The importance of context in extreme value analysis with application to extreme temperatures in the USA and Greenland

Daniel Clarkson

Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom.

E-mail: d.clarkson@lancaster.ac.uk

Emma Eastoe

Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom.

Amber Leeson

Data Science Institute, Lancaster University, Lancaster, United Kingdom.

Summary. Statistical extreme value models allow estimation of the frequency, magnitude and spatio-temporal extent of extreme temperature events in the presence of climate change. Unfortunately, the assumptions of many standard methods are not valid for complex environmental data sets, with a realistic statistical model requiring appropriate incorporation of scientific context. We examine two case studies in which the application of routine extreme value methods result in inappropriate models and inaccurate predictions. In the first scenario, record-breaking temperatures experienced in the US in the summer of 2021 are found to exceed the maximum feasible temperature predicted from a standard extreme value analysis of pre-2021 data. Incorporating random effects into the standard methods accounts for additional variability in the model parameters, reflecting shifts in unobserved climatic drivers and permitting greater accuracy in return period prediction. The second scenario examines ice surface temperatures in Greenland. The temperature distribution is found to have a poorly-defined upper tail, with a spike in observations just below 0°C and an unexpectedly large number of measurements above this value. A Gaussian mixture model fit to the full range of measurements is found to improve fit and predictive abilities in the upper tail when compared to traditional extreme value methods.

1. Introduction

The roots of the statistical modelling of extreme events lie in the theoretical work of Von Mises (1936) and Gnedenko (1943) who sought to understand the asymptotic behaviour of sample maxima. From the 1950s, this probability theory began to be translated into the statistical modelling framework now known as the block maxima method (Gumbel, 1958). Both theory and modelling framework were later extended to a Peaks Over Threshold (PoT) approach (Pickands, 1975; Davison and Smith, 1990). It was recognised early on that both block maxima and PoT frameworks could make major contributions to the study of natural hazards, leading to a long-standing relationship with the environmental sciences and civil engineering. Specific application areas include hydrology (Katz et al., 2002; Jonathan and Ewans, 2013), air pollution (Eastoe and Tawn, 2009; Reich et al., 2013; Gouldsborough et al., 2021), temperatures (Acero et al., 2014; Winter et al., 2016), precipitation (Katz, 1999), drought (Burke et al., 2010), wind speeds (Fawcett and Walshaw, 2006), statistical downscaling (Friederichs, 2010; Towe et al., 2017), space weather (Thomson et al., 2011; Rogers et al., 2020) and climate change problems (Sterl et al., 2008; Cooley, 2009; Cheng et al., 2014).

Both theory and models have evolved considerably since the mid-twentieth century. In particular, there have been massive advances in methodology for multivariate extreme events (Heffernan and Tawn, 2004; Ramos and Ledford, 2009; Rootzén et al., 2018) and the related areas of spatial (Cooley et al., 2012; Huser and Wadsworth, 2019), temporal (Ledford and Tawn, 2003; Winter and Tawn, 2017) and spatio-temporal (Economou et al., 2014; Simpson and Wadsworth, 2021) extreme value modelling. These developments are in part a response to the increasing availability of large spatio-temporal environmental data sets obtained, for example, from satellite images, high resolution process-based models and online measurement tools.

Nevertheless the continued development of methods for univariate extremes remains a crucial research topic, not least due to the absence of a unified best practice methodology for modelling univariate extremes in the presence of complex trends, such as those seen in the presence of climate change. There is also still no single standard to inform the quantification and communication of risk in the presence of trends, whether or not this is climate change driven. Whilst such a standard may ultimately prove to be unobtainable, the commonly employed approach of reporting a single risk measure to cover all scenarios, regardless of long-term trends or other changes, can, and should, be improved upon. Lastly, specification of a model that describes the marginal model well is an essential first step in the very large number of copula-based multivariate extreme value models (Joe, 1994).

This paper examines extreme temperature data for two regions with contrasting climates and very different geophysical features. The first case study assesses the heatwave experienced in the western United States during the summer of 2021 to ascertain whether the record-breaking temperatures from this event are consistent with historical temperature extremes. This study was motivated by a recent analysis of ERA5 reanalysis model output (Philip et al., 2021) which used a statistical extreme value model fitted to pre-2021 temperature extremes to estimate the upper bound of the temperature distribution. They found that the 2021 observations exceeded this upper bound at multiple locations. We propose some adaptations to their analysis to better investigate if this is the case

and, if so, to understand why. The second case study analyses spatio-temporal records of ice surface temperature for the Greenland ice sheet, motivated by the need to find an appropriate univariate model for the measurements in each cell as the first stage of a spatial extreme value model. Given the large number of cells (1139) the marginal model should ideally be designed to permit automatic fitting, however the characteristics of the upper tail of the temperature distribution differ considerably across the ice sheet and so careful consideration is required to achieve this objective.

A feature common to both case studies is that the tail behaviour of the univariate distribution of interest cannot be described using out-of-the-box extreme value methods. Instead, by applying a pragmatic, contextualised and data-driven approach, much improved models can be developed whilst avoiding the need to incorporate huge numbers of parameters or highly complex structures. This approach has several advantages. Firstly, since the models are computationally both cheap and stable, analysis of large data sets is entirely feasible. Secondly, simple models are less prone to errors in fitting and interpretation, making them particularly useful for those without statistical training. Lastly, a simple model results in straightforward estimates of risk which makes it easier for the end-user to put these estimates to use.

2. Extreme Value Analysis

This section provides background on univariate extreme value modelling and introduces the random effects model that is subsequently used in Section 3. Univariate extreme value methods encompass models for both block maxima and PoT data sets. Central to this paper is the PoT approach which characterises the tail behaviour of the underlying process by modelling only those observations in the data set that lie above a pre-selected high threshold. In contrast, block maxima methods describe the distribution of the maxima of pre-defined and non-overlapping blocks; for environmental studies this block is usually a year. Each method has a theoretically-motivated probability distribution at its core: the generalised extreme value distribution (block maxima) and the generalised Pareto distribution (PoT). Whilst theoretical justification of these distributions is beyond the scope of the paper, note that, as with the central limit theorem, the justification relies on asymptotic arguments. For finite sample sizes the distributions therefore only approximate the tail behaviour and are not guaranteed to describe the data exactly.

2.1. Vanilla block maxima and PoT models

The theory on which the vanilla versions of both block maxima and PoT models are based starts with the assumption of a sequence $Y_1, \ldots, Y_m$ of independent replicates of a random variable Y, with distribution $F(y) = \Pr[Y \leq y]$. Under some non-restrictive conditions on F, Von Mises (1936) and Gnedenko (1943) showed that in the limit as $m \to \infty$, the distribution of the normalised block maximum $(M_m - b_m)/a_m = (\max\{Y_1, \ldots, Y_m\} - b_m)/a_m$ is one of the three types of max-stable distribution: Weibull, Fréchet or Gumbel. Whilst both the rate of convergence to the limiting distribution and the sequences $\{a_m > 0\}$ and $\{b_m\}$ are determined by the underlying distribution F which, in a modelling context is unknown, this result is nevertheless used to justify the

use of the Generalised Extreme-value Distribution (GEV) to model the maxima of large but finite blocks.

