kth-order Markov extremal models for assessing heatwave risks

Hugo C. Winter · Jonathan A. Tawn

Abstract Heatwaves are defined as a set of hot days and nights that cause a marked short-term increase in mortality. Obtaining accurate estimates of the probability of an event lasting many days is important. Previous studies of temporal dependence of extremes have assumed either a first-order Markov model or a particularly strong form of extremal dependence, known as asymptotic dependence. Neither of these assumptions is appropriate for the heatwaves that we observe for our data. A first-order Markov assumption does not capture whether the previous temperature values have been increasing or decreasing and asymptotic dependence does not allow for asymptotic independence, a broad class of extremal dependence exhibited by many processes including all non-trivial Gaussian processes. This paper provides a kth-order Markov model framework that can encompass both asymptotic dependence and asymptotic independence structures. It uses a conditional approach developed for multivariate extremes coupled with copula methods for time series. We provide novel methods for the selection of the order of the Markov process that are based upon only the structure of the extreme events. Under this new framework, the observed daily maximum temperatures at Orleans, in central France, are found to be well modelled by an asymptotically independent third-order extremal Markov model. We estimate extremal quantities, such as the probability of a heatwave event lasting as long as the devastating European 2003 heatwave event. Critically our method enables the first reliable assessment of the sensitivity of such estimates to the choice of the order of the Markov process.

H.C. Winter, J.A. Tawn

Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K.
Present address: of H.C. Winter
EDF Energy R&D UK Centre, Cardinal Place, 80 Victoria Street, London, SW1E 5JL, U.K.
E-mail: hugo.winter@edfenergy.com

1 Introduction

Many devastating natural hazards are caused by events that are extreme and
rare. Extreme value theory provides a general framework for modelling such
extreme values. In many situations a singular extreme observation does not
have a great effect, whereas combinations and runs of extreme values can
cause widespread devastation. A heatwave is defined as a set of consecutive
days and/or nights that lead to an increase in mortality. So when estimating
risks attributed to heatwaves we need to account for the fact that one very
hot day may not cause a large increase in mortality whereas a run of consec-
utive less hot days can be far more damaging. Therefore any extreme value
model utilised to help assess the risk of heatwaves must be able to model such
behaviour reliably. In the terminology of extreme value theory, this requires
a model that can capture the extremal temporal dependence structure along-
side marginal tail characteristics. The data that we will model in this paper
relates to summer daily maximum temperatures from a single site observed
over a number of years. Therefore we want to model the extreme events of a
univariate stationary series.

Let $\{Y_t\}$ be a stationary time-series with upper endpoint y_F. We are interested
in modelling the behaviour for $\{Y_t\}$ above some high threshold u_Y. Following
copula time series methods (Joe 1997), our approach is to separately model
the margins and dependence structure of $\{Y_t\}$. The most common approach to
modelling the marginal distributions of extreme values is to fit a generalized
Pareto distribution (GPD) to exceedances of u_Y. The GPD takes the form

$$P(Y_t - u_Y > y \mid Y_t > u_Y) = \left(1 + \frac{\xi y}{\sigma_{u_Y}}\right)_+^{-1/\xi} \qquad \text{for} \quad y \geq 0, \quad (1)$$

where $c_+ = \max(c, 0)$, $\sigma_{u_Y} > 0$ and $\xi \in \mathbb{R}$ are the scale and shape param-
eters of the GPD respectively (Coles 2001), with the scale parameter being
threshold dependent. The justification for this model is an asymptotic result
of Pickands (1971) that showed that, under weak conditions on Y_t, the dis-
tribution of suitably scaled exceedances of a threshold by Y_t converges to the
GPD as the threshold tends to the upper endpoint y_F. Thus the GPD model in
equation (1) assumes that the limiting result holds exactly for a large enough
threshold u_Y.

