Bayesian spatial clustering of extremal behaviour for hydrological variables

Christian Rohrbeck[1,2] and Jonathan A. Tawn[1]

[1] Department of Mathematics and Statistics, Lancaster University

[2] Department of Mathematical Sciences, University of Bath

Abstract

To address the need for efficient inference for a range of hydrological extreme value problems, spatial pooling of information is the standard approach for marginal tail estimation. We propose the first extreme value spatial clustering methods which account for both the similarity of the marginal tails and the spatial dependence structure of the data to determine the appropriate level of pooling. Spatial dependence is incorporated in two ways: to determine the cluster selection and to account for dependence of the data over sites within a cluster when making the marginal inference. We introduce a statistical model for the pairwise extremal dependence which incorporates distance between sites, and accommodates our belief that sites within the same cluster tend to exhibit a higher degree of dependence than sites in different clusters. By combining the models for the marginal tails and the dependence structure, we obtain a composite likelihood for the joint spatial distribution. We use a Bayesian framework which learns about both the number of clusters and their spatial structure, and that enables the inference of site-specific marginal distributions of extremes to incorporate uncertainty in the clustering allocation. The approach is illustrated using simulations, the analysis of daily precipitation levels in Norway and daily river flow levels in the UK.

1 Introduction

Statistical models for estimating the frequency and size of flood and severe rainfall events are required by decision makers to construct effective protection measures and by risk analysts to set insurance premiums. Since such extreme events occur rarely at a site of interest, model-based estimates for the behaviour of extremes at a site are usually derived using a small number of observations, inducing high uncertainty. With a view to obtaining more reliable estimates, pooling of information from other sites can be used; this

1

is the basis of regional methods which are widely used for environmental, meteorological or hydrological hazards (Hosking and Wallis, 1993; Casson and Coles, 1999; Sang and Gelfand, 2009; Asadi et al., 2018). Natural questions that arise in this context are what criteria should be used to group sites, and how should the uncertainty in the clustering allocation, given this selected criterion, be reflected in the uncertainty in the estimated site-specific marginal distributions, and what do the estimated clusters look like?

We propose a novel approach for spatial clustering of extremal behaviour and the subsequent inference which coherently addresses these questions. Our methodology is motivated by two applications described in Section 2: modelling weather-related insurance claims and flood risk analysis. In both applications, we are interested in the extremal behaviour of a hydrological variable across multiple spatial locations. Each of these problems requires the marginal analysis of extreme values at different locations, whilst accounting for spatial structure both in the marginal distributions and in the dependence of the data from across sites.

Our approach for modelling both marginal and dependence structures is to use statistical models that have been asymptotically justified by extreme value theory (Coles, 2001; Beirlant et al., 2004) and that have, for instance, been applied in climatology (Blanchet and Davison, 2011; Reich et al., 2014) and finance (Chavez-Demoulin et al., 2014; Hilal et al., 2014). When considering the extremes of a univariate random variable X, we adopt the peaks-over threshold approach. Exceedances by X of a high threshold u are then modelled using the generalized Pareto distribution $\text{GPD}(\psi, \nu)$ with

$$\mathbb{P}(X \leq x + u \mid X > u) \; = \; 1 - \left(1 + \nu \frac{x}{\psi}\right)_+^{-1/\nu} \qquad \text{for} \quad x > 0, \qquad (1)$$

where $y_+ = \max(y, 0)$, and $\psi > 0$ and $\nu \in \mathbb{R}$ are scale and shape parameters. In practice, the threshold u is often selected using graphical diagnostic tools, such as the mean residual life and threshold stability plots (Coles, 2001). More recent techniques are described by Wadsworth (2016) and Northrop et al. (2017). When fitting extreme value models over a range of variables (e.g., the same physical variable measured at different sites, or different variables measured at the same time), it is natural to model the tails of the marginal distributions as GPD, with the parameters (ψ, ν) potentially changing over variables (Davison and Smith, 1990; Cooley et al., 2007; Wang and So, 2016; Sharkey and Winter, 2019).

There are a number of ways for modelling dependence in multivariate and spatial extremes. Multivariate approaches have included fitting parametric models for multivariate extreme value copula (Tawn, 1988) and various threshold methods (Ledford and Tawn, 1997; Rootzén et al., 2018). In a spatial context, max-stable processes are the most widely used (Davison et al., 2012; Reich and Shaby, 2012). However, they have a number of inference and model limitations. Inference issues are mostly overcome by instead using

Pareto processes (Ferreira and de Haan, 2014; Dombry and Ribatet, 2015) or exploiting graphic structure within a Bayesian inference framework (Vettori et al., 2019; Engelke and Hitz, 2020); however these models give either a strong form of extremal dependence (corresponding to assuring asymptotic dependence (Ledford and Tawn, 1997), i.e., that $\chi_{k,k'} > 0$ in (3)) across all sites or give independence. Alternative methods that allow for spatial dependence that weakens with extremal level also exist (Wadsworth and Tawn, 2012, 2019).

All these full spatial extreme models are not ideal for the class of problems that include our two applications as these models all require the following conditions: marginal distributions changing smoothly over space with the level of smoothness predetermined; a parametric spatial dependence structure model, typically requiring assumptions of stationarity and higher order distributions to be well modelled; and small numbers of data sites for feasible inference. In contrast we need methods for identifying regions of similar behaviour of the spatial process (so identical margins and strong spatial dependence); marginal parameters that can change rapidly over space to reflect physical features, i.e., altitude and prevailing weather directions; that are robust to complex spatial dependence structure; and that scale well as the number of sites gets large. We believe that the clustered spatial extreme model we develop meets these needs.

Two broad clustering approaches have emerged in the literature: methods that aim purely to find clusters of similarly distributed variables for purposes of interpretation; and methods that aim to pool information over similarly distributed variables to enhance inference efficiency. The first category of methods tend to evaluate extreme value theory summary statistics (e.g., the GPD shape parameter from each site) and apply widely used generic clustering techniques (e.g., k-means or k-medoids (Kaufman and Rousseeuw, 2005)) to form clusters. Such approaches have been proposed by Rubio et al. (2018) to analyze extreme stock market losses, and by Bernard et al. (2013), Bador et al. (2015) and Mornet et al. (2017) to explore different levels of pairwise dependence across sites for environmental variables.

The latest approaches in the second category of clustering methods use hierarchical modelling. First versions go back to the pooling methods used by the Flood Studies Report (1975) (Institute of Hydrology (Great Britain), 1975) which selected a hydrologically coherent region and assumed a common GPD shape parameter for all sites within the region. The methods evolved to also account for the dependence structure (Coles and Tawn, 1990, 1996), but these methods do not account for the uncertainty in the process of identifying the regions/clusters of similar variables. The first hierarchical method which accounted for the cluster uncertainty in the inference for extremes was by Smith and Goodman (2000), later extended to a Bayesian mixture model by Bottolo et al. (2003), who consider borrowing information across multiple types of independent insurance claims. For spatial extreme value problems, hierarchical clustering models have been proposed

3

by Carreau et al. (2017) and Reich and Shaby (2019), with both approaches having limitations. The former does not account for dependence over space. Although this feature is somewhat addressed by Reich and Shaby (2019), they model dependence over variables in the same cluster with a restrictive exchangeable parametric extreme value copula, which is likely to be too simplistic and as a consequence bias the marginal inference (Dupuis and Tawn, 2001). Furthermore, the approach has a high computational cost and their Bayesian framework does not allow the number of clusters to change.

The existing methods for spatial clustering of extremes either focus solely on criteria based on the similarity of the marginal distributions or they only consider dependence features. No methods look at both aspects yet, from a physical perspective, spatial dependence of a process is the key determinant of the marginal distributions being similar. Furthermore, the vast majority of approaches ignore the clustering uncertainty (both numbers of clusters and the allocation of variables to clusters) and neglect the spatial dependence between variables within a cluster in their marginal inference.

Our proposed approach is the first to address all of these features in a single Bayesian framework and it is applicable to both areal and geostatistical data. It learns about both the number of clusters and their spatial structure with respect to the site-wise distribution of the peaks-over threshold and spatial dependence in the extremes. Critically, we assume the following features: within localised clusters the process is assumed to be identically marginally distributed but can change abruptly between clusters; that spatial dependence can be accounted for nonparametrically; and that the modelling of a summary measure of spatial pairwise extremal dependence between sites is sufficient, with dependence within and across clusters handled differently. This local clustering enables large scale spatial application, clusters that are identified to have value to practitioners, and for uncertainty in clustering to be accounted for leading to marginal tail models which model average over a range of GPDs.

Unlike Reich and Shaby (2019) we do not attempt to estimate a full model for the spatial dependence structure, instead, similar to Bernard et al. (2013), we account for dependence through a widely used pairwise measure of extremal dependence (Coles et al., 1999). We introduce a statistical model for the extremal dependence measure which incorporates both distance between sites and our belief that sites within the same cluster tend to exhibit a higher degree of dependence than sites in different clusters. Our approach to impose spatial structure on the parameters of the site-wise GPD is similar to Bottolo et al. (2003), but the additional consideration of spatio-temporal dependence makes it a more general and harder problem. Posterior samples are obtained using a reversible jump MCMC algorithm (Green, 1995) which allows: the number of spatial clusters to vary, the analysis of the site-wise marginal tail behaviour, and the derivation of a point estimate for the cluster structure using Bayesian decision theory.

The paper is organized as follows: Section 2 describes the motivating applications and

introduces the data; Sections 3 and 4 detail the statistical modelling framework and the inference procedure; there is a simulation study in Section 5 and Section 6 presents a data analysis for the applications; and we conclude with a discussion in Section 7.

2 Motivating examples and data

2.1 Weather-related property insurance claims in Norway

Insurance companies are interested in the distribution of the number of claims they are likely to receive over different areas as a consequence of rainfall and/or snow-melt. Scheel et al. (2013) showed that the upper tail of this distribution is exceptionally difficult to model for individual administrative areas (termed municipalities) in Norway. Rohrbeck et al. (2018) developed an extremal regression model that shows how the largest numbers of insurance claims per weather event in a municipality can be clearly linked to extreme precipitation in that municipality, and that from such a model, marginalizing over the precipitation gives a good model for the marginal distribution of the number of claims.

While their approach achieves good results for three highly populated Norwegian municipalities, its complexity requires a high number of insurance claims in order to obtain reliable estimates. As this criterion is not satisfied by most Norwegian municipalities, there is a need to exploit spatial structure to borrow information across municipalities and instead model the total claims over clusters of municipalities. The question then is what is the precipitation value to use and how to model that? Although the probability of a claim given an extreme precipitation is likely to change very slowly over space, the distribution of extreme precipitation varies more rapidly due to geographical reasons. Therefore, we need to identify clusters of municipalities which have both the same distribution of extreme precipitation and have similar actual values in each extreme event (i.e., strong extremal dependence), so that the average precipitation over the cluster can be used as a covariate for the aggregated claims across the municipalities within the cluster.