The GEV is a class of distributions combining the three types of max-stable distribution. The resulting cumulative distribution function is,

$$G(z) = \exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]^{-1/\xi}\right\}, \quad -\infty < z < y^+$$

where $y^+ < \infty$ in the case $\xi < 0$. This distribution has location μ, scale $\sigma > 0$ and shape ξ parameters, with the Weibull, Fréchet and Gumbel distributions corresponding to $\xi < 0$, $\xi > 0$ and $\xi \to 0$ respectively. As a model the GEV has the advantage of increased flexibility over choosing one of the three types a priori.

The vanilla PoT model is motivated by a result of Pickands (1975) who showed that, under some non-restrictive regularity conditions, as $u \to y^+$ for $y > u$,

$$F(y) = 1 - \Pr[Y > u]\Pr[Y > y | Y > u] \approx 1 - \phi\bar{G}(y) = 1 - \phi\left[1+\frac{\xi}{\psi}(y-u)\right]_+^{-1/\xi} \quad (1)$$

where $a_+ = \max\{a, 0\}$, $0 \leq \phi \leq 1$, $\psi > 0$ and $\bar{G}(y)$ is the survival function of the generalised Pareto distribution. The parameters ϕ, ψ and ξ are known as the rate, scale and shape respectively; the rate $\phi = \Pr[Y > u]$ is sometimes called the exceedance probability. The value of the shape parameter is determined by the rate at which $F(y) \to 1$ as $y \to y^+$; exponential decay corresponds to $\xi = 0$, and heavy (light) decay to $\xi > 0$ ($\xi < 0$). For $\xi \geq 0$ there is no finite end-point, whilst $y^+ = u - \sigma/\xi < \infty$ if $\xi < 0$.

By assuming that this limit result is a suitable approximation to the behaviour of the exceedances of a finite, but high, level u, it can be used to motivate a model in which the magnitudes of the threshold exceedances $\{y_i - u : y_i > u, i = 1, \ldots, n\}$ are a random sample from the generalised Pareto distribution and the exceedance times a random sample from a homogeneous Poisson process. Whilst not essential an assumption of independence is often made about the exceedances. If there is evidence of serial dependence, a routine and theoretically-justified approach is to identify all clusters of extremes and model only the frequency and magnitude of the cluster maxima (Smith and Weissman, 1994; Laurini and Tawn, 2003; Drees, 2008; Davis et al., 2013). In either case, the parameters (ϕ, ψ, ξ) are estimated with standard statistical inference procedures e.g maximum likelihood or Bayesian inference.

A crucial modelling consideration is the choice of threshold u, which requires a compromise between retaining a sufficiently large number of exceedances such that estimation of the model parameters is possible, whilst still ensuring that the assumption of the limiting approximation is plausible. Various tools are available to aid threshold selection. Most of these use two key threshold stability properties of the generalised Pareto distribution to identify the minimum plausible threshold above which the distribution is a suitable description of the data.

2.2. Random Effects PoT Model

The most common extension of the vanilla PoT model is to model one or more of the distribution parameters (ϕ, ψ, ξ) as a function of covariates. The motivation for this is

that most physical processes display changes in their behaviour over time. Such changes may be cyclic, smooth or sudden and often arise in response to changes in one or more of climate, weather conditions or other geophysical processes. Whilst many commonly used regression modelling frameworks, both parametric (Smith, 1989; Davison and Smith, 1990; Eastoe and Tawn, 2009) and semi-parametric (Hall and Tajvidi, 2000; Yee and Stephenson, 2007; Jones et al., 2016; Youngman, 2019), have been incorporated into the PoT framework, there are several barriers that prevent their successful implementation. Firstly, since most environmental processes are driven by multiple interacting factors, appropriate drivers will often be either unknown or unobserved, or both. Secondly, PoT data sets are small by definition and there is frequently insufficient information to identify the often subtle covariate relationships. Lastly, whilst the motivating extreme value theory provides a justification to extrapolate into the underlying tail distribution given values of the covariates, there is no similar justification to extrapolate the covariate-response relationship beyond the measurement period. This is a serious limitation for climate change forecasts.

A plausible alternative is the PoT random effects model which overcomes the above limitations for cases where changes in extremal behaviour occur between, but not within, years. Previously applied to hydrological data by Eastoe (2019) and Towe et al. (2020), inter-year variability in both the magnitudes and frequency of the extremes is modelled by latent variables (or 'random effects') which vary between years. The advantage of this framework is that, unlike a parametric regression, it requires no prior specification of the functional form for the inter-year variability. In this sense, the model is a compromise between the rigid fully parametric regression model and the flexible non-parametric regression. It can be used to identify long term trends, unusually extreme years, and groupings of years in which the extremal behaviour is similar. The model also has the capability to make predictions about extreme episodes over a period of time (e.g. a number of decades) but not to do so for a specific out-of-sample year. Although it is beyond the scope of the current paper, such predictions could be achieved by incorporating serial dependence in the random effects.

For generality, we define a threshold u_i for year i. This allows for cases in which it is appropriate for the threshold, and hence the definition of an extreme observation, to vary between years. Let Y_{ij} be the jth observation in year i then, for $y > u_i$,

$$\Pr[Y_{ij} \leq y | Y_{ij} > u_i] = 1 - \phi_i \left[1 + \frac{\xi_i}{\psi_i}(y - u_i)\right]_+^{-1/\xi_i} \tag{2}$$

where

$$\psi_i \sim F_\psi(\psi_i|\theta_\psi), \quad \xi_i \sim F_\xi(\xi_i|\theta_\xi), \quad \phi_i \sim F_\phi(\phi_i|\theta_\phi) \tag{3}$$

and F_ψ, F_ξ and F_ϕ are pre-specified distributions with parameters θ_ψ, θ_ξ and θ_ϕ respectively. In hierarchical modelling terminology, equations (2) and (3) are data and latent process models respectively, and (ψ_i, ξ_i, ϕ_i) are random effects for year i. We assume throughout that the responses Y_{ij} are independent given the random effects and that the random effects are mutually independent and serially uncorrelated.

Let y denote the full vector of data, N be the number of years in the observation period and n_i be the number of observations in each year. Denote the vectors of random

effects by $\psi = (\psi_1, \ldots, \psi_N)$, $\xi = (\xi_1, \ldots, \xi_N)$, $\phi = (\phi_1, \ldots, \phi_N)$. The starting point for inference is the joint likelihood function

$$L(y|\psi, \xi, \phi, \theta_\psi, \theta_\xi, \theta_\phi) = L(y|\psi, \xi, \phi) f(\psi|\theta_\psi) f(\xi|\theta_\xi) f(\phi|\theta_\phi)$$

where

$$L(y|\psi, \xi, \phi) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left\{ \phi_i \psi_i^{-1} \left[1 + \frac{\xi_i}{\psi_i}(y_{ij} - u_i) \right]_+^{-1/\xi_i - 1} \right\}^{I[y_{ij} > u_i]} (1 - \phi_i)^{I[y_{ij} \leq u_i]} \quad (4)$$

and $f(\psi|\theta_\psi)$, $f(\xi|\theta_\xi)$ and $f(\phi|\theta_\phi)$ are each the product of the relevant distribution taken from equation (3), for example $f(\psi|\theta_\psi) = \prod_{i=1}^{N} f_\psi(\psi_i|\theta_\psi)$ where f_ψ is the density associated with F_ψ. Note that the dimension of the full parameter vector is $3N + |\theta_\psi| + |\theta_\xi| + |\theta_\phi|$.