For heatwaves it is important to be able to model the distribution of the
number of exceedances of a critical level during a block of time. It is also
necessary to be able to estimate other extremal quantities of heatwave events,
here named cluster functionals. Methods exist to split a time-series of tem-
perature data into independent clusters of exceedances of the threshold u_Y,

where within each cluster groups of dependent exceedances occur and between clusters values are treated as independent. Clusters are not necessarily consecutive exceedances, in fact the most popular technique for cluster identification is the runs method (Smith and Weissman 1994), with run length l, which takes a cluster to be exceedances of u_Y that are not separated by a run of l consecutive non-exceedances of u_Y. The value of l can be selected automatically using methods of Ferro and Segers (2003). The number of clusters is Poisson distributed (Davison and Smith 1990). We wish to accurately model the temporal dependence of the within cluster values, i.e., the local time-series during an extreme event. Empirical distributions of cluster functionals could be used for inference of within cluster behaviour, but they have major limitations for extrapolation and so are only really suitable for model checking. Our approach is to use a kth-order Markov chain for $\{Y_t\}$, using only values of Y_t within lag k of an exceedance of u_Y. We term such a model an extremal Markov chain.

Many different approaches exist for modelling the multivariate dependence structure of extreme values. First, consider two random variables (Y_0, Y_τ) at a time lag τ. A key way to discriminate between approaches is through the lag τ extremal dependence measure χ_τ, often termed the tail coefficient, where

$$\chi_\tau = \lim_{y \to y_F} P(Y_\tau > y \mid Y_0 > y). \tag{2}$$

When $\chi_\tau > 0$, i.e., the largest values of the variables can occur together, the pair are termed asymptotically dependent. Asymptotic dependence arises when the conditions for multivariate regular variation hold and for max-stable distributions/processes; see de Haan and Ferreira (2006), Resnick (1987) and Davison et al (2012). When $\chi_\tau = 0$, i.e., the largest values of the variables cannot occur together, the pair are termed asymptotically independent. Asymptotic independence arises for all non-trivial Gaussian processes and for a broad range of examples identified by Ledford and Tawn (1997) and Heffernan (2000). The conditional extremes approach of Heffernan and Tawn (2004) currently is the only model that has the flexibility to capture both of these extremal dependence classes whilst being generalisable to higher-dimensional problems. We shall base our inference on this class.

A range of temporal dependence models for extreme values have been proposed, with some specific to heatwave applications. Smith et al (1997) provide a framework using first-order Markov chain approaches for modelling threshold exceedances and analysing the distribution of cluster functionals of extreme events. A weakness with their approach is that it assumes that at lag 1 only asymptotic dependence is possible, and this implies asymptotic dependence holds at all lags. Yun (2000), Fawcett and Walshaw (2006) and Ribatet et al (2009) outline extensions of this approach to kth-order Markov chains but also they are restricted to assuming asymptotic dependence for all lags. More recently, Reich et al (2014) formulate an asymptotically dependent max-stable process using random effects within a Bayesian framework, incorporating dependence within 10 day windows. A range of asymptotically independent

Markov processes have been assumed. After marginal preprocessing, Dupuis (2012) models heatwaves using an asymptotically independent AR(8) model. However this model is fitted to the whole series, not simply the extremes, so may lead to bias when applied to the extremes. Bortot and Tawn (1998) use theory from Ledford and Tawn (1997) to derive a class of models for first-order Markov chains that permits both asymptotic independence and asymptotic dependence. However these models are only justified when consecutive values are large, i.e., $Y_t > u_Y$ and $Y_{t+1} > u_Y$, which is restrictive for our application.

Winter and Tawn (2016) built a first-order Markov approach, based upon the conditional extremes approach of Heffernan and Tawn (2004), that can account for both asymptotic dependence and asymptotic independence and applies if at least one component of (Y_t, Y_{t+1}) is greater than u_Y. The limit theory for this model has been studied by Papastathopoulos et al (2017). They find the limiting joint behaviour of $(Y_{t+1}, \ldots, Y_{t+m})|Y_t > u_Y$, after suitable normalisation, as $u_Y \to y_F$, for any integer $m \geq 1$. For the daily maximum temperature data that are analysed in our paper, Winter and Tawn (2016) found that standard time series diagnostics, e.g., PACF and a comparison of observed and modelled cluster functionals, suggest that the first-order Markov assumption was reasonable. However, the physical mechanisms of heatwaves suggest that this is perhaps an oversimplification that could lead to an underestimation of the risk of a heatwave event. They also found strong evidence of asymptotic independence, with significant positive dependence, and that falsely assuming a first-order Markov model with asymptotic dependence leads to overestimation of heatwave characteristics. This paper seeks to take advantage of the higher-order structure of the extreme values of the process through a kth-order Markov model for extremes to provide more accurate estimates of the risk of a heatwave event.