We consider precipitation across the 343 municipalities in South Norway. These areal units, shown in Figure 1 left panel, differ substantially in size, ranging from a few through to several hundred square kilometres. The data were produced by the Norwegian Meteorological Institute (`www.met.no`) and provide the daily amount of precipitation (in mm), including both rain and snowfall, between 1997 and 2006. Data for each municipality are obtained by a two-step process. First, point observations from more than 200 measurement stations across Norway are spatially interpolated to a regular grid of 1km^2. Then, for each municipality, the precipitation is obtained by a weighted averaging over the grid within the municipality's boundary, with weights proportional to the population density (Haug et al., 2011). The data exhibit a small number of missing values; 326 of the 343 municipalities have a full record and only six have more than 20 observations missing.
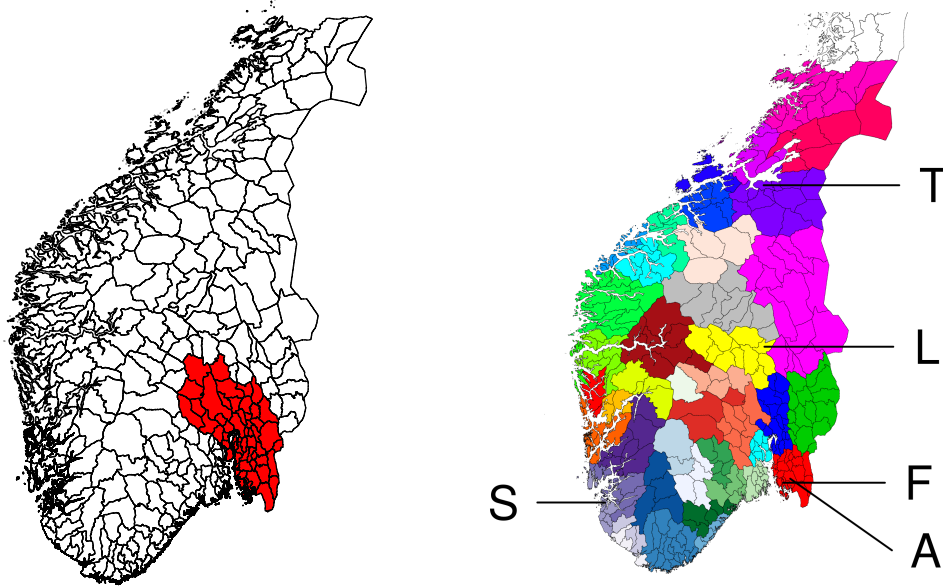
Figure 1: Map of South Norway showing the boundaries of the municipalities (left) and the estimated cluster structure using our approach (right). The coloured region in the left panel is considered in Section 5. The highlighted municipalities in the right panel are Fredrikstad (F), Sarpsborg (A), Lillehammer (L), Stavanger (S) and Trondheim (T).

We find substantial spatial variation in the average daily precipitation; a detailed plot is provided in the Supplementary Material. The highest average values are recorded along the west coast and the averages decrease typically with increasing easterly coordinates and distance from the coast. All municipalities exhibit seasonality, with the largest average daily precipitation typically in September and October, but with the west coast having higher average precipitation levels in January to March, while these are the driest months, on average, for the municipalities in the south-east.

To focus on the larger events, we explore the exceedances of a threshold, corresponding to the annual 90% quantile for each municipality. For each municipality, the average number of events exceeding the threshold varies across the year, however, the excess values themselves are found not to exhibit seasonality. Similar results are found for all higher thresholds. Thus site-wise peaks over threshold can be considered as identically distributed over time but they do exhibit spatial variation.

2.2 Flood risk analysis for the UK

Improving river flood risk analysis in the UK is of high priority given that it has experienced several severe and widespread flood events over the past years. For instance, the floods related to Storm Desmond, Storm Eva and Storm Frank in 2015/2016 caused an estimated economic damage of between £1.3–1.9 billion (Environment Agency, 2018). Practitioners are interested in the potential size of future flood events, as well as their spatial extent. Detecting groups of sites which have similar dynamics in terms of extreme
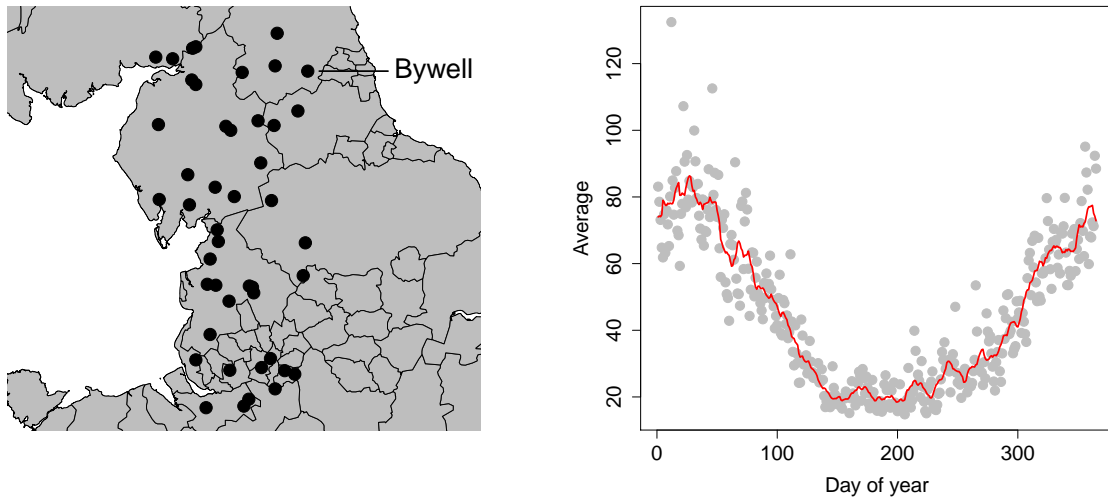
Figure 2: Locations of the 45 river flow gauges considered in Section 2.2 (left) and the daily average river flow for the Bywell gauge on the River Tyne (right). The solid line in the right panel shows the two-week moving average over the year for Bywell.

river flow levels helps to address this question as we can combine them in the inference to improve statistical efficiency and hence produce estimates that give more reliable extrapolations to rarer events. However, extreme rainfall events in different seasons have different spatial characteristics due to frontal rainfall in summer and convective rainfall in winter; so clusterings may differ between seasons.

Daily mean river flow levels (in m³/s) for the years 1980 to 2011 for 45 gauges are obtained from the UK's National River Flow Archive (`nrfa.ceh.ac.uk`) and cover northern England and southern Scotland, with the majority being in North West England; see Figure 2 left panel. Observations for several years are missing for two gauges and other stations also exhibit some missing values. The data include the river flow levels for the floods in Cumbria in November 2009. Hydrological distances which account for catchment closeness are available; these provide a better emulation of the spatial dependence than the geographical distances (Asadi et al., 2015).

Figure 2 right panel indicates strong seasonality for one of the gauges, a feature typical of all gauges in the region; the highest average river flow levels are observed for November through to March, while June through to August record the lowest averages. As such, a river flow level which is considered very high in summer may be rather standard in winter. Moreover, the data exhibit strong autocorrelation. Therefore, an extreme weather event may cause extreme river flow levels over consecutive days. The data values vary substantially across gauges; the site-wise daily average ranges from 0.2m³/s up to 54.9m³/s. All these aspects, that is seasonality, spatio-temporal dependence and difference in scale, are considered in our analysis in Section 6.

7

3 Cluster model

3.1 Introduction

Consider K sites with spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_K \in \mathbb{R}^2$, where $\mathbf{s}_k$ $(k = 1, \ldots, K)$ refers to the centroid of the k-th areal unit (Section 2.1) or the point location of the k-th site (Section 2.2). Spatial proximity of any pair (k, k') of sites is measured via a suitable metric which provides a distance $d_{k,k'} \geq 0$ based on $\mathbf{s}_k$ and $\mathbf{s}_{k'}$. For each site, we have data for the variable we wish to draw inference on the distribution of its largest values.

Section 2 shows that hydrological processes usually exhibit seasonality and spatio-temporal dependencies. To account for temporal dependence, we use declustering prior to any model fitting, with each of the K time series being split into subperiods such that extremes occurring in different subperiods can be assumed to be independent (Ferro and Segers, 2003). We select the same set of subperiods for each site and only use the highest observation per subperiod and site in our analysis.

Let $R_{k,1}, \ldots, R_{k,T}$ denote the time series for site k after declustering, i.e., $R_{k,t}$ and $R_{k',t'}$ are assumed to be independent for any $t \neq t'$ $(t, t' = 1, \ldots, T)$. If $d_{k,k'}$ is small, it is typically reasonable to assume that the marginal distributions for $R_{k,t}$ and $R_{k',t}$ are very similar, therefore they should have similar GPD parameter values in (1), with exceptions occuring when geophysical features change rapidly. Furthermore, spatial dependence leads to the same extreme event being present at different sites. Thus, if $R_{k,t}$ is large then the chances of $R_{k',t}$ being large presumably increase if $\mathbf{s}_k$ and $\mathbf{s}_{k'}$ are close, relative to them being further apart.

We aim to group the K sites, such that sites in the same cluster have both similar marginal distributions and the spatial dependence is greater between sites in the same cluster than between sites in different clusters. To represent the cluster structure, we introduce K latent random variables $\mathbf{Z} = (Z_1, \ldots, Z_K)$. Let $J \in \{1, \ldots, K\}$ denote the number of clusters. Then, $Z_k = j$ $(j \in \{1, \ldots, J\})$ corresponds to the k-th site being allocated to the j-th cluster. Conditional on $\mathbf{Z}$, we propose separate models for the marginal distributions of $R_{k,t}$ and the spatial dependence in Sections 3.2 and 3.3 respectively. We later combine these models in Section 4.

3.2 Model for Marginal Clustering

We first select thresholds $u_1, \ldots, u_K$ using graphical diagnostic tools and model $R_{k,t} - u_k \mid (R_{k,t} > u_k)$ as following a $\text{GPD}(\psi_k, \nu_k)$ $(k = 1, \ldots, K)$ as in (1). To group sites with very similar marginal tail behaviour, the peaks-over threshold for all sites within a cluster are modelled by a GPD with the same scale and shape parameters. The distribution of the

peaks-over threshold for the k-th site, conditional on Z_k, is then

$$R_{k,t} - u_k \mid (Z_k = j\,,\, R_{k,t} > u_k) \ \sim \ \text{GPD}\,(\sigma_j\,,\, \xi_j)\,, \tag{2}$$

where $\sigma_j > 0$ and $\xi_j \in \mathbb{R}$ $(j = 1, \ldots, J)$ denote the cluster-specific scale and shape parameters, so $(\psi_k, \nu_k) = (\sigma_j, \xi_j)$ if $Z_k = j$. We denote the parameters of the model given $\mathbf{Z}$ by $\boldsymbol{\theta}_{\text{M}}^{(J)} = \big\{\boldsymbol{\sigma}^{(J)}, \boldsymbol{\xi}^{(J)}\big\}$, where $\boldsymbol{\sigma}^{(J)} = (\sigma_1, \ldots, \sigma_J)$ and $\boldsymbol{\xi}^{(J)} = (\xi_1, \ldots, \xi_J)$.