Since the random effects cannot be integrated out of the joint likelihood function, maximum likelihood parameter estimates must be obtained through numerical optimisation of the very high-dimensional log-likelihood function. A more pragmatic approach is to use Bayesian inference; to do so, prior distributions $\pi_\psi(\psi)$, $\pi_\xi(\xi)$ and $\pi_\phi(\phi)$ must now be specified for θ_ψ, θ_ξ and θ_ϕ respectively. We recommend that these are chosen to be as uninformative as possible, e.g. flat Gaussian distributions, in order to avoid undue impact on the analysis. Combining the prior distributions with the likelihood function (4), the joint posterior is

$$\pi(\theta_\psi, \theta_\xi, \theta_\phi, \psi, \xi, \phi) = L(y|\psi, \xi, \phi) f(\psi|\theta_\psi) f(\xi|\theta_\xi) f(\phi|\theta_\phi) \pi_\psi(\psi) \pi_\xi(\xi) \pi_\phi(\phi).$$

Since closed forms cannot be found for the joint and marginal posterior distributions, the posterior distribution is estimated using a sampling algorithm. We use an adaptive Markov chain Monte Carlo (MCMC) algorithm to enable automatic tuning of the proposal step-sizes and achieve fast convergence. Our algorithm is based on the adaptive Metropolis-Within-Gibbs scheme proposed by Roberts and Rosenthal (2009). Each parameter is updated separately via a Metropolis-Hastings Random Walk (MHRW) step, allowing separate evolution of the MHRW scaling parameter associated with each of the model parameters. Earlier work by Eastoe (2019) showed that of a number of possible alternatives, this algorithm worked the best for this particular model.

3. Case study 1: US air temperatures

This case study is motivated by the 2021 heatwave which affected three western states in the United States. An analysis of climate model annual maxima by Philip et al. (2021) showed that the observed 2021 annual maximum were inconsistent with historical extremes. Further, this finding recurred across a number of stations and held despite adjustments made to the data to account for long-term trends in global temperature. Our objectives are to examine whether (a) an analysis of observational annual maxima is consistent with the ERA5 analysis of Philip et al. (2021), (b) an analysis of PoT data points to the same conclusions as an analysis of annual maxima, and (c) a more flexible model for long term trends captures better the 2021 event.

The three states most affected by the heatwave were Oregon, Washington and British Columbia. For these states, we obtain maximum daily temperatures from Version 3 of the

Global Historical Climatology Network (Menne et al., 2012). Stations are included only if they came online in or before 1950, were still online at the time of the 2021 heatwave, and have at most 20% of observations missing. This leaves the majority of stations in Oregon (62) and Washington (61). Whilst there are fewer in British Columbia (17), the majority are situated in the region most affected by the heatwave, i.e. the south of the state close to the Washington border. The heatwave epicentre is clearly shown in Figure 1. Areas including Portland and The Dalles recorded some of the highest temperatures, with the highest observation across all stations being 47.8°C at The Dalles Municipal Airport (lat 45.62, long -121.16) on 28/06/21. This date appears to correspond to the peak of the heatwave, with station-wide average temperature of 39.6°C. Whilst temperatures at coastal stations are lower than at the epicentre, extremity is defined relative to local climate norms and so measurements from these stations are included in the analysis when the temperature exceeds the station specific threshold.

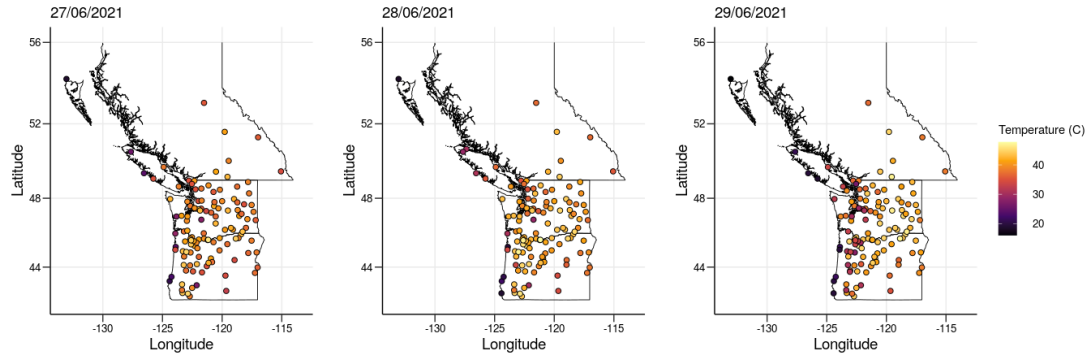

Fig. 1: Maximum temperature observations for weather stations in Oregon (South), Washington (Centre), British Columbia (North) on 27/06/21, 28/06/21, and 29/06/21. Stations with missing value(s) on these days are omitted.

3.1. Annual maxima analysis

We start by considering annual maxima. These are adjusted by removal of the 4-year rolling average of Global Mean Surface Temperature (GMST) to account for long term climate changes and ensure that the maxima are representative of local climate conditions. The GEV is fitted to the adjusted maxima using maximum likelihood estimation. Following Philip et al. (2021), we fit to pre-2021 data only and use the resulting model to assess the likelihood of the 2021 temperatures being predicted prior to the heatwave.

From the GEV model fits, it is immediately clear why there is concern: 41% of the stations have a 2021 annual maximum which exceeds the finite upper end point implied by the GEV model fitted to pre-2021 data, i.e. these 2021 observations are apparently infeasible. Even more concerning is that this occurs despite the adjustments for GMST to account for global climate variation - without this adjustment, 48% of stations exhibit

the same behaviour. Note that these are relatively coarse statistics - the upper limit can only be estimated at locations for which $\xi < 0$ and the uncertainty in the estimated limit has not been taken into consideration - however they suggest that climate variation, at least as measured by GMST, does not fully explain the variation in annual maxima, and indeed may play only a small role. Further, annual maxima only capture the temperature on a single day and are not therefore capable of informing a broader understanding of shifts in frequency and magnitude of extreme heatwaves. Such events are likely to be influenced by many more factors than simply the overall climate trends of the region. Capturing all such factors is made difficult by the complexity of meteorological and climate processes and incomplete recording of relevant physical variables.