We also seek to develop diagnostic tests to choose an appropriate order for the Markov process to fit to extreme events. Standard time-series diagnostics for choosing an appropriate Markov process are potentially misleading when considering the behaviour of extremes. If the process is kth-order Markov, then its extreme states will follow a Markov process with order of at most k. Ledford and Tawn (2003) developed diagnostic tools to test long and short range dependence assumptions within extreme events of both asymptotically dependent and asymptotically independent processes. However, these methods were unable to detect the order of the process. For asymptotically dependent processes Fawcett and Walshaw (2006) and Ribatet et al (2009) explore heuristic methods proposed in Smith et al (1997) for identifying the order. Here we seek to extend these tools to select the order of an extremal Markov process irrespective of whether it is asymptotically independent or asymptotically dependent. There are natural connections with the equivalent issue of identifying graphical structures in multivariate extremes, see Papastathopoulos and Tawn (2013) and Hitz and Evans (2015), but in these cases again the focus to date has been on asymptotically dependent variables.

Section 2 sets out the copula formulation for kth-order stationary Markov chains, with the asymptotic representations for these processes when in extreme states being identified in Section 3. Our asymptotically justified model for kth-order chains is set out in Section 4 and the inference for this is discussed in Section 5. A discussion of diagnostic methods for the choice of the order of the extremal Markov process is given in Section 6. Section 7 gives results for our temperature data set, from Orleans in central France, and includes comparisons with the results of Winter and Tawn (2016) for a first order Markov model. Discussion and conclusions are presented in Section 8.

2 Copula formulations for stationary Markov processes

We shall model the stationary time-series $\{Y_t\}$ by a kth-order Markov chain using copula time series methods. Under the assumption that a stationary time-series $\{Y_t\}$ follows a kth-order Markov process, the joint density function $f_{1:n}$ of $\mathbf{Y}_{1:n} = (Y_1, \ldots, Y_n)$ can be written as

$$f_{1:n}(\mathbf{y}_{1:n}) = f_{1:k}(\mathbf{y}_{1:k}) \prod_{t=1}^{n-k} f_{k+1|1:k}(y_{t+k} \mid \mathbf{y}_{t:t+k-1}),$$

where $f_{k+1|1:k}(\cdot \mid \cdot)$ is the conditional density function of $Y_{k+1}|\mathbf{Y}_{1:k}$. Here and throughout we subscript densities and vector variables to denote the associated indices of $\{Y_t\}$. We also use the notation $i : j$ to denote $(i, i+1, \ldots, j)$. For stationarity the joint density $f_{1:k+1}(\mathbf{y}_{1:k+1})$ must satisfy the property that its m-dimensional joint margins satisfy the condition

$$f_{i_1,\ldots,i_m}(\mathbf{y}_{1:m}) = f_{i_1+\tau,\ldots,i_m+\tau}(\mathbf{y}_{1:m}), \tag{3}$$

for all $m < k+1$, $\tau \in \mathbb{N}$, $i_j \in \mathbb{N}$ for $j = 1, \ldots, m$, with $1 \leq i_1 < \ldots < i_m + \tau \leq k+1$ and $\mathbf{y}_{1:m} \in \mathbb{R}^m$ (Joe 1997). As a consequence of condition (3) the margins f_i must be identical and we subsequently denote them by f. Additional dependence conditions must also hold, e.g., (Y_i, Y_j) and $(Y_{i+\tau}, Y_{j+\tau})$ have identical joint distributions.

We shall adopt a copula framework for modelling $f_{1:k+1}$, with associated joint distribution function $F_{1:k+1}$ satisfying

$$\begin{aligned} F_{1:k+1}(\mathbf{y}_{1:k+1}) &= C(F_Y(y_1), \ldots, F_Y(y_{k+1})) \\ &= C^X(F_X^{-1}\{F_Y(y_1)\}, \ldots, F_X^{-1}\{F_Y(y_{k+1})\}), \end{aligned} \tag{4}$$

where C is a copula with uniform margins and C^X is the associated joint distribution with identical marginal distribution functions F_X, where F_X^{-1} is the inverse of F_X. The copula C and joint distribution function C^X inherit the

stationarity conditions (3) that are required for $f_{1:k+1}$. Specifically, for C^X we require that its m-dimensional marginal $C^X_{i_1,\dots,i_m}$ satisfies

$$C^X_{i_1,\dots,i_m}(\mathbf{x}_{1:m}) = C^X_{i_1+\tau,\dots,i_m+\tau}(\mathbf{x}_{1:m}), \qquad (5)$$

for all $m < k+1$, $\tau \in \mathbb{N}$, $i_j \in \mathbb{N}$ for $j = 1,\dots,m$, with $1 \le i_1 < \dots < i_m + \tau \le k+1$ and $\mathbf{x}_{1:m} \in \mathbb{R}^m$.