The model (2) assumes stationarity of the tail of $R_{k,1}, \ldots, R_{k,T}$. However, our data analysis in Section 2 indicates that the marginals are not identically distributed due to seasonality. There are a range of established approaches to handle this: to split the time series into shorter time periods for which the assumption of stationarity seems reasonable; model the threshold and GPD parameters in terms of a set of temporal predictors (Davison and Smith, 1990; Chavez-Demoulin and Davison, 2005); or to preprocess the data to remove non-stationary to the overall series (Eastoe and Tawn, 2009). For simplicity, we assume that $R_{k,1}, \ldots, R_{k,T} \sim R_k$ are identically distributed when presenting the methods, and discuss the handling of non-stationarity later in Sections 6 and 7.

3.3 Model for Spatial Dependence Clustering

Instead of attempting to model the joint occurrence of extreme events across all subsets of sites, we only consider pairwise extremal dependence. The most widely used extremal dependence measure is the coefficient of asymptotic dependence χ (Coles et al., 1999). Formally, for the random variables R_k and $R_{k'}$ $(k, k' = 1, \ldots, K)$, define $\chi_{k,k'} = \lim_{v \to 1} \chi_{k,k'}(v)$ where

$$\chi_{k,k'}(v) \ = \ \mathbb{P}\,[\,F_k(R_k) > v \mid F_{k'}(R_{k'}) > v\,], \qquad v \in [0,1], \tag{3}$$

with $F_k(\cdot)$ and $F_{k'}(\cdot)$ denoting the cumulative distribution functions of R_k and $R_{k'}$, respectively. When $\chi_{k,k'} > 0$, R_k and $R_{k'}$ are termed asymptotically dependent, with increasing values corresponding to stronger extremal dependence, whilst we say that R_k and $R_{k'}$ are asymptotically independent if $\chi_{k,k'} = 0$. There are strong parallels between $\chi_{k,k'}$ and the extremogram (Davis and Mikosch, 2009) at distance $d_{k,k'}$, with the key difference being that inference for the extremogram pools together all the data from pairs of sites with the same separation under the assumption of stationarity.

We model $\chi_{k,k'}$ for all k and k' as the realization of an underlying process, conditional on a clustering $\mathbf{Z}$, which explains varying spatial extremal dependence within and between clusters. First consider what properties the expected value of this conditional distribution should possess. Since sites (k, k') within the same cluster are expected to have a higher probability of joint extreme events, $\mathbb{E}(\chi_{k,k'} \mid \mathbf{Z})$ should be larger for $Z_k = Z_{k'}$ than when $Z_k \neq Z_{k'}$, that is,

$$\mathbb{E}\,(\chi_{k,k'} \mid Z_k = Z_{k'}) \ \geq \ \mathbb{E}\,(\chi_{k,k'} \mid Z_k \neq Z_{k'})\,, \tag{4}$$

for all $k, k' = 1, \ldots, K$. Furthermore, $\chi_{k,k'}$ should decrease with increasing distance $d_{k,k'}$, irrespective of whether the sites k and k' are in the same cluster or not. We assume that $\mathbb{E}(\chi_{k,k'} \mid \mathbf{Z})$ decays exponentially with $d_{k,k'}$, but with a varying rate across each cluster and with a common, but faster, decay rate between clusters. These properties are reflected in the formulation

$$
\mathbb{E}\left(\chi_{k,k'} \mid \mathbf{Z}\right) = \begin{cases} \exp\left(-\gamma_j \ d_{k,k'}\right) & \text{if} \quad Z_k = Z_{k'} = j, \\ \exp\left(-\gamma_0 \ d_{k,k'}\right) & \text{if} \quad Z_k \neq Z_{k'}, \end{cases} \tag{5}
$$

where $\gamma_j > 0$ $(j = 1, \ldots, J)$ is cluster-specific and $\gamma_0 > \max(\gamma_1, \ldots, \gamma_J) \geq 0$ to ensure condition (4). Our approach is not limited to an exponential decay as in (5), and any other correlation function is a valid alternative. If $J = 1$, the expectation in (5) is fully defined by the parameter γ_1 with $\mathbb{E}(\chi_{k,k'} \mid \mathbf{Z}) = \exp(-\gamma_1 d_{k,k'})$. For $J \geq 2$, we ensure that $\gamma_0 > \max(\gamma_1, \ldots, \gamma_J)$ by defining parameters $\epsilon_1, \ldots, \epsilon_J$ such that

$$
\log\left(\gamma_j\right) = \log(\gamma_0) - \epsilon_j, \qquad \epsilon_j \geq 0, \qquad (j = 1, \ldots, J). \tag{6}
$$

To complete, we define the distribution of $\chi_{k,k'} \mid \mathbf{Z}$. As $\chi_{k,k'} \in [0, 1]$ may differ between two pairs of sites in the same cluster with the same $d_{k,k'}$, due to factors such as topology, we choose a beta distribution model with

$$
\chi_{k,k'} \mid \mathbf{Z} \sim \begin{cases} \mathrm{Beta}\left(\frac{\beta \exp\left(-\gamma_j \ d_{k,k'}\right)}{1 - \exp\left(-\gamma_j \ d_{k,k'}\right)} \ , \ \beta\right) & \text{if} \quad Z_k = Z_{k'} = j, \\ \mathrm{Beta}\left(\frac{\beta \exp\left(-\gamma_0 \ d_{k,k'}\right)}{1 - \exp\left(-\gamma_0 \ d_{k,k'}\right)} \ , \ \beta\right) & \text{if} \quad Z_k \neq Z_{k'}, \end{cases} \tag{7}
$$

which has expectation given by (5) and where $\beta > 0$ controls the variance of $\chi_{k,k'}$; higher values of β correspond to $\chi_{k,k'}$ being less variable. Note, $\chi_{k,k'} \to 1$ as $d_{k,k'} \to 0$, and the spatial variation of the coefficient of asymptotic dependence, conditional on $\mathbf{Z}$, is given by the $J + 2$ parameters γ_0, $\boldsymbol{\epsilon}^{(J)} = (\epsilon_1, \ldots, \epsilon_J)$ and β for $J \geq 2$, while it is fully specified by γ_1 and β for $J = 1$. We denote the dependence parameters given $\mathbf{Z}$ by $\boldsymbol{\theta}_{\mathrm{D}}^{(J)} = (\gamma_0, \boldsymbol{\epsilon}^{(J)}, \beta)$.

The distribution (7) could be applied to model $\chi_{k,k'}$ for all pairs of sites (k, k'). However, this may not be optimal. For $Z_k \neq Z_{k'}$, $\chi_{k,k'}$ may vary strongly, depending on whether sites k and k' belong to adjacent clusters or whether there are multiple clusters along the path between the two sites. Such differences can probably not be captured by the single parameter γ_0. We thus consider $\chi_{k,k'}$ for adjacent pairs of sites only. In case of the sites being point locations, we first derive the Voronoi partition of the study area and then define sites as being adjacent if their corresponding Voronoi cells are adjacent.

4 Bayesian Inference

4.1 Introduction

We use Bayesian inference for the number of clusters, J, the latent variables $\mathbf{Z}$, and the marginal and dependence structure parameters $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$ using the declustered data $\mathcal{D} = \{\,(r_{k,1}, \ldots, r_{k,T}) \,:\, k = 1, \ldots, K\,\}$; the model structure is illustrated graphically in the Supplementary Material. The posterior, and the algorithm to sample from it, are developed in this section. The derivation of the marginal and dependence structure likelihood contributions L_{M} and L_{D}, given $\mathbf{Z}$ and $\mathcal{D}$, are provided in Section 4.2. Critically, the data that are used to analyze the marginal and dependence structures are different, in that for marginal distributions we use all marginal exceedances of a threshold at all sites, whereas for the dependence model, only ranks of the variables are used. These two forms of the data are only weakly dependent. Furthermore, Genest et al. (1995) and Genest and Segers (2009) have shown that inference for dependence parameters is largely unaffected by marginal parameter estimation. Therefore we use the following approximation to the joint likelihood, given by the independent likelihood

$$L\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}, \mathbf{u}\right) \;=\; L_{\mathrm{M}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}, \mathbf{u}\right) \times L_{\mathrm{D}}\left(\boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right), \qquad (8)$$

where $\mathbf{u} = (u_1, \ldots, u_K)$ denotes the vector of selected thresholds. Section 4.3 presents our priors, including a spatial prior for $\mathbf{Z}$ given J. Section 4.4 details our algorithm to sample from the posterior distribution, and Section 4.5 outlines the analysis of the posterior samples and the estimation of the underlying cluster structure.

4.2 Likelihood Components

4.2.1 Marginal component

If the peaks-over threshold data were independent over all sites, the likelihood function for $\boldsymbol{\theta}_{\mathrm{M}}$, conditional on $\mathbf{Z}$, the data $\mathcal{D}$ and thresholds $\mathbf{u}$ would be

$$L_{\mathrm{M}}^{\mathrm{ind}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}, \mathbf{u}\right) \;=\; \prod_{k=1}^{K} \prod_{\{t \,:\, r_{k,t} > u_k\}} \frac{1}{\sigma_{Z_k}} \left(1 + \xi_{Z_k} \frac{r_{k,t} - u_k}{\sigma_{Z_k}}\right)_{+}^{-1/\xi_{Z_k} - 1}. \qquad (9)$$

However, spatial independence is not a valid assumption for our motivating problems since severe weather events usually affect a number of sites (municipalities/gauges). Therefore, the likelihood function in expression (9) corresponds to a misspecified model. Inference under the misspecified model, i.e., when there is spatial dependence, would underestimate the variance of the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)}$. With respect to our Bayesian framework, this would correspond to credible intervals of the parameters being too narrow.

Under suitable regularity conditions, Kent (1982) shows that a general theory for the asymptotic distribution of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)}$ based on the misspecified likelihood (9) is

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)} - \boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right) \sim \mathrm{Normal}\left(\mathbf{0}, \Sigma = H\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right)^{-1} V\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right) H\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right)^{-1}\right), \qquad (10)$$

where $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ are the true model parameters, $H\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right) = -\mathbb{E}\left[\nabla^2 \log L_{\mathrm{M}}^{\mathrm{ind}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)\right]$ denotes the Fisher information, $V\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right) = \mathrm{Cov}\left[\nabla \log L_{\mathrm{M}}^{\mathrm{ind}}\left(\boldsymbol{\theta}_{\mathrm{M}} v \mid \mathcal{D}, \mathbf{Z}\right)\right]$ and ∇^i refers to the i-th derivative. The limiting variance in (10) is different from the classic Fisher information if the model was misspecified, but that if we have no spatial dependence, $H\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right) = V\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right)$ and then the classic asymptotic result is obtained.