3.2. Peaks Over Threshold analysis

Progressing to the PoT model, the 4-year rolling average of GMST is subtracted from the station-specific 97.5% quantile to give year- and station-specific thresholds. This results in between 419 and 639 pre-2021 threshold exceedances at each station - considerably more data than the 71 annual maxima. To assess the extremity of the 2021 event under the PoT model fitted to pre-2021 data, Figure 2 shows predicted return periods associated with the 2021 annual maxima. The estimated period is finite at all stations, with a maximum of 559 years (Estevan Point, British Columbia). The return period is greater than 5 (10) years for 24.2% (8.3%) of the stations. Thus the PoT analysis suggests that the 2021 maximum temperatures are extreme, but not as extreme as the GEV fit implies. However, while the return periods determine the rarity of a single observation, they do not quantify whether the heatwave itself was exceptionally extreme.

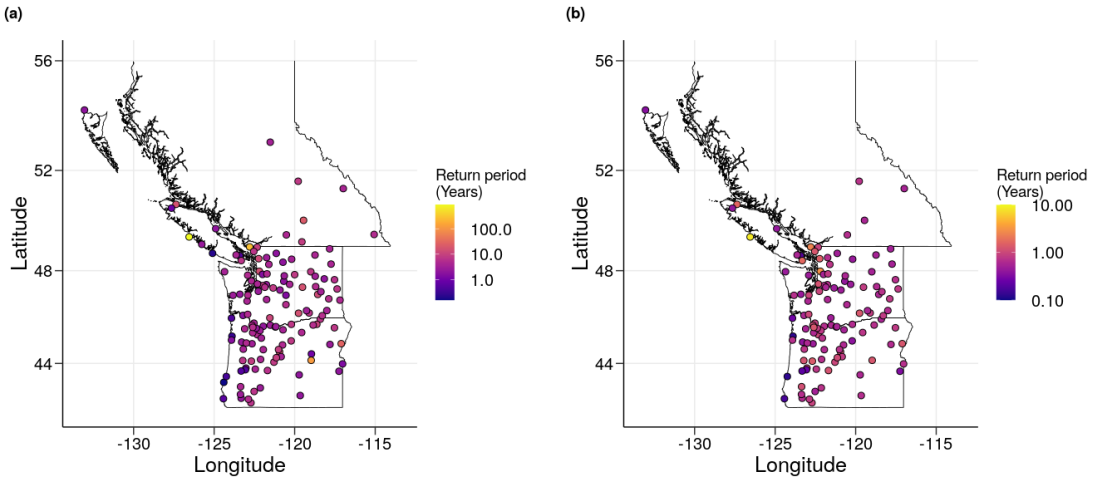

Fig. 2: (a) Return periods of the 2021 maxima at each weather station using the fitted Generalised Pareto Distributions (GPDs), and (b) Return periods of the 2021 maxima at each weather station using the fitted random effects PoT model. 16 stations are not included as their 2021 maxima are below their 97.5% quantile.

The PoT random effects model uses the same thresholds as the vanilla PoT model. As discussed earlier, the model cannot predict the extremal behaviour in a specific year and therefore, in contrast to the vanilla fit, we include the 2021 data when estimating the parameters. A simplified version of the model described in equations (2) and (3) was used with only the scale parameter allowed to vary between years. The justification for a constant shape parameter is that estimation of a varying shape parameter requires so much information that it is common practice to model it as either constant or piece-wise constant. A varying rate parameter was found to be unnecessary due to the year-specific threshold. Each site was modelled separately.

To ensure that the scale parameter estimates are within the parameter space, we use a log link function:

$$\tau_i = \log \psi_i \sim \text{Normal}(\theta_{\psi,0}, \theta_{\psi_1}).$$

We used independent Normal$(0, 20)$ priors for ξ, $\theta_{\psi,0}$ and $\theta_{\psi,1}$. The adaptive MCMC algorithm was run for 100000 iterations at each site with a burn-in of 10000. The output was thinned by retaining only every 50th point in each chain. Trace plots and posterior distributions at a random selection of sites were examined for mixing and convergence. An example of these plots is shown for the shape and scale parameter estimates for the Cedar Lake station in Figure 7 of the Appendix.

Figure 3 shows the estimated scale parameter random effects by station and by year, highlighting years with relatively high extreme temperatures. Note that the inter-year variability in the random effects does not show evidence of a long-term trend, suggesting that the unidentified drivers which cause this variability are unlikely to themselves have long term trends that are substantially different to that seen in GMST. The random effects in a given year are reasonably consistent across stations, with a maximum cross-station standard deviation of 0.28. The overall range is $(-1.301, 1.073)$ with 17 of the top 20 estimates occurring in 2021.

The random effects give a much clearer picture of the relative magnitude of the 2021 heatwave. Considering inter-year variability much reduces the return periods of each 2021 maxima for those that exceed the threshold (Figure 3), and in particular, the cross-station mean of the random effects is highest in 2021 (0.468) compared to the next highest (0.250) in 1961. The highest variance in random effects (0.270) also occurs in 2021, perhaps due to different parts of the region experiencing the heatwave to varying degrees, with extremely high relative temperatures at some stations and more typical temperatures at others, and whilst previous years have isolated random effects at individual stations that reach the same levels as 2021, consistency of the high random effects in 2021 distinguishes it from previous warm years, e.g. 1961.

4. Case Study 2: Greenland ice surface temperatures

One impact of the rise in global temperatures caused by climate change has been an increased volume of melt water generated by the world's ice sheets. This has led to a sea level rise that has already increased the risk of flooding in low-lying areas and will continue to do so whilst temperatures continue to rise. The two largest ice sheets, Greenland and the Antarctic, have had an accelerating contribution to sea level rise since

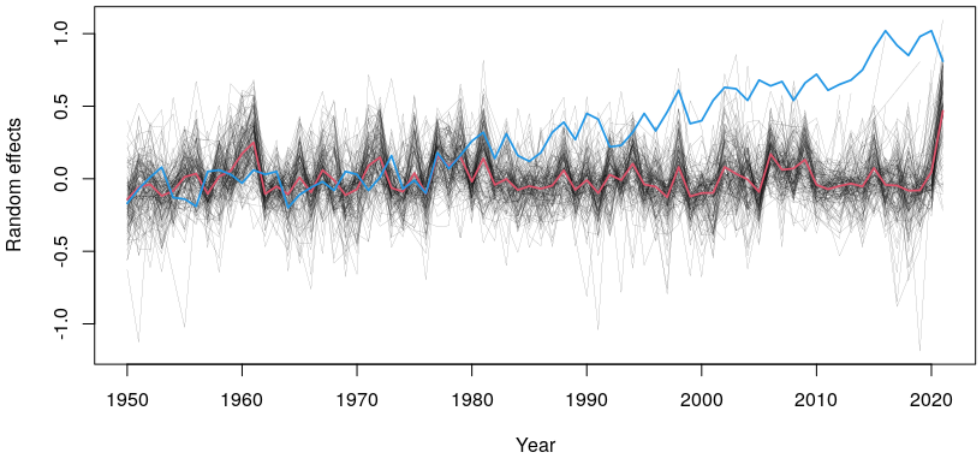

Fig. 3: Estimated scale parameter annual random effects for each weather station (black lines), the annual mean for all stations (red line), and GMST (blue line, in degrees). Results cover the period 1950 to 2021 inclusive.

the early 1990's (Rignot et al., 2018), and are likely to continue to play a key role in the near future (Rahmstorf et al., 2017). Since melt occurs when the ice surface temperature exceeds 0°C, understanding the spatial extent, frequency, magnitude and trends of such events is vital to quantify melt risk both now and in the future, not least because the consequences of increased melt are likely to persist for years to come (Kulp and Strauss, 2019).

