol. Meteorol. Oceanogr.* **6**(2), 177–203 (2020). <https://doi.org/10.5194/ascmo-6-177-2020>
- R Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2022). <https://www.R-project.org/>
- Ribatet, M.: SpatialExtremes: Modelling Spatial Extremes. R package version 2.1-0 (2022). <https://CRAN.R-project.org/package=SpatialExtremes>
- Stephenson, A.G.: EVD: Extreme value distributions. *R News.* **2**(2) (2002). <https://CRAN.R-project.org/doc/Rnews/>
- Tradowsky, J.S., Philip, S.Y., Kreienkamp, F., Kew, S.F., Lorenz, P., Arrighi, J., Belleflamme, A., Bettmann, T., Caluwaerts, S., Chan, S.C., Ciavarella, A., De Cruz, L., de Vries, H., Demuth, N., Ferrone, A., Fischer, E.M., Fowler, H.J., Goergen, K., Heinrich, D., Henrichs, Y., Lenderink, G., Kaspar, F., Nilson, E., Otto, F.E.L., Ragone, F., Seneviratne, S.I., Singh, R.K., Skålevåg, A., Termonia, P., Thalheimer, L., van Aalst, M., Van den Bergh, J., Van de Vyver, H., Vannitsem, S., van Oldenborgh, G.J., Van Schaeybroeck, B., Vautard, R., Vonk, D., Wanders, N.: Attribution of the heavy rainfall events leading to severe flooding in Western Europe during July 2021. *Clim. Chang.* **176**(7), 1–38 (2023). <https://doi.org/10.1007/s10584-023-03502-7>
- van der Vaart, A.W.: Asymptotic Statistics, volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1998). ISBN 0-521-49603-9; 0-521-78450-6
- van der Wiel, K., Kapnick, S.B., van Oldenborgh, G.J., Whan, K., Philip, S., Vecchi, G.A., Singh, R.K., Arrighi, J., Cullen, H.: Rapid attribution of the August 2016 flood-inducing extreme precipitation in south Louisiana to climate change. *Hydrol. Earth Syst. Sci.* **21**(2), 897–921 (2017). <https://doi.org/10.5194/hess-21-897-2017>
- van Oldenborgh, G.J., van der Wiel, K., Sebastian, A., Singh, R., Arrighi, J., Otto, F., Haustein, K., Li, S., Vecchi, G., Cullen, H.: Attribution of extreme rainfall from Hurricane Harvey, August 2017. *Environ. Res. Lett.* **12**(12), 124009 (2017). <https://doi.org/10.1088/1748-9326/aa9ef2>
- Vautard, R., Yiou, P., van Oldenborgh, G.J., Lenderink, G., Thao, S., Ribes, A., Planton, S., Dubuisson, B., Soubeyroux, J.M.: Extreme fall 2014 precipitation in the Cévennes Mountains. *Bull. Am. Meteorol. Soc.* **96**, S56–S60 (2015). https://doi.org/10.1175/BAMS-EEE-2014_ch12.1

Zanger, L.: findpoolreg: Find a possible pooling region for precipitation extremes. R package version 0.1.0 (2022). <https://github.com/leandrazan/findpoolreg>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.