We follow an approach of Ribatet et al. (2012), who propose an adjustment of the curvature of the likelihood (9) around its mode using the asymptotic behaviour in expression (10). The adjusted likelihood is

$$L_{\mathrm{M}}^{\mathrm{adj}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}, \mathbf{u}\right) = L_{\mathrm{M}}^{\mathrm{ind}}\left(\hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)} + B\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} - \hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)}\right) \mid \mathcal{D}, \mathbf{Z}, \mathbf{u}\right), \qquad (11)$$

where the $2J \times 2J$ matrix B depends on $\mathbf{Z}$ and is

$$B = \left\{\left[H\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}\right)\right]^{1/2}\right\}^{-1}\left[\Sigma^{-1}\right]^{1/2} \qquad (12)$$

with $[\cdot]^{1/2}$ denoting the matrix square root and Σ is defined in (10). To compute B, the evaluation of H and V matrices uses the observed information matrix and an estimate for the covariance matrix of the score function respectively, both evaluated with $\boldsymbol{\theta}_{\mathrm{M}}^{(J)} = \hat{\boldsymbol{\theta}}_{\mathrm{M}}^{(J)}$. Note, B is a block diagonal matrix consisting of J lots of 2×2 blocks, each block corresponds to one cluster (i.e., for the j-th block this corresponds to the terms for σ_j and ξ_j); this feature enables efficient computation of B.

Using this adjusted likelihood has a number of key properties: not changing the maximum likelihood estimate relative to using likelihood $L_{\mathrm{M}}^{\mathrm{ind}}$, suitably inflating the variances to be consistent with (10) when there is spatial dependence, and leaving the inference unchanged from using likelihood $L_{\mathrm{M}}^{\mathrm{ind}}$ in the absence of spatial dependence. We therefore take our likelihood contribution for the marginal parameters to be $L_{\mathrm{M}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right) = L_{\mathrm{M}}^{\mathrm{adj}}\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)$.

4.2.2 Dependence component

With $\chi_{k,k'}(v)$ defined as in expression (3), we assume that there exists a value $0 \leq \tilde{v} < 1$ such that $\chi_{k,k'}(v) = \chi_{k,k'}$, i.e., it is equal to its limit form, for all $\tilde{v} < v < 1$ and for all pairs of sites (k, k'). Techniques for the selection of $\tilde{v}$, for a pair (k, k'), are available (de Haan

and de Ronde, 1998; Wan and Davis, 2019). Our data for estimating $\chi_{k,k'}$ are derived from exceedances of the $100\tilde{v}\%$ quantiles at sites k and k', as determined by their respective empirical distribution estimates $\hat{F}_k$ and $\hat{F}_{k'}$. First let $\mathcal{Q}_{k,k'} = \left\{ t : \hat{F}_{k'}(r_{k',t}) > \tilde{v} \right\}$, be the set of times when there is both an exceedance of the quantile threshold at site k' and the data for site k are available, and let $Q_{k,k'} = |\mathcal{Q}_{k,k'}|$ denote the cardinality of this set. Further let

$$P_{k,k'} \;=\; \# \left\{ t \in \mathcal{Q}_{k,k'} \,:\, \hat{F}_k(r_{k,t}) > \tilde{v} \right\}, \qquad k, k' = 1, \ldots, K, \tag{13}$$

be the number of times when both site k and site k' simultaneously observed an extreme event. Then it follows that $P_{k,k'}$ is distributed as

$$P_{k,k'} \mid \chi_{k,k'} \;\sim\; \text{Binomial}\,(\, Q_{k,k'} \,,\, \chi_{k,k'} \,). \tag{14}$$

By combining (14) and the information about $\chi_{k,k'}$, from our cluster model (7), and then integrating over $\chi_{k,k'}$, we obtain that $P_{k,k'} \mid (\mathcal{D}, \mathbf{Z})$ follows a beta-binomial distribution of the form

$$P_{k,k'} \mid (\mathcal{D}, \mathbf{Z}) \;\sim\; \begin{cases} \text{Beta-binomial}\left(Q_{k,k'} \,,\, \frac{\beta}{\exp\left(\gamma_j d_{k,k'}\right)-1} \,,\, \beta \right) & \text{if} \quad Z_k = Z_{k'} = j, \\[2ex] \text{Beta-binomial}\left(Q_{k,k'} \,,\, \frac{\beta}{\exp\left(\gamma_0 d_{k,k'}\right)-1} \,,\, \beta \right) & \text{if} \quad Z_k \neq Z_{k'}. \end{cases} \tag{15}$$

Following the discussion in Section 3.3 this model is assumed to hold only for pairs of sites (k, k') which are adjacent. We let $k \sim k'$ denote the distinct pairs of sites (k, k') which are adjacent. Denote the density function of the beta-binomial distribution in (15) by g, then for a pair of adjacent sites (k, k') the likelihood contribution to $L_{\mathrm{D}}\left(\boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)$ is

$$L_{\mathrm{D}}^{k,k'} = \left[g\left(P_{k,k'} \mid Q_{k,k'}, \mathbf{Z}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \right) \times g\left(P_{k',k} \mid Q_{k',k}, \mathbf{Z}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \right) \right]^{0.5},$$

as critically we have two estimates for $\chi_{k,k'}$ which contain almost exactly the same information, each of equal value, and so we weight both observations by 0.5.

Under an assumption of independence of distinct pairs, the likelihood function for the spatial dependence model is then

$$L_{\mathrm{D}}\left(\boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right) \;=\; \prod_{k \sim k'} L_{\mathrm{D}}^{k,k'}. \tag{16}$$

This likelihood is misspecified since $\{P_{k,k'} : k \neq k'\}$ are not independently distributed; for instance, when there are no missing data, if $P_{k,k'}$ and $P_{k,k''}$ are very large, then $P_{k',k''}$ cannot be small. We could use the Ribatet et al. (2012) adjustment again but we opted against it due to the following reasons. Firstly, in our data examples, the values for

$P_{k,k'}$ are not too close to $Q_{k,k'}$, so dependence between $P_{k,k'}$ and $P_{k,k''}$ will not be strong. Secondly, we are not directly interested in the marginal posterior distributions of $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$ and as for our posterior, given $\mathbf{Z}$ and $\mathcal{D}$, $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$ are independent. Finally, the observed information matrix which we would require for the curvature adjustment is dense due to the parameter β being present in all terms of the likelihood function and of dimension $(J+2) \times (J+2)$. Therefore, we would have to invert a potentially high-dimensional matrix, which leads to a substantial computational cost since we have to compute this matrix many times.

4.3 Priors

To perform Bayesian inference, we specify priors for the number of clusters J, the cluster labels $\mathbf{Z} \mid J$, and the model parameters $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$. Since $J \geq 1$, we define $J - 1 \sim$ Poisson(κ) and set a weakly informative Gamma prior for κ, $\kappa \sim$ Gamma$(1, 0.001)$.

We wish to impose that clusters are contiguous; otherwise, two sites which are far apart may be grouped together, despite the probability of them jointly facing an extreme event being close to zero. Our prior is similar to Knorr-Held and Raßer (2000) and only gives positive mass to contiguous clusters. The idea is to represent the spatial structure of the J clusters via a set of centres $\mathbf{C}^{(J)} = (C_1, \ldots, C_J) \in \{1, \ldots, K\}$, $C_i \neq C_j$ if $i \neq j$; $C_i = j$ corresponds to $\mathbf{s}_j$ being the centre of the i-th cluster. Each site k is assigned to the closest cluster centre in terms of $d_{k,k'}$ $(k, k' = 1, \ldots, K)$, i.e., we take

$$Z_k \mid \mathbf{C}^{(J)} = \underset{j=1,\ldots,J}{\arg\min} \ d_{k,C_j}. \tag{17}$$

To ensure that $\mathbf{Z} \mid \mathbf{C}^{(J)}$ is well-defined, the site k is allocated to the cluster with lowest index if multiple cluster centres have minimum distance to the site. Relationship (17) implies that we can assign a prior for $\mathbf{Z} \mid J$ via one for $\mathbf{C}^{(J)}$. We choose a uniform prior with

$$\mathbb{P}\left(\mathbf{C}^{(J)} \mid J\right) = \frac{(K - J)!}{K!}.$$

We conclude by assigning priors for the parameters $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$. For the GPD parameters in (2), a log-normal prior is defined for the scale parameter, $\sigma_j \sim$ Lognormal$(\mu^\sigma, \theta^\sigma)$, while a normal prior is set for the shape parameter, $\xi_j \sim$ Normal(μ^ξ, θ^ξ) $(j = 1, \ldots, J)$. The priors for the parameters describing the spatial dependence are defined as exponentially distributed. Specifically, $\epsilon_j \sim$ Exponential(θ^ϵ), $\gamma_0 \sim$ Exponential(0.001) and $\beta \sim$ Exponential(0.01). The exponential prior for ϵ_j represents a prior preference to small spatial differences in the extremal dependence. To complete the model setup, we specify independent conjugate priors for the hyperparameters: $\mu^\sigma \sim$ Normal$(0, 1)$, $\mu^\xi \sim$ Normal$(0, 0.2)$, $\theta^\sigma, \theta^\xi \sim$ Inverse-Gamma$(1, 0.1)$ and $\theta^\epsilon \sim$ Gamma$(5, 2)$.

4.4 Implementation

We wish to sample from the posterior distribution defined by the likelihood function (8) and the prior distributions in Section 4.3. Since the dimension of the parameter space changes with the number of clusters, J, we use a reversible jump MCMC algorithm (RJMCMC) (Green, 1995). Given a current sample with J clusters, we propose one of the following seven moves:

Birth: Introduce a new cluster centre C^* with parameters ϵ^*, σ^* and ξ^*.

Death: Remove one of the existing cluster centres $C_1, \ldots, C_J$.

Shift: Move one of the cluster centres to an adjacent site which is not a cluster centre.

Sigma: Update the scale parameters $\boldsymbol{\sigma}^{(J)}$ of the GPD in (2).

Xi: Update the shape parameters $\boldsymbol{\xi}^{(J)}$ of the GPD in (2).

Chi: Update the dependence parameters $\boldsymbol{\epsilon}^{(J)}$, γ_0 and β in (15).

Hyper: Update the hyperparameters κ, (μ^σ, μ^ξ) and $(\theta^\sigma, \theta^\xi, \theta^\epsilon)$.

The birth and death moves are comparable to the split and merge moves defined in other Bayesian clustering approaches, see Bottolo et al. (2003) for instance, but they potentially affect more than one cluster. For the examples in Sections 5 and 6, birth, death and shift are proposed with probability 0.2 each while the remaining four moves are each proposed with probability 0.1.

We briefly describe some features of our implementation and more details are provided in Appendix A. For a birth move, the new cluster centre C^* is uniformly sampled from the $K - J$ sites which are not currently cluster centres; the proposal is rejected immediately if $K = J$. In addition to C^*, the index at which to insert C^* in the vector $\mathbf{C}^J$ is sampled with equal probability. The proposal distributions for the cluster parameters ϵ^*, σ^* and ξ^* are close to the priors in Section 4.3 but some spatial information is incorporated. A death move ensures reversibility and one of the existing J cluster centres is removed with equal probability. If $J = 1$, the death move is rejected immediately. A shift move reallocates an existing cluster centre to one of the adjacent sites which are not currently cluster centres, without changing the cluster parameters and the indexing of the cluster centres. Note, the matrix B in expression (11) has to be updated in the case of a birth, death or shift move; due to B being a block diagonal matrix, we only have to update the 2×2 blocks of the clusters which are affected by the proposal. The model parameters in $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$ are updated via a Metropolis-within-Gibbs algorithm, while we sample from the corresponding full conditional distributions when updating the hyperparameters.