Spatial extreme value models provide an ideal tool to analyse the widespread melt events that are of most concern. This case study developed from what was initially intended as a routine first step in fitting such a model to a spatio-temporal Ice Surface Temperature (IST) data set from a multilayer Greenland Moderate Resolution Imaging Spectroradiometer (MODIS)-based product (Hall et al., 2018), the MODIS/Terra Sea Ice Extent 5-Min L2 Swath 1km, Version 6 (MOD29). The spatial resolution of the data is 0.78 km $\times$ 0.78 km over the period 01/01/2001 to 31/12/2019, and the temporal resolution is daily observations. The IST data product uses a cloud mask to filter observations that occur in cloudy conditions since water vapour in the clouds can interfere with infrared radiation leading to measurement inaccuracies. Since cloudy days are generally warmer than clear days (Koenig and Hall, 2010) the data is missing not at random. Consequently we do not attempt imputation of missing values and instead limit the inferences from our model to clear days only.

Since melt occurs when the IST exceeds 0°C, positive temperatures are of most concern. For the majority of cells, particularly inland and/or at higher altitudes, there are too few positive observations to take 0°C as the PoT modelling threshold. However, when taking a modelling threshold below 0°C, there are problems with the PoT fit, as

seen in Figure 4. Many cells have a secondary mode corresponding to a large mass of observations close to 0°C and a much lower-weighted tail covering small positive temperature values. The most likely explanation of this is that once the temperature exceeds 0°C the ice melts. This change in state from ice to melt water, which is no longer considered the surface of the ice sheet, creates a soft upper limit close to 0°C on ISTs. We cannot determine from the data which of the positive observations are water temperatures and which are due to measurement or recording uncertainty (±1°C Hall et al. (2018)). Regardless of the reason, the soft upper limit results in a distribution mode close to 0°C, rendering the generalised Pareto distribution unsuitable. Lastly we note that the shape of the tail distribution, and especially the behaviour at or near 0°C, varies primarily in response to altitude and proximity to the coast.

The objective now is to build a model to reflect these insights and establish how the melt threshold impacts the distribution of ISTs. To facilitate ease of application at all locations we propose a truncated Gaussian mixture model for the full distribution of ISTs. This circumvents the need to define a tail region - as discussed previously, for this particular data set such a definition is non-trivial due to the varying cross-site behaviour of the upper tail. Utilising information on behaviour in sub-tail regions should help better identify behaviour close to the melt threshold. Furthermore, a mixture model allows joint modelling of the ice and melt observations; this is critical since we have no information on which category each observation belongs to.

The truncated Gaussian mixture model has separate model components for ice and melt temperatures. Observations are not pre-assigned to a component, rather the fit provides an estimate of the probability with which each observation belongs to each component. For n_I ice components and a single melt component, let ϕ_i be the weight associated with ice component i and ϕ_M be the weight associated with the melt component, such that $\sum_{i=1}^{n_I} \phi_i + \phi_M = 1$. Let $f_i(x)$ and $f_M(x)$ be the probability density functions of the ice $i \in \{1, \ldots, n_I\}$ and melt components respectively, such that for each i, $X \sim TN(\mu_i, \sigma_i^2, a_i, b_i)$, where μ_i is the mean, $\sigma_i > 0$ the standard deviation, and a_i and b_i the lower and upper truncation points. Then the mixture model probability density function of IST x is:

$$p(x) = \sum_{i=1}^{n_I} \phi_i f_i(x) + \phi_M f_M(x). \tag{5}$$

Ice temperature components are bounded above by $b = 0$, since these should not exceed 0°C, and have no lower bound since the only physical lower limit is absolute zero, which is far below the range of the data. For the melt component, there is no upper bound as ISTs are not feasibly physically limited on the ice sheet. The lower bound is set to -1.65 as this has previously been estimated as the melting point of saline ice and thereby acts as a conservative minimum estimate for the ice sheet. It follows that data in the range $[-1.65, 0]$ can be modelled by either the ice or melt components, capturing the uncertainty in the true nature of the ice sheet when these observations were measured. The probability of melt $\rho(x)$ for given IST x is defined as the ratio of the densities of the ice and melt components, such that:

$$\rho(x) = \frac{f_M(x)}{f_M(x) + \sum_{i=1}^{n_I} f_i(x)}. \tag{6}$$

As a result of the model truncation points, $\rho(x) = 0$ below $-1.65°$C and $\rho(x) = 1$ above $0°$C. ISTs between these values vary smoothly within this range. Small discontinuities in the melt probability may arise at $-1.65°$C and $0°$C due to the truncation of the mixture components at these points; the magnitude of the discontinuity depends on the severity of the truncation. To find the most appropriate number of ice components for the model and allow for the diversity of distributions across the ice sheet, we fit the model with n_I from 3 to 6 and compare the fits using Bayesian Information Criterion (BIC). In contrast, a single melt component is used since, in theory, the melt temperatures should show much less cross-cell variability than the ice temperatures.

We use the Expectation-Maximisation (EM) algorithm to estimate the parameters ϕ, μ and σ^2 for each model component. We found 800 iterations of the algorithm optimal in terms of both convergence and computational time. To improve the stability of the algorithm, particularly for cells with few or no observations above $-1.65°$C, we set additional restrictions on two parameters: $\phi \geq 0.0005$ and $\sigma \geq 0.35$. This allows a melt component - however small - to be fitted even if there are no observations in the melt temperature range, and avoids melt components with $\sigma \approx 0$ since this results in a point-mass density for cells with only a single observation above $-1.65°$C.

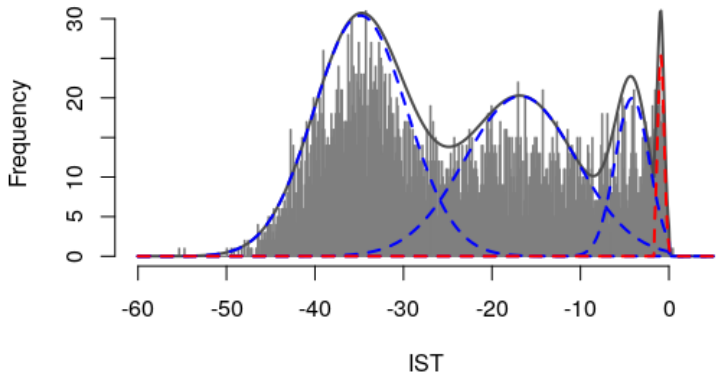

Fig. 4: Example mixture model fit for a single cell (82.47N, -37.50E). Data (grey), ice components densities (blue), melt component density (red), and overall density (black). Density function has been scaled to the maximum value of the histogram data.