The reason for considering the joint distributions C^X with non-uniform identical margins, instead of copulas with uniform margins, is that the extremal properties are more simply expressed for some non-uniform marginal choices. The most convenient choice of F_X depends on the context: the Fréchet or Pareto distributions are typically assumed for max-stable distributions; for conditional extremes Heffernan and Tawn (2004) use Gumbel margins; whereas for joint tail modelling Wadsworth and Tawn (2013) used exponential margins. Keef et al (2013) showed that a more comprehensive approach arises for Laplace margins with

$$F_X(x) = \begin{cases} \frac{1}{2}\exp(x), & x < 0, \\ 1 - \frac{1}{2}\exp(-x), & x \ge 0. \end{cases}$$

Trivially, if $\{Y_t\}$ is a stationary kth-order Markov chain, then $\{X_t\}$, defined by

$$X_t = \begin{cases} \log\{2F(Y_t)\} & \text{if} \quad F(Y_t) < 1/2, \\ -\log\{2[1 - F(Y_t)]\} & \text{if} \quad F(Y_t) \ge 1/2, \end{cases}$$

is a stationary kth-order Markov chain with Laplace margins. With this formulation it follows that the likelihood is

$$f_{1:n}(\mathbf{y}_{1:n}) = f_{1:k}(\mathbf{y}_{1:k}) \prod_{t=1}^{n-k} \frac{c^X_{1:k+1}(\mathbf{x}_{t:t+k})}{c^X_{1:k}(\mathbf{x}_{t:t+k-1})} \frac{f(y_{t+k})}{f_X(F_X^{-1}(F_Y(y_{t+k})))}, \qquad (6)$$

where $c^X_{1:k+1}$ and $c^X_{1:k}$ are the copula densities for $\mathbf{X}_{t:t+k}$ and $\mathbf{X}_{t:t+k-1}$ respectively and $x_t = F_X^{-1}\{F_Y(y_t)\}$ for all t. Here $c^X_{1:k+1}$ satisfies the joint density condition that arises from the stationary copula condition (5). Note that the term in the product of the likelihood factorises into separate terms for the marginal and dependence structure of the time series.

3 Asymptotic representations for stationary extremal Markov chains

3.1 Preliminaries

We now restrict the stationary kth-order Markov chain $\{Y_t\}$ to only extreme events, i.e., what we call extremal Markov chains. We focus on the extreme

events as we don't want the dependence structure of the body of the process to influence the fitting of the model to the extreme events. In particular we do this by defining an extreme event to be when the first component of the vector $\mathbf{Y}_{t:t+k}$ is extreme, i.e., $Y_t > u_E$, where u_E is a high threshold. Here u_E is not necessarily equal to the marginal modelling threshold u_Y, defined in Section 1, as it is a dependence structure definition of an extreme which can differ from a marginal definition as the rate of convergence to the limiting form can be different for these two characteristics.

So we only model the conditional distribution $Y_{t+k}|\mathbf{Y}_{t:t+k-1}$ parametrically when $Y_t > u_E$, and we model this conditional distribution non-parametrically when $Y_t \leq u_E$. Equivalently, this corresponds to modelling $X_{t+k}|\mathbf{X}_{t:t+k-1}$ when $X_t > u$, where $u = F_X^{-1}(F_Y(u_E))$. Similarly, marginally we model the distribution F_Y parametrically for $Y_t > u_Y$ and non-parametrically for $Y_t \leq u_Y$. If we use the copula formulation (4) then for large n, the likelihood (6) for our model simplifies to

$$f_{1:n}(\mathbf{y}_{1:n}) \approx \prod_{t:x_t>u} \frac{c_{1:k+1}^X(\mathbf{x}_{t:t+k})}{c_{1:k}^X(\mathbf{x}_{t:t+k-1})} \prod_{t:y_t>u_Y} f(y_t), \tag{7}$$

as the terms that have been dropped contribute little or nothing to the likelihood for the extreme value model. Details of the marginal model for F_Y, that is required for the second product in this likelihood, are given in Section 4.1. The first term requires the conditional distribution of $X_{t+k} \mid \mathbf{X}_{t:t+k-1}$ when $X_t > u$. Section 3.2 presents our main method for modelling extremal dependence based upon the conditional approach outlined in Heffernan and Tawn (2004). However, this model gives a limiting representation only for $\mathbf{X}_{t+1:t+k} \mid (X_t > u)$ as $u \to \infty$. Therefore we propose an extension to give the behaviour of the distribution of $X_{t+k} \mid \mathbf{X}_{t:t+k-1}$ when $X_t > u$ for $u \to \infty$.