4.5 Analysis of the posterior samples

Suppose that we generated N posterior samples using the RJMCMC algorithm in Section 4.4. From these samples, we want to estimate the underlying cluster structure, the marginal distributions of the K sites and the measure $\chi_{k,k'}$ $(k, k' = 1, \ldots, K)$.

We employ Bayesian decision theory to derive a point estimate $\mathbf{Z}^*$ for the cluster structure. Following Wade and Ghahramani (2018), we derive $\mathbf{Z}^*$ based on minimizing the posterior expected loss,

$$\mathbf{Z}^* = \arg\min_{\hat{\mathbf{Z}}} \sum_{\mathbf{Z}} \left[\text{VI}(\mathbf{Z}, \hat{\mathbf{Z}}) \times \pi(\mathbf{Z} \mid \mathcal{D}) \right], \tag{18}$$

where the sum is over all possible clusterings, $\pi(\mathbf{Z} \mid \mathcal{D})$ is the posterior probability of the clustering $\mathbf{Z}$, and $\text{VI}(\mathbf{Z}, \mathbf{Z}^*)$ is the variation of information loss by Meilă (2007), which compares the information in $\mathbf{Z}$ and $\mathbf{Z}^*$ with the information shared between the two clusterings. The variation of information loss consists of three terms, the first two are the entropies of the clusterings $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, and the final term measures the mutual information. We apply the `mclust.ext` R package (Wade, 2015) to estimate $\mathbf{Z}^*$ based on our posterior samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(N)}$.

We estimate the marginal distribution of the extremes for site k $(k = 1, \ldots, K)$ by averaging over the sampled GPDs, so the estimated tails of R_k do not necessarily follow a GPD. The i-th sampled scale and shape parameters for site k, $\psi_k^{(i)}$ and $\nu_k^{(i)}$, are given by the sampled cluster-specific scale and shape parameters, $\sigma_{Z_k^{(i)}}^{(i)}$ and $\xi_{Z_k^{(i)}}^{(i)}$, where $Z_k^{(i)}$ is the cluster label for site k in the i-th sample $(i = 1, \ldots, n)$. In addition to the GPD parameters, we characterize the marginal extremal behaviour using return levels. The τ-year return level is the value which is exceeded on average once every τ years. For the sample $\left\{ \psi_k^{(i)}, \nu_k^{(i)} \right\}$, the τ-year return level of R_k is given by

$$u_k + \frac{\psi_k^{(i)}}{\nu_k^{(i)}} \left[(\lambda_u \tau)^{\nu_k^{(i)}} - 1 \right], \tag{19}$$

where λ_u is the expected number of times R_k exceeds u_k per year. We fix λ_u to its empirical estimate, and we estimate the τ-year return level for site k by averaging over the N τ-year return levels associated to the sampled GPD parameters.

Similarly to the marginal distributions, we derive estimates for the pairwise extremal dependence measure. Given the posterior samples for γ_0, $\boldsymbol{\epsilon}$ and β, we average over the beta distributions in (7) to obtain posterior estimates of $\chi_{k,k'}$ for adjacent sites k and k' $(k, k' = 1, \ldots, K)$. To assess model fit, we compare the posterior mean estimate of $\chi_{k,k'}$ to its empirical estimate. In the absence of knowledge of our model in Section 3.3 for $\chi_{k,k'}$

over sites, the empirical estimate for $\chi_{k,k'}$ is given by

$$\hat{\chi}_{k,k'} \;=\; \frac{P_{k,k'}}{Q_{k,k'}},$$

which, if there were no missing data, is simply the proportion of the exceedances of the $100\tilde{v}\%$ for site k' that also exceed this quantile threshold for site k.

For illustration and comparison purposes, we also consider an alternative approach which imposes a spatial cluster structure and assumes that clusters are fixed to the point estimate $\mathbf{Z}^*$ in (18). This produces estimates that replicate many regional methods which ignore cluster uncertainty. The posterior distribution is then given by the likelihood $L_{\mathrm{M}}^{\mathrm{adj}}(\cdot \mid \mathcal{D}, \mathbf{Z}^*)$ in (11) and the priors in Section 4.3. Posterior samples of the GPD parameters are generated using only the moves Sigma, Chi and Hyper in Section 4.4. Thus, the sampled GPD parameters for sites k and k' are identical if these sites belong to the same cluster according to $\mathbf{Z}^*$. Posterior return level estimates are again derived by averaging over the τ-year return levels (19) associated to the sampled scale and shape parameters. We refer to this approach as the fixed cluster (FC) approach and we apply it later in Section 6.

5   Simulation Examples

Via three simulation studies, we study our approach's performance to infer the number of clusters, the spatial cluster structure and the marginal GPD parameters. Figure 1 left panel highlights the $K = 55$ areal units considered in each of these studies. The distance $d_{k,k'}$ $(k, k' = 1, \ldots, 55)$ is computed using the coordinates of the municipalities' centroids and by accounting for the earth's curvature. Distances are standardized such that $0 \leq d_{k,k'} \leq 1$.

In each study, we specify the number of clusters, J, and the cluster-specific parameters, $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$, and then generate 100 data sets using a two-step process. In the first step, we sample the J cluster centres uniformly amongst the 55 areal units and allocate the remaining sites according to rule (17). Next, observations for the marginal distributions and the extremal dependence are simulated independently. The marginal distributions are set to $R_{k,t} \sim \mathrm{GPD}(\psi_k, \nu_k)$ $(k = 1, \ldots, 55;\ t = 1, \ldots, T)$ and all sites within a cluster have the same marginal GPD parameters given by $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$. Since the marginal distribution follows a GPD, we set $u_k = 0$ in (1); in a practical analysis, we usually require a higher threshold which we select using the graphical diagnostic tools mentioned in Section 1. We simulate $T = 100$ observations per site from the marginal GPD. For the extremal dependence, we fix $Q_{k,k'} = 25$ $(k, k' = 1, \ldots, 55)$ and sample $P_{k,k'}$ from the beta-binomial distribution (14) with the specified parameters $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$.

Posteriors for the spatial clusters and model parameters herein are based on running

the RJMCMC algorithm in Section 4.4 for 2×10^6 iterations with the initial number of clusters set to $J = 10$; the first 5×10^5 iterations are discarded as burn-in and then every 50-th sample is stored for analysis. Examples of the sampled Markov chains for J and $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ are provided in the Supplementary Material.

Studies 1 and 2 consider the case of all sites having the same marginal distribution and common pairwise extremal dependence structure, i.e., $J = 1$. The studies differ in that the data are independent (strongly dependent) in Study 1 (Study 2). By investigating the case $J = 1$, we check whether our approach tends to the parsimonious estimate $J = 1$ or the extreme case $J = 55$, i.e., each site forms its own cluster. Study 3 then considers the case of $J = 5$ clusters and with spatially dependent data.

Study 1: The marginal GPD parameters are $\psi_k = 2$ and $\nu_k = 0.2$ $(k = 1, \ldots, 55)$, and the spatial dependence parameters are $\gamma_1 = 2$ and $\beta = 10$; we do not specify γ_0 since $J = 1$. The true spatial cluster structure is recovered for 99 of the 100 data sets; for the remaining data set, the estimated number of clusters is $J = 2$. Across the 99 data sets for which we identify the correct clustering, the posterior probability of $J = 1$ ranges from 0.6 up to almost 1. Regarding the marginal GPD parameters, the true values of ψ_k and ν_k lie within the central 90% credible interval for 93 and 87 of the data sets; the average lengths of the credible intervals for ψ_1 and ν_1 are 0.15 and 0.06 respectively.

Study 2: For the data as generated in Study 1, we induce strong spatial dependence by matching ranks across sites, i.e., the m-th highest observation $(m = 1, \ldots, 100)$ occurs simultaneously at all sites. The true cluster structure is recovered for 97 of the 100 data sets. Furthermore, the posterior median is $J = 1$ in 95 cases, and the posterior probability of $J = 1$ ranges from 0.5 to 0.97 for these 95 data sets. Regarding the marginal GPD parameters, the true values of ψ_k and ν_k always lie within the central 90% credible interval, and the average lengths of the credible intervals for ψ_1 and ν_1 are 1.0 and 0.35. The credible intervals are wider than in Study 1, as expected given there is less information in the pooled data than in Study 1 due to the strong dependence. Furthermore, the width of the credible intervals is very similar to the ones when estimating ψ_1 and ν_1 solely based on the observations for site $k = 1$ in Study 1; so we obtain almost no additional information from pooling due to the strong spatial dependence, apart from the centre of the credible interval being closer to the true value. The latter feature arises as our method pools information across the cluster in contrast to when only data from site 1 is used to estimate (ψ_1, ν_1).

Study 3: The cluster-specific marginal GPD parameters are $\boldsymbol{\sigma} = (1.5, 1.75, 2.0, 2.25, 2.5)$ and $\boldsymbol{\xi} = (0.1, 0.15, 0.2, 0.25, 0.3)$. As in Study 2, we induce spatial dependence in the marginal observations. Given the underlying clustering $\mathbf{Z}$, we first sample from a multivariate normal distribution, with stronger dependence for sites in the same cluster than
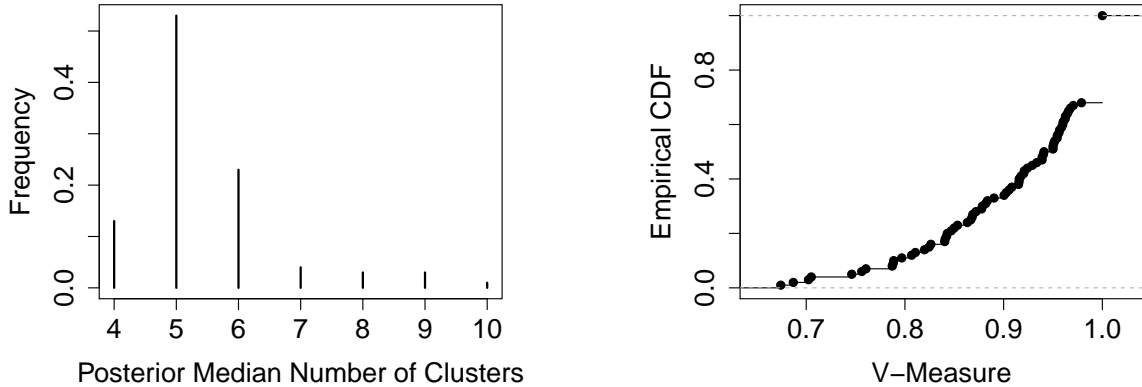
Figure 3: Histogram of the estimated posterior median number of clusters for the 100 data sets in Study 3 (left). Empirical cumulative distribution function of the V-measure for the 100 data sets of Study 3 (right).

sites in different clusters, and we then apply the probability integral transformation to obtain marginal observations which are GPD with the given scale and shape parameters. The value $P_{k,k'}$ $(k, k' = 1, \ldots, 55)$ is sampled from the beta-binomial distribution (14) with parameters $\gamma_0 = 2$, $\gamma_j = 1$ $(j = 1, \ldots, 5)$ and $\beta = 10$.