We apply the model to 1139 cells across the ice sheet; this represents a subsample of 1 in every 50 available cells in both horizontal and vertical directions. The cells are equally spaced in distance and are taken from those cells identified as ice by the MODIS ice mask. The number and spread of the sub-sampled cells accurately reflects the variety in glaciological and climatological conditions across the ice sheet. A range of average temperatures, surface conditions, latitudes, longitudes, elevations and distances to the coast are all observed within the sample.

The impact of the soft upper limit on ISTs can be seen from the number of ice components selected by the BIC to model the data at different cells. Figure 5 shows the number of ice components in the best fitting model to be higher at cells close to the coastline than those closer to the centre of the ice sheet. There are also trends within coastal areas, with cells on the south west coast generally having the highest number of

model components and areas in the north west coast having fewer. Cells closer to the coast generally have higher average temperatures, experience more melt and therefore have more frequently truncated observations. Furthermore, melt estimates from the model agree with empirical estimates of melt well using previously established methods of identifying melt with a single threshold Clarkson et al. (2022). This coastal temperature trend is consistent with local climate trends of higher temperatures occurring closer to the coast and demonstrates the impact of the upper limit from the truncated observations. This also demonstrates the difficulty mentioned above of finding a single distribution to describe the data at all locations of the ice sheet.

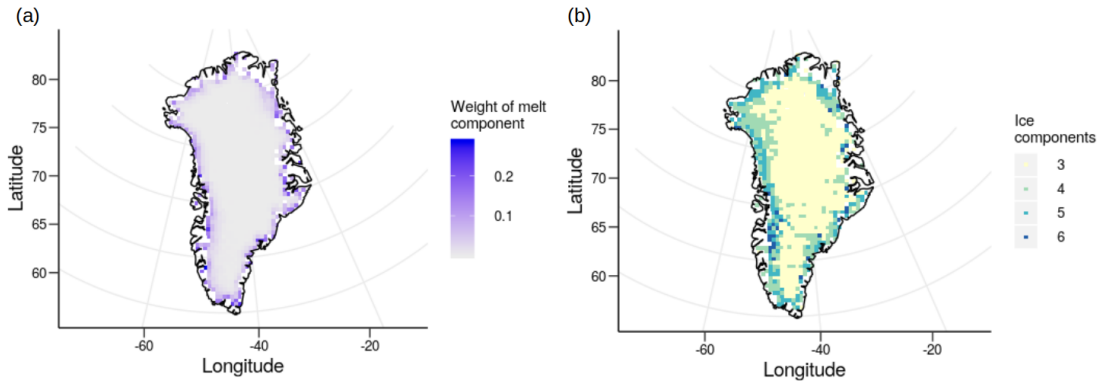

Fig. 5: (a) Estimated weight $\hat{\phi}$ of the melt component of each fitted mixture model. Due to restrictions during the fitting process the minimum possible value of this estimate is 0.0005. (b) The number of ice components in the best fitting mixture model at each cell in our study area.

While there is value in examining how each parameter varies across the ice sheet, to understand melt it is most useful to consider variability and trends in the weight of the melt component. This gives an overview of melt observed across the ice sheet, and reflects broad climate trends and the coastal effect. Figure 5 shows coastal areas having more of their distribution associated with melt temperatures, with cells in the south having a slightly higher weight than those in the north. Melt probability (6) is estimated for each observation and used to calculate the expected number of melt observations at each cell. The cross-cell average is then used to estimate the number of melt observations across the ice sheet in each year. These values are likely to be under-estimates since the expected melt at each cell would be higher at almost all cells if it were possible to also include cloudy days in the analysis. Figure 6 shows the majority (90.7%) of cells have a non-trivial expectation of melt (> 0.001) in an average year. Although melt is more common on the coasts, it is clear from the expected melt estimates that this is not exclusive with some inland cells also experiencing melt with relatively high frequency, and the main factors that appear to determine the melt extent are distance to the coast and elevation.

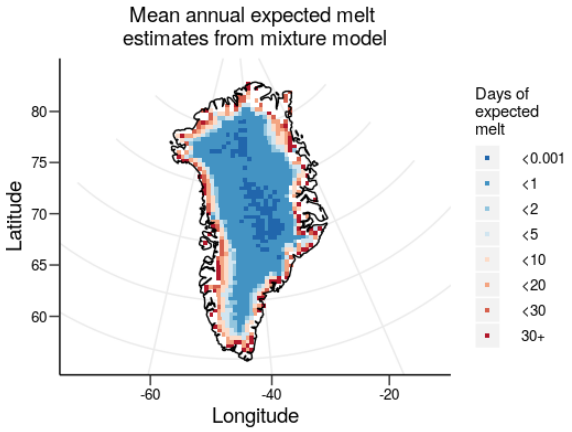

Fig. 6: Expected melt calculated from the fitted mixture models. The probability of each observation representing a melt temperature is calculated then averaged to the expectation in a single year.

5. Discussion

Extreme value analysis has a long-established link with hazard modelling, risk management and climate change analysis. Whilst historically focus has been on the tail behaviour of univariate responses, more recently considerable attention has been given to advancing extreme value theory and models for multivariate responses, due to (a) an increasing awareness of the potential for impact from multiple hazards and (b) an explosion in the availability of high-dimensional spatio-temporal data products, e.g. satellite images and numerical climate model output. Analysis of the extreme events of such data sets raises many challenges, from defining and identifying jointly extreme events to developing both statistical models and inference tools and the probability theory which underpins them. In this paper, we step back to assess ongoing challenges in univariate extreme value modelling. Such challenges are not restricted to model development, and it is our opinion that for them to be met successfully requires sustained dialogue between statisticians and fellow scientists. Evaluating impacts of climate change on natural hazards requires an understanding of the physical behaviour of both the hazard and any driving processes. Only once this is appropriately incorporated into the statistical model can attribution and prediction of the effects of climate change begin.

The two analyses presented in this paper demonstrate the necessity for novel, bespoke, context-driven models to address questions driven by climate change. Whilst IST and air temperature data sets share some similarities - both measure the same physical variable on or near the Earth's surface - the most suitable modelling approach for each is quite different due to differences in the physical behaviour of the processes and their drivers, the data collection methods and the underlying research questions. For ISTs, the physical properties of both the ice sheet and the melt process drive the model. Seemingly straightforward tasks - such as defining melt - are non-trivial and, despite

representing the tail of the sample, the distribution of temperatures around the melt point does not comply with the assumptions necessary for an out-of-the-box extreme value analysis. Whilst the proposed Gaussian mixture model is not consistent with more routine methods for modelling tail data, it does incorporate an understanding of the melt process that cannot be included in a PoT model. In contrast, the core challenge in the US air temperature study arises from the 2021 heatwave which, despite accounting for global temperature trends, cannot be explained through an extreme value analysis of historical data. To improve our understanding of the extremity of the 2021 event, a PoT random effects model is used to give a better indication of the relative size of the 2021 heatwave in the context of previous extreme temperatures, climate trends and inter-year variability due to unobserved climate drivers.