We first analyze the sampling distribution of the estimated posterior median number of clusters over the 100 data sets. Figure 3 left panel shows that the estimated posterior median of J is usually close to the truth, with being $J = 5$ in more than half of the replications. When analyzing the posterior samples for ψ_k and ν_k, the true marginal GPD parameters lie within the estimated 90% central credible interval for most sites.

Next, we derive the point estimate $\mathbf{Z}^*$ for each of the 100 data sets and we apply the V-measure (Rosenberg and Hirschberg, 2007) to quantify the similarity between the true cluster structure and $\mathbf{Z}^*$ numerically. The V-measure takes values between 0 and 1, with higher values indicating higher resemblance between the true and estimated clusters, and a value of 1 corresponding to matching clusterings. Figure 3 right panel shows the sampling distribution of the V-measure, illustrating that our point estimate $\mathbf{Z}^*$ performs well in terms of the V-measure; all values are within 0.35 of 1 and we correctly identify the spatial cluster structure for 32 of the 100 data sets. We also find that the true clustering is visited in 69 of the 100 sampled Markov chains for $\mathbf{Z}$. Spatial plots illustrating examples of correctly identified clusterings, and the true and estimated clusterings with the three lowest V-measures are provided in the Supplementary Material.

6 Data analysis

We now apply our methodology to analyze the data described in Section 2. After an appropriate burn-in period, we perform 1.5×10^7 iterations of the RJMCMC sampling
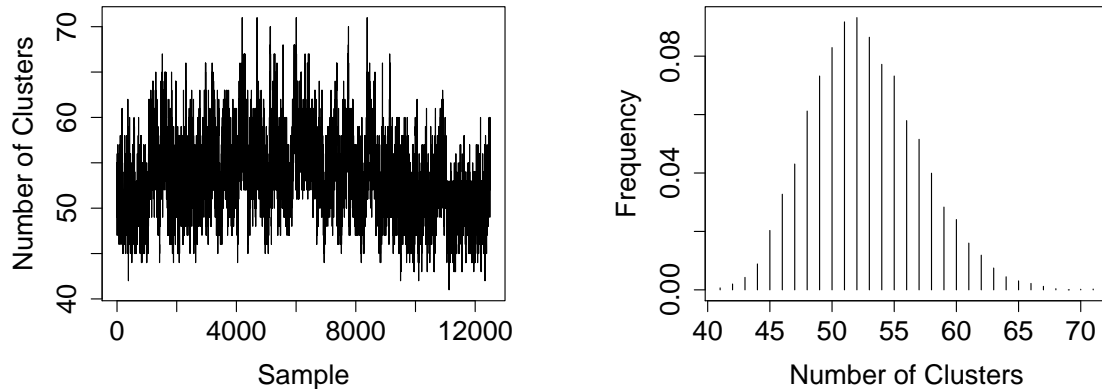
Figure 4: Trace plot (left) and posterior mass function (right) of the number J of clusters for the precipitation data in Section 6.1.

scheme in Section 4.4, with every 1000-th sample being stored. Initially, 10% of the sites are cluster centres. The acceptance probabilities for birth (and death) were 0.03 and 0.05 for the data in Sections 6.1 and 6.2 respectively. Our C++ implementation took about 60 minutes per 10^6 iterations on a standard laptop for the larger data set in Section 6.1.

6.1 Daily precipitation in South Norway

We analyse the extremal behaviour of the daily amount of precipitation $R_{k,t}$ ($k = 1, \ldots, 343$; $t = 1, \ldots, 3651$) across the 343 municipalities constituting South Norway. Analysis of the threshold stability plots suggests setting u_k as the empirical 97% quantile; this gives about 100 peaks-over thresholds per municipality. The empirical measures $P_{k,k'}$ and $Q_{k,k'}$ ($k, k' = 1, \ldots, 343$) are derived based on the dependence threshold $\tilde{v} = 0.99$.

Figure 4 left panel indicates appropriate convergence and mixing of the sampled Markov chain for the number of clusters, J, and Figure 4 right panel shows that the posterior distribution of J is unimodal with mean $J = 52$ and central 80% credible interval $(47, 58)$.

We then derive the estimate $\mathbf{Z}^*$ for the cluster structure as described in Section 4.5. The estimate $\mathbf{Z}^*$ in Figure 1 right panel comprises 43 clusters; the largest clusters contain about 20 municipalities, and several clusters are made up of five or less municipalities. We can see that the point estimate $\mathbf{Z}^*$ proposes a lower number of clusters than most of the posterior samples, thus suggesting a more parsimonious model. The derived spatial clusters agree with known climatology. Clusters along the west coast regularly observe very high amounts of precipitation which are often related to the Gulf Stream. We also find that the drier municipalities in central Norway are grouped together.

Next, we analyze the posterior samples for the remaining model parameters and we find very good mixing for all; trace plots of the sampled marginal GPD parameters for five municipalities, and the parameters γ_0 and β are provided in the Supplementary Material.
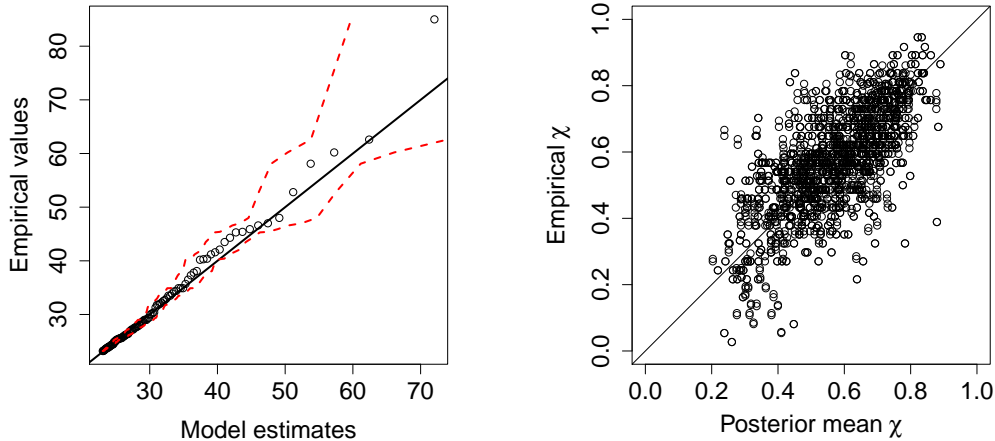
20

Figure 5: QQ plot for the peaks-over threshold for Stavanger (left) and posterior mean estimates of $\chi_{k,k'}$ versus their empirical estimate $\hat{\chi}_{k,k'}$ for all adjacent pairs of municipalities (right).

The posterior samples for the site-specific GPD scale parameter, ψ_k, show strong spatial variation, similar to the average amount of precipitation illustrated in the Supplementary Material, while we find little spatial variation in ν_k; spatial plots of the posterior mean estimates for ψ_k and ν_k are provided in the Supplementary Material. Figure 5 left panel indicates good model fit of the marginal observations for one of the municipalities; QQ plots for four other municipalities are provided in the Supplementary Material.

Figure 5 right panel shows that our model estimates capture the core variations in pairwise extremal dependence across short spatial distances. The agreement is far from perfect but it must be remembered that the empirical estimates are themselves highly uncertain. One aspect where the fit is least good is for the pairs of sites with the empirical estimates of $\chi_{k,k'}$ close to 0. These sites lie in central Norway and they are separated by the Scandinavian Mountains, which leads to lower spatial dependence, given their separation. Since we assumed that γ_0 is common across space, we overestimate the extremal dependence for these pairs of sites.

We conclude by estimating return levels for the five municipalities highlighted in Figure 1: Fredrikstad (F), Sarpsborg (A), Lillehammer (L), Stavanger (S) and Trondheim (T); spatial maps for the estimated return levels across South Norway are provided in the Supplementary Material. The selected municipalities cover different climates and belong to medium to large sized clusters in Figure 1 right panel, with Fredrikstad and Sarpsborg being in the same cluster. We consider three approaches to estimate return levels: (i) estimates obtained individually using only the observed peaks-over threshold of the municipality, (ii) our method and (iii) the FC approach described in Section 4.5.

Table 1 shows that the spatial pooling methods provide shorter credible intervals for the considered municipalities than method (i); the central 90% credible interval for method (i) generally contains the credible intervals for methods (ii) and (iii). This increased preci-

τ	Method	Fredrikstad	Sarpsborg	Lillehammer	Stavanger	Trondheim
25	(i)	51 (43,63)	57 (47,71)	61 (49,81)	97 (66,150)	67 (52,91)
	(ii)	57 (51,64)	57 (51,64)	61 (53,70)	86 (70,109)	67 (56,80)
	(iii)	58 (52,64)	58 (52,65)	60 (53,68)	97 (71,133)	64 (55,76)
100	(i)	60 (47,80)	68 (52,93)	75 (56,109)	135 (76,247)	85 (60,129)
	(ii)	67 (58,79)	68 (59,79)	75 (63,92)	113 (83,156)	88 (70,112)
	(iii)	69 (60,80)	69 (60,80)	74 (63,89)	135 (86,214)	84 (69,105)

Table 1: Posterior mean (central 90% credible interval) of the τ-year return level in mm for five municipalities with $\tau = (25, 100)$. The considered methods are (i) Individual estimation for each municipality (ii) our approach and (iii) the FC approach described in Section 4.5.

sion of estimates shows the benefits of spatial pooling. The three approaches give almost identical return level point estimates for Sarpsborg, Lillehammer and Trondheim which shows that in these cases our clustering approach groups neighbouring municipalities with very similar characteristics. Fredikstad is on the coast, with Sarpsborg being an adjacent municipality with similar climate and topography. For Fredrikstad, methods (ii) and (iii) broadly agree, and produce similar values to Sarpsborg, but give different estimates than method (i) without giving estimates outside that associated 90% credible interval. In contrast, for Stavanger methods (i) and (iii) give similar estimates with method (ii) giving lower estimates, but with all estimates falling inside the other methods' associated 90% credible intervals. These features indicate that our cluster allocation method gives a point estimate which groups municipalities with similar marginal tail behaviour, but that accounting for cluster uncertainty gives rise to the inclusion of clusters with shorter GPD tails.

6.2 Daily river flow in the UK

The data in Section 2.2 exhibit a strong seasonal pattern for all $K = 45$ gauges. We consider separately the maximum weekly river flow for November-March and May-September for which in each case the assumption of stationarity seems reasonable. The observations in each season are standardized site-wise to mean 0 and variance 1; while this affects the GPD scale parameter ψ_k, it is well known that this leaves the shape parameter ν_k unchanged. A common threshold across all gauges, $u_1 = u_2 \cdots = u_K$, is selected individually for the two seasons, giving between 25 and 40 peaks-over threshold per season for most sites. The threshold $\tilde{v}$ in (13) is set to $\tilde{v} = 0.965$ for November-March and $\tilde{v} = 0.95$ for May-September. This gives $Q_{k,k'} \approx 35$ (25) for most pairs of sites (k, k') $(k, k' = 1, \ldots, 45)$ over the summer (winter) season. Here, the distance $d_{k,k'}$ between sites (k, k') is set to
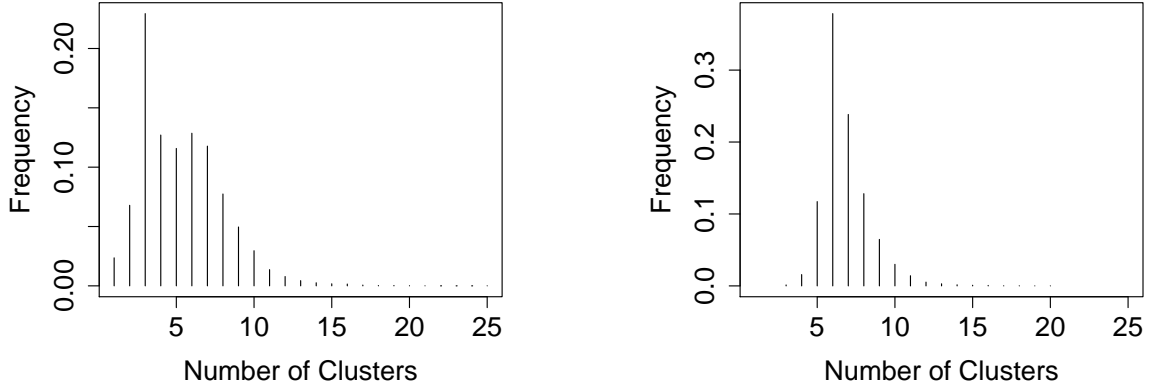
Figure 6: Posterior mass functions of the number of clusters, J for the time periods November-March (left) and May-September (right).
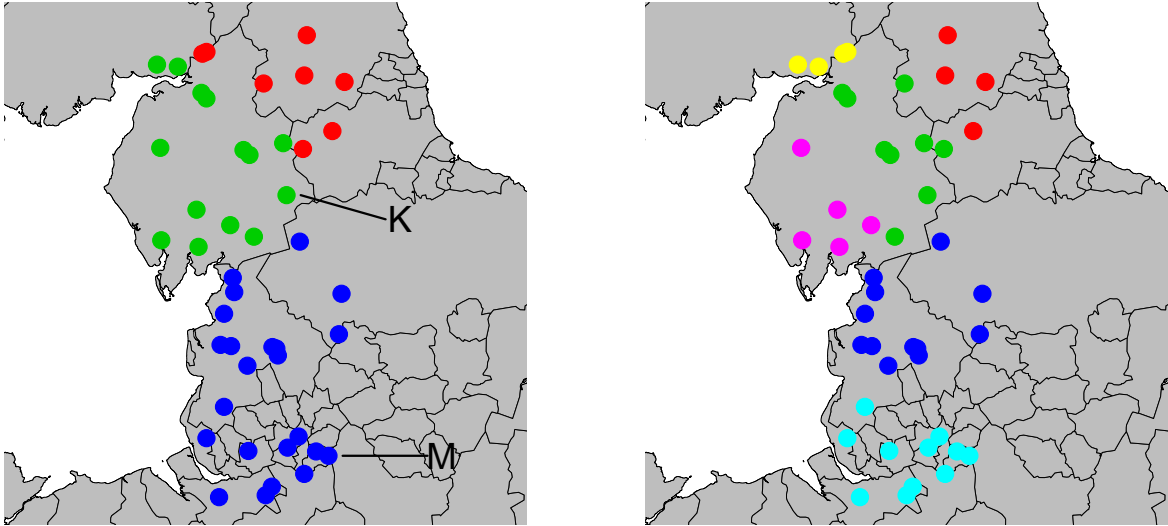

Figure 7: Point estimates for the underlying spatial cluster structure for November-March (left) and May-September (right). The solid black lines are the boundaries of the metropolitan and non-metropolitan counties, and the highlighted gauges are Kirkby Stephen (K) and Marple Bridge (M).

their hydrological distance (Section 2) which we scale such that $0 \leq d_{k,k'} \leq 1$.

Figure 6 indicates a higher number of clusters for May-September than for November-March. This result agrees with known climatology. Extreme river flows in winter in the UK are often caused by extratropical cyclones which affect larger areas; we thus expect larger clusters for November-March than for May-September. Trace plots for J and the GPD parameters, and QQ plots for the two gauges highlighted in Figure 7 are provided in the Supplementary Material.

We then derive the point estimate $\mathbf{Z}^*$ in (18) for the two seasons. Here, we focus on the results with respect to the geographical coordinates of the gauges. Hydrological maps illustrating the hydrological distances between gauges, the adjacency structure applied and

23

τ	Model	Kirkby Stephen		Marple Bridge	
		Winter	Summer	Winter	Summer
100	(i)	112 (85,159)	47 (36,66)	73 (55, 104)	35 (26,50)
	(ii)	101 (83,130)	46 (37,57)	75 (56, 105)	36 (31,42)
	(iii)	112 (84,152)	49 (41,58)	61 (54, 70)	36 (32,41)
500	(i)	152 (98, 258)	61 (40,102)	105 (66,179)	49 (31,84)
	(ii)	137 (101,200)	60 (44,82)	104 (69,178)	49 (39,64)
	(iii)	166 (107,265)	64 (51,84)	79 (66,97)	48 (38,62)

Table 2: Posterior mean (central 90% credible interval) of the τ-year return level in m^3/s for two gauges with $\tau = (100, 500)$. The considered methods are (i) Individual estimation for each municipality (ii) our approach and (iii) the FC approach described in Section 4.5.

the clusterings with respect to hydrological coordinates are provided in the Supplementary Materials. Figure 7 shows an interesting match between the point estimates for the two seasons. The southern cluster for November-March splits into two equally-sized clusters in May-September. Similarly, the other large cluster (green) for the winter season splits into two clusters in summer. The estimated clusters are not identical to the river networks. For instance, we have five gauges for the River Tyne in North-East England but only four of these are allocated to the same cluster (the red cluster for the summer season). Finally, we note that some clusters are split along administrative boundaries, in particular, in winter, which is not surprising as these boundaries were determined by topography.

To conclude, we estimate τ-year return levels for $\tau = (100, 500)$ for the two gauges highlighted in Figure 7 left panel: Kirkby Stephen (K) and Marple Bridge (M). Kirkby Stephen was selected because multiple of its adjacent sites are allocated to different clusters, while Marple Bridge lies centrally in the southern cluster in winter and at the edge of the most southern cluster in the summer season. We consider the same inference approaches as in Section 6.1.

Table 2 shows that the three methods produce similar return level estimates for the period May-September, which shows that our clustering approach indeed groups gauges with very similar characteristics. As in Section 6.1, methods (ii) and (iii) give narrower credible intervals than model (i), again showing the benefits of spatial pooling. When analyzing the estimates for Kirby Stephen for November-March, we see that method (ii) produces different estimates than methods (i) and (iii). Similar to Stavanger in Table 1, this indicates that our point estimate groups gauges with similar characteristics, but that accounting for cluster uncertainty gives rise to the inclusion of clusters with shorter GPD tails. For Marple Bridge and November-March, we see strong differences in Table 2, in particular, for methods (i) and (iii). This indicates that our point estimate groups too many gauges in this case, but we have to remember that estimates are highly uncertain given the strong spatial correlation and small amount of data.

7 Discussion

We introduced a Bayesian clustering approach which groups geographical sites based on their marginal tail behaviour and their extremal dependence. The likelihood for the peaks-over threshold accounts for the spatial dependence usually found in hydrological applications. Our model for the extremal dependence postulates that sites within the same cluster exhibit higher pairwise extremal dependence than sites in different clusters, and that the degree of dependence decreases with an increasing distance between sites. Clusters are represented by their centre, which imposes them to be contiguous and leads to a good computational performance. Samples from the posterior distribution were obtained using a reversible jump MCMC algorithm. A point-estimate for the spatial cluster structure was derived from the pairwise posterior probabilities of sites being in the same cluster using Bayesian decision theory.

We applied our approach to analyze precipitation levels across South Norway and the derived clusters agree with climatology. The cluster approach was further applied to river flow data in the UK, and we found that clusters are not identical to the river networks and that the spatial extent of extreme river flow levels is larger over winter than summer. The results also showed that our approach efficiently pools spatial information to improve return level estimates.

Throughout the paper we have assumed that the same physical extreme events occur simultaneously at different sites and that each extreme event is only present at a single time point after declustering. In practice there can be temporal lags between the same event at different sites. Our methodology can be adapted to handle such situations provided that we have a method for identifying these time lags, see for example Tawn and Ledford (2003) and Davis and Mikosch (2009). We can then account for these lags both in the derivation of $P_{k,k'}$ in (14) and the estimation of $V(\boldsymbol{\theta}_{\mathrm{M}}^{(J)})$ in (10).

One of the referees highlighted that a random design for $Z_k \mid \mathbf{C}^{(J)}$ may be preferred to the deterministic design in (17); this would indeed yield a larger number of possible sampled clusterings. Then, $Z_1, \ldots, Z_K$ are resampled in each iteration for each of the proposed moves in Section 4.2.2. However, the likelihood in (8) would then have to be evaluated fully, while the deterministic approach (17) implies that the clustering changes only locally, reducing the computational cost. Since our main aim is to present a computationally efficient approach, we opted for the deterministic definition of $Z_k \mid \mathbf{C}^{(J)}$.

There are various ways to extend our cluster approach. Firstly, we focused on modelling threshold exceedances, but we could have worked with an extremal mixture model (Behrens et al., 2004; MacDonald et al., 2011) comprising separate distributions for observations below and above the threshold u if the full marginal distribution was of interest. Secondly, we model extremal dependence of a pair of adjacent sites (k, k') in different clusters via the single parameter γ_0. Instead of γ_0 being constant in (5), γ_0 may be defined as

a function of the cluster-specific parameters. Consider a pair of adjacent sites (k, k') with $Z_k = j$ and $Z_{k'} = j'$, $j \neq j'$. High values of γ_j and $\gamma_{j'}$ may then imply a high value for γ_0. Another possible extension is the consideration of temporal variations in the distribution of the peaks-over threshold and/or the extremal dependence. Such an extension should then also allow for a potential change of the spatial cluster structure across seasons; the results for the UK river flow data indicate the presence of such a temporal variation in the number of clusters.

Acknowledgement

A   Details of the reversible jump MCMC algorithm

Birth and death

Suppose there are currently J clusters with centres $\mathbf{C}^{(J)} \subseteq \{1, \ldots, K\}$ and parameters $\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \right)$. In a birth, we first uniformly sample a new cluster centre C^* together with the index j^* at which to insert it in the set $\mathbf{C}^{(J)}$; the probabilities are $\mathbb{P}\left(C^* = k \mid \mathbf{C}^{(J)} \right) = (K - J)^{-1}$ $(k \in \{1, \ldots, K\} \setminus \mathbf{C}^{(J)})$ and $\mathbb{P}(j^* \mid \mathbf{C}^{(J)}) = (J + 1)^{-1}$ $(j^* = 1, \ldots, J + 1)$. For the new set of cluster centres, $\mathbf{C}^* = (C_1, \ldots, C_{j^*}^*, \ldots, C_{J+1})$, we then derive the cluster labels $\tilde{\mathbf{Z}}$ according to (17).