All statistical models are a compromise between model parsimony and functionality. Further, the scope for generalisation from any statistical model is entirely dependent on the quality and quantity of the data to which the model is fitted. With ever more powerful computing power at our disposal it can be tempting to develop high-dimensional models that aim to mimic their process-based counterparts, regardless of the size and quality of the data set. It is our belief that parsimony is in fact to be preferred, and that a key advantage of statistical models is their comparative cheapness to fit and ease of interpretation. At the same time, some compromise may be desirable; where incorporation of additional model features significantly improves interpretation, predictive abilities or our understanding of the physical process, then such features should be included where the information in the data supports this. Ultimately the utility of the model should be assessed by its ability to describe the data set, with both the most suitable modelling strategy and the necessary level of complexity judged against this benchmark.

The models proposed here are no exception, having both advantages and limitations. The random effects model provides a more flexible description of the data than permitted under a parametric regression model. However, unlike a parametric regression model, it does not permit estimation of the extreme behaviour in a specific year. It further assumes that, conditional on the annual random effects, the observations are independent. In practice there is likely to be some short-range serial dependence in the largest values. The model could be extended to incorporate serial dependence in either, or both, of the random effects and observations; the consequences of this would be an increased number of model parameters and a much more complex inference process. The mixture model could be criticised for diverging from the traditional approach of analysing only the tail data. Preliminary investigations using this approach provided such a poor fit to the data that it was immediately clear that an alternative approach - taking into account the unusual shape of the tail distribution - was required. The resulting model provides a much better description of the upper tail and, indeed, is a good fit to the full range of values. Neither the random effects nor the mixture model accounts for the inherent spatial structure in the physical processes and whilst this is beyond the scope of the current research it is the subject of ongoing research by the authors.

We end by discussing some broader open questions for the application of extreme value analysis in climate change scenarios. The first of these concerns guidance for scientists on the most appropriate way to incorporate change in either univariate or multivariate models. We conjecture that it is unlikely that there will ever be a definitive way to do

this due to the diverse physical properties exhibited by different environmental hazards. Instead, statisticians should strive for an increased awareness of the appropriateness, advantages and limitations of the available methods through direct comparisons of these different strategies. Such comparisons should also support an increased understanding of the benefits of trialling multiple modelling approaches.

Secondly, extreme value models are asymptotically justified in the sense of extrapolation into the tail of an underlying distribution. Most existing approaches to incorporating change assume the model parameters vary in time and/or in response to driving physical processes. Such regression-based models, whether semi- or fully parametric, have no similar justification for the extrapolation of trends into the future. Further, the predictions of future events will be very sensitive to the form assumed for the change. Again, this is strong motivation to search for more reliable, robust methods.

Thirdly, most routine multivariate, spatio- and temporal extreme value models require a separate model for the univariate marginal distributions. It is therefore critical that these marginal models reflect as faithfully as possible both the empirical properties of the data and physical properties of the underlying processes. As in the case of the ice surface temperatures discussed in Section 4, this might mean reconsidering the use of an extreme value model for the marginal distributions.

Finally, much work remains to identify risk measures which still have an interpretation under climate change; return levels (periods) make less sense if there is a long-term trend - what is deemed extreme now may no longer be extreme in 100 years. Return levels and periods which condition on a time period or covariates (Eastoe and Tawn, 2009) are more realistic but harder to interpret. Related to this is the problem of engaging with fellow scientists and the general public on the topic of changes in the risk of extreme events. It is vital to emphasise that acknowledgement of such change should not imply a loss of credibility in the predictions, but rather a more accurate reflection of a changing world.

Acknowledgements

We thank Paul Young for an introduction to the US temperature data, and David Leslie and Rachael Duncan for helpful comments on the draft. DC, EE and AL were supported by the EPSRC-funded Data Science of the Natural Environment project (EP/R01860X/1).

A. MCMC trace plots

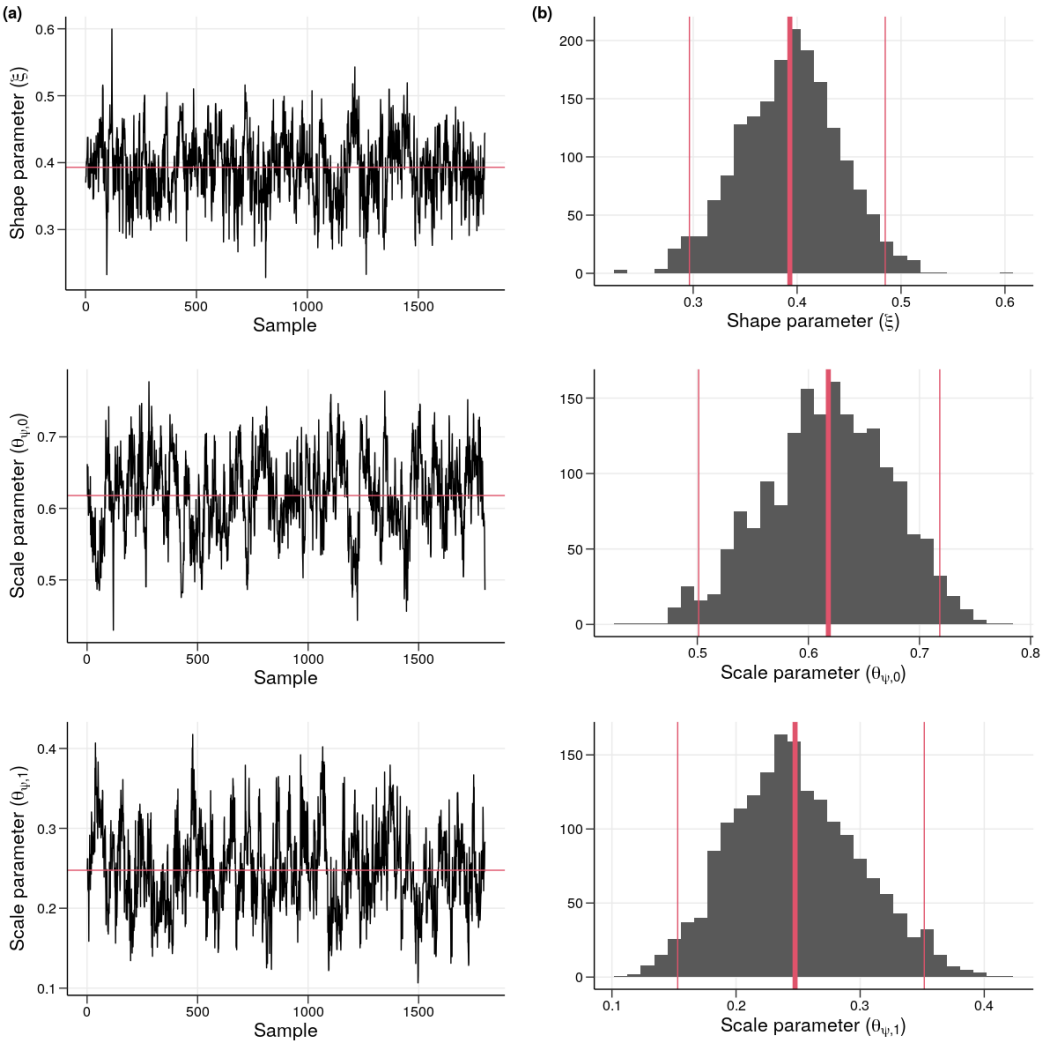

Fig. 7: (a) Trace plot of the shape parameter for weather station 'Cedar Tree' with the mean of the estimates (red line). (b) Histogram of the sampled shape parameters from the same MCMC chain as the trace plot, with the mean (bold red line) and 2.5% and 97.5% quantiles (normal red lines).