To complete the proposal, we sample the parameters $(\epsilon^*, \sigma^*, \xi^*)$ of the new cluster. The mean of each proposal distribution is set to the current average value of that parameter across the sites allocated to the new cluster, while the variance of the proposal is set to the one of the corresponding prior. Let $\mathcal{C} = \left\{ k : \tilde{Z}_k = j^* \right\}$ denote the sites allocated to the new cluster. The proposal ξ^* is sampled from a normal distribution with

$$\xi^* \mid \left(\boldsymbol{\xi}^{(J)}, \mathbf{Z}, \tilde{\mathbf{Z}} \right) \sim \text{Normal}\left(\frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} \xi_{Z_k}, \, \theta^{\xi} \right).$$

The parameter σ^* is sampled from a log-normal distribution with

$$\mathbb{E}\left(\sigma^*\right) \;=\; \frac{1}{|\mathcal{C}|}\sum_{k\in\mathcal{C}}\sigma_{Z_k} \qquad \text{and} \qquad \mathrm{Var}\left(\sigma^*\right) \;=\; [\exp(\theta^\sigma)-1]\exp(2\mu^\sigma+\theta^\sigma).$$

The proposal varies ϵ^* with respect to the current number J of clusters. If $J > 1$, ϵ^* is sampled with

$$\epsilon \;\sim\; \mathrm{Exponential}\left(\frac{|\mathcal{C}|}{\sum_{k\in\mathcal{C}}\epsilon_{Z_k}}\right).$$

In case $J = 1$, the model for $P_{k,k'}$ $(k, k' = 1, \ldots, K)$ is fully described by the parameters γ_1 and β. If a second cluster is proposed, we have to sample two proposals to derive $(\gamma_0, \gamma_1, \gamma_2)$, and in order to satisfy the dimension matching condition (Green, 1995). We sample ϵ_1^* and ϵ_2^* independently from an Exponential(θ^ϵ) and define γ_0^* as

$$\gamma_0^* \;=\; \gamma_1\exp\left(\epsilon_1^*\right), \quad \gamma_1^* \;=\; \gamma_1, \quad \gamma_2^* \;=\; \gamma_1\exp\left(\epsilon_1^* - \epsilon_2^*\right).$$

We compute the acceptance probability as given in Green (1995). The determinant of the Jacobian is equal to 1, unless $J = 1$, since the defined mapping is the identity function. For $J = 1$, the determinant of the Jacobian is $\exp(\epsilon_1^*)$. For the reverse move, death, we select one of the existing cluster centres with equal probability. The acceptance probability for the case $J > 1$ is then

$$\min\left\{1\,,\; \frac{L\left(\boldsymbol{\theta}_{\mathrm{M}}^*, \boldsymbol{\theta}_{\mathrm{D}}^* \mid \mathcal{D}, \tilde{\mathbf{Z}}\right)}{L\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)} \times \frac{\pi(\sigma^*, \xi^*, \epsilon^*)}{q(\sigma^*, \xi^*, \epsilon^*, \mid \boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)}, \mathbf{Z}, \tilde{\mathbf{Z}})} \times \frac{\kappa}{J} \times \frac{p_D}{p_B}\right\},$$

where p_D and p_B are the probabilities for proposing a death and birth move, respectively, π is the joint prior density and q is the joint proposal density.

In case of a death, we first update the cluster labels to obtain $\tilde{\mathbf{Z}}$ for the proposed set of $J - 1$ cluster centres. Let $(\sigma_*, \xi_*, \epsilon_*)$ denote the parameters of the cluster which is proposed to be removed, and $\boldsymbol{\theta}_{\mathrm{M}}^* = \boldsymbol{\theta}_{\mathrm{M}}^{(J)} \setminus (\sigma_*, \xi_*)$ and $\boldsymbol{\theta}_{\mathrm{D}}^* = \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \setminus (\epsilon_*)$. The acceptance probability for the case $J > 2$ is then

$$\min\left\{1\,,\; \frac{L\left(\boldsymbol{\theta}_{\mathrm{M}}^*, \boldsymbol{\theta}_{\mathrm{D}}^* \mid \mathcal{D}, \tilde{\mathbf{Z}}\right)}{L\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)} \times \frac{q(\sigma_*, \xi_*, \epsilon_* \mid \boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)}, \tilde{\mathbf{Z}}, \mathbf{Z})}{\pi(\sigma_*, \xi_*, \epsilon_*)} \times \frac{J-1}{\kappa} \times \frac{p_B}{p_D}\right\}.$$

Remaining moves

In case of a shift move, we first select one of the current cluster centres with equal probability and propose to reallocate it to one of the adjacent sites which is not currently a cluster centre. Let j^* be the index of the sampled cluster centre. We first derive the set $\mathcal{N}$ of potential sites and then select one with equal probability as the new cluster

centre C^*. This reallocation of a cluster centre usually changes the spatial clusters and we thus derive the set of updated cluster labels $\tilde{\mathbf{Z}}$ via (17). To calculate the acceptance probability, we also require the set $\mathcal{N}^*$ of sites adjacent to C^* which are not cluster centres; this set includes the current cluster centre C_{j^*}. Since $\boldsymbol{\theta}_{\mathrm{M}}^{(J)}$ and $\boldsymbol{\theta}_{\mathrm{D}}^{(J)}$ do not change, the prior densities cancel and the acceptance probability is

$$\min \left\{ 1 \, , \, \frac{L\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \tilde{\mathbf{Z}}\right)}{L\left(\boldsymbol{\theta}_{\mathrm{M}}^{(J)}, \boldsymbol{\theta}_{\mathrm{D}}^{(J)} \mid \mathcal{D}, \mathbf{Z}\right)} \times \frac{|\mathcal{N}|}{|\mathcal{N}^*|} \right\}.$$

We specified three moves to update $\boldsymbol{\theta}_{\mathrm{M}}^*$ and $\boldsymbol{\theta}_{\mathrm{D}}^*$. Since the parameters for cluster j and j', $j \neq j'$, are independent given $\mathbf{Z}$, each cluster parameter is updated separately via an independence sampler; the proposal distribution is equal to the prior in this case and the acceptance probability is thus equal to the likelihood ratio. The last move updates the hyperparameters by drawing from the full conditional distributions using Gibbs sampling.

References

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050.

Asadi, P., Engelke, S., and Davison, A. C. (2018). Optimal regionalization of extreme value distributions for flood estimation. *Journal of Hydrology*, 556:182–193.

Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T. (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. *Weather and Climate Extremes*, 9:17–24.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons Chichester.

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, 26(20):7929–7937.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1725.

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6(1):25–47.

Carreau, J., Naveau, P., and Neppel, L. (2017). Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resources Research*, 53(5):4407–4426.

Casson, E. and Coles, S. G. (1999). Spatial regression models for extremes. *Extremes*, 1(4):449–468.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):207–222.

Chavez-Demoulin, V., Embrechts, P., and Sardy, S. (2014). Extreme-quantile tracking for financial time series. *Journal of Econometrics*, 181(1):44–52.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London.

Coles, S. G., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. G. and Tawn, J. A. (1990). Statistics of coastal flood prevention. *Philosophical Transactions of the Royal Society: Physical and Engineering Sciences (1990-1995)*, 332(1627):457–476.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):463–478.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840.

Davis, R. A. and Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

de Haan, L. and de Ronde, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes*, 1(1):7–45.

Dombry, C. and Ribatet, M. (2015). Functional regular variations, Pareto processes and peaks over threshold. *Statistics and Its Interface*, 8:9–17.

Dupuis, D. J. and Tawn, J. A. (2001). Effects of mis-specification in bivariate extreme value problems. *Extremes*, 4(4):315–330.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):25–45.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, To appear.

Environment Agency (2018). Estimating the economic costs of the winter floods 2015 to 2016. Ref: LIT 10736, https://www.gov.uk/government/publications/floods-of-winter-2015-to-2016-estimating-the-costs.

Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737.

Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):545–556.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C. and Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas. *The Annals of Statistics*, 37(5B):2990–3022.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Haug, O., Dimakos, X. K., Vårdal, J. F., Aldrin, M., and Meze-Hausken, E. (2011). Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, 2011(1):1–20.

Hilal, S., Poon, S.-H., and Tawn, J. A. (2014). Portfolio risk assessment using multivariate extreme value methods. *Extremes*, 17(4):531–556.

Hosking, J. R. M. and Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(2):271–281.

Institute of Hydrology (Great Britain) (1975). *Flood studies report*. Natural Environment Research Council, London.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley Hoboken, N.J.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69(1):19–27.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

Meilă, M. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Mornet, A., Opitz, T., Luzi, M., Loisel, S., and Bailleul, B. (2017). Wind storm risk management: sensitivity of return period calculations and spread on the territory. *Stochastic Environmental Research and Risk Assessment*, 31(8):1977–1995.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1):93–120.

Reich, B. J. and Shaby, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, 6(4):1430–1451.

Reich, B. J. and Shaby, B. A. (2019). A spatial Markov model for climate extremes. *Journal of Computational and Graphical Statistics*, 28:117–126.

Reich, B. J., Shaby, B. A., and Cooley, D. (2014). A hierarchical model for serially-dependent extremes: A study of heat waves in the western US. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(1):119–135.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Rohrbeck, C., Eastoe, E. F., Frigessi, A., and Tawn, J. A. (2018). Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics*, 12(1):246–282.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117 – 131.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Rubio, R., de Carvalho, M., and Huser, R. G. (2018). Similarity-based clustering of extreme losses from the London stock exchange. *Unpublished manuscript*.

Sang, H. and Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2013). A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100.

Sharkey, P. and Winter, H. C. (2019). A Bayesian spatial hierarchical model for extreme precipitation in Great Britain. *Environmetrics*, 30(1):e2529.

Smith, R. L. and Goodman, D. J. (2000). Bayesian risk analysis. In Embrechts, P., editor, *Extremes and Integrated Risk Management*, chapter 17, pages 235–251. Risk Books, London.

Tawn, J. and Ledford, A. (2003). Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):521–543.

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415.

Vettori, S., Huser, R., and Genton, M. G. (2019). Bayesian modeling of air pollution extremes using nested multivariate max-stable processes. *Biometrics*, 75(3):831–841.

Wade, S. (2015). mcclust.ext: Point estimation and credible balls for Bayesian cluster analysis. https://www.researchgate.net/publication/279848500_mcclustext-manual.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wadsworth, J. L. and Tawn, J. A. (2012). Dependence modelling for spatial extremes. *Biometrika*, 99(2):253–272.

Wadsworth, J. L. and Tawn, J. A. (2019). Higher-dimensional spatial extremes via single-site conditioning. *Submitted*.

Wan, P. and Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166.

Wang, Y. and So, M. K. (2016). A Bayesian hierarchical model for spatial extremes with multiple durations. *Computational Statistics & Data Analysis*, 95:39–56.