References

Acero, F., García, J. A., Gallego, M. C., Parey, S., and Dacunha-Castelle, D. (2014). Trends in summer extreme temperatures over the Iberian Peninsula using nonurban station data. *Journal of Geophysical Research: Atmospheres*, 119(1):39–53.

Burke, E. J., Perry, R. H., and Brown, S. J. (2010). An extreme value analysis of uk drought and projections of change in the future. *Journal of Hydrology*, 388(1-2):131–143.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic change*, 127(2):353–369.

Clarkson, D., Eastoe, E., and Leeson, A. (2022). Melt probabilities and surface temperature trends on the Greenland ice sheet using a Gaussian mixture model. *The Cryosphere*, 16(5):1597–1607. Publisher: Copernicus GmbH.

Cooley, D. (2009). Extreme value analysis and the study of climate change. *Climatic change*, 97(1):77–83.

Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O., and Sun, Y. (2012). A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *Revstat*, 10(1):135–165.

Davis, R. A., Mikosch, T., and Zhao, Y. (2013). Measures of serial extremal dependence and their estimation. *Stochastic Processes and their Applications*, 123(7):2575–2602.

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):393–425.

Drees, H. (2008). Some aspects of extreme value statistics under serial dependence. *Extremes*, 11(1):35–53.

Eastoe, E. F. (2019). Nonstationarity in peaks-over-threshold river flows: A regional random effects model. *Environmetrics*, 30(5):e2560.

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.

Economou, T., Stephenson, D. B., and Ferro, C. A. (2014). Spatio-temporal modelling of extreme storms. *The Annals of Applied Statistics*, 8(4):2223–2246.

Fawcett, L. and Walshaw, D. (2006). A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(5):631–646.

Friederichs, P. (2010). Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13(2):109–132.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pages 423–453.

Gouldsborough, L., Hossain, R., Eastoe, E., and Young, P. (2021). A Temperature Dependent Extreme Value Analysis of UK Surface Ozone, 1980 – 2019. Submitted.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press.

Hall, D. K., Cullather, R. I., DiGirolamo, N. E., Comiso, J. C., Medley, B. C., and Nowicki, S. M. (2018). A multilayer surface temperature, surface albedo, and water vapor product of greenland from modis. *Remote Sensing*, 10(4).

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, pages 153–167.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22(1):47–64.

Jonathan, P. and Ewans, K. (2013). Statistical modelling of extreme ocean environments for marine design: a review. *Ocean Engineering*, 62:91–109.

Jones, M., Randell, D., Ewans, K., and Jonathan, P. (2016). Statistics of extreme ocean environments: Non-stationary inference for directionality and other covariate effects. *Ocean Engineering*, 119:30–46.

Katz, R. W. (1999). Extreme value theory for precipitation: sensitivity analysis for climate change. *Advances in water resources*, 23(2):133–139.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304.

Koenig, L. S. and Hall, D. K. (2010). Comparison of satellite, thermochron and air temperatures at Summit, Greenland, during the winter of 2008/09. *Journal of Glaciology*, 56(198):735–741.

Kulp, S. A. and Strauss, B. H. (2019). New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. *Nature Communications*, 10(1).

Laurini, F. and Tawn, J. A. (2003). New estimators for the extremal index and other cluster characteristics. *Extremes*, 6(3):189–211.

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

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910. Publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology.

Philip, S. Y., Kew, S. F., van Oldenborgh, G. J., Yang, W., Vecchi, G. A., Anslow, F. S., Li, S., Seneviratne, S. I., Luu, L. N., Arrighi, J., Singh, R., van Aalst, Hauser, M., Schumacher, D. L., Marghidan, C. P., Ebi, K. L., Bonnet, R., Vautard, R., Tradowsky, J., Coumou, D., Lehner, F., Rodell, C., Stull, R., Howard, R., Gillett, N., and Otto, F. E. L. (2021). Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021. page 37.

Pickands, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

Rahmstorf, S., Foster, G., and Cahill, N. (2017). Global temperature evolution: recent trends and some pitfalls. 12(5):054001. Publisher: IOP Publishing.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

Reich, B., Cooley, D., Foley, K., Napelenok, S., and Shaby, B. (2013). Extreme value analysis for evaluating ozone control strategies. *The annals of applied statistics*, 7(2):739.

Rignot, E., Velicogna, I., Broeke, M. R. v. d., Monaghan, A., and Lenaerts, J. T. M. (2018). Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophysical Research Letters*.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of computational and graphical statistics*, 18(2):349–367.

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., and Thomson, A. W. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, 10:5.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145.

Simpson, E. S. and Wadsworth, J. L. (2021). Conditional modelling of spatio-temporal extremes for Red Sea surface temperatures. *Spatial Statistics*, 41:100482.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):515–528.

Sterl, A., Severijns, C., Dijkstra, H., Hazeleger, W., van Oldenborgh, G. J., van den Broeke, M., Burgers, G., van den Hurk, B., van Leeuwen, P. J., and van Velthoven, P. (2008). When can we expect extremely high surface temperatures? *Geophysical Research Letters*, 35(14).

Thomson, A. W., Dawson, E. B., and Reay, S. J. (2011). Quantifying extreme behavior in geomagnetic activity. *Space Weather*, 9(10).

Towe, R., Eastoe, E., Tawn, J., and Jonathan, P. (2017). Statistical downscaling for future extreme wave heights in the North Sea. *The Annals of Applied Statistics*, 11(4):2375–2403.

Towe, R., Tawn, J., Eastoe, E., and Lamb, R. (2020). Modelling the clustering of extreme events for short-term risk assessment. *Journal of Agricultural, Biological and Environmental Statistics*, 25(1):32–53.

Von Mises, R. (1936). La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique*, 1:141–160.

Winter, H. C. and Tawn, J. A. (2017). kth-order Markov extremal models for assessing heatwave risks. *Extremes*, 20(2):393–415.

Winter, H. C., Tawn, J. A., and Brown, S. J. (2016). Modelling the effect of the El Nino-Southern Oscillation on extreme spatial temperature events over Australia. *The Annals of Applied Statistics*, 10(4):2075–2101.

Yee, T. W. and Stephenson, A. G. (2007). Vector generalized linear and additive extreme value models. *Extremes*, 10(1):1–19.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds

with an application to return level estimation for US wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.