3.2 Asymptotics for conditional extremes

Heffernan and Tawn (2004) propose an asymptotically justified conditional multivariate extremes approach for modelling the extremes of a vector $\mathbf{X}_{t:t+m}$, for any integer $m > 0$, with all variables having Laplace margins and a joint density. We present that model and then discuss the additional conditions required for $\{X_t\}$ to be a stationary kth-order Markov process. Throughout the rest of the paper all vector calculations are to be interpreted componentwise.

To explore the conditional distribution $P\{\mathbf{X}_{t+1:t+m} \leq \mathbf{x} \mid X_t > u\}$ for large u we use an asymptotically justified form for this distribution as $u \to \infty$. If $\mathbf{x}$ is fixed, in general the limit distribution will be a degenerate distribution. Hence $\mathbf{X}_{t+1:t+m}$ needs to be normalised appropriately so that the limiting conditional distribution is non-degenerate as $u \to \infty$. Heffernan and Resnick (2007) propose that $\mathbf{X}_{t+1:t+m}$ is linearly normalised as a function of either X_t

or u. For statistical purposes it is most simple to use the approach of Heffernan and Tawn (2004) and to normalise by X_t, so that is the approach that we will take.

Heffernan and Tawn (2004) assume that there exist functions $a \colon \mathbb{R} \to \mathbb{R}^m$ and $b \colon \mathbb{R} \to \mathbb{R}_+^m$, such that

$$\mathrm{P}\left(\frac{\mathbf{X}_{t+1:t+m} - a(X_t)}{b(X_t)} \leq \mathbf{z}_{1:m}, X_t - u > x \;\middle|\; X_t > u\right) \to G_{1:m}(\mathbf{z}_{1:m})\exp(-x),$$
(8)

as $u \to \infty$ with $\mathbf{z}_{1:m} \in \mathbb{R}^m$, where $G_{1:m}$ is a joint distribution function that is non-degenerate in each margin, i.e., for $j = 1, \ldots, m$ the jth margin G_j of $G_{1:m}$ is non-degenerate. There is no finite parametric form for $G_{1:m}$.

Under weak assumptions on the joint distribution of $\mathbf{X}_{t:t+m}$, Heffernan and Resnick (2007) show that componentwise a and b must be regularly varying functions satisfying certain constraints, which for Laplace margins corresponds to each of the components of a (respectively b) being regularly varying functions of index 1 (respectively less than 1). Within this structure Heffernan and Tawn (2004) found that a simple form for a and b holds for a very broad range of copulas. In particular, they assume that

$$a(X_t) = \boldsymbol{\alpha}_{1:m} X_t \text{ and } b(X_t) = X_t^{\boldsymbol{\beta}_{1:m}}$$
(9)

where $\boldsymbol{\alpha}_{1:m} = (\alpha_1, \ldots, \alpha_m) \in [-1, 1]^m$ and $\boldsymbol{\beta}_{1:m} = (\beta_1, \ldots, \beta_m) \in (-\infty, 1)^m$. This canonical parametric subfamily of a and b provides a parsimonious yet flexible family for statistical modelling.

A key property of the limit (8) is that the limiting distribution factorises, corresponding to large values of X_t being independent of the associated normalised $\mathbf{X}_{t+1:t+m}$. Here, stationarity of $\{X_t\}$ ensures that $\boldsymbol{\alpha}_{1:m}$, $\boldsymbol{\beta}_{1:m}$ and $G_{1:m}$ do not depend on t. When $\beta_i < 0$ then X_{t+i} is asymptotically a multiple of X_t for all $X_t > u$ as $u \to \infty$. As this deterministic structure is unlikely to occur in practice, we take a pragmatic approach and restrict the parameter space for $\boldsymbol{\beta}_{1:m}$ so that $(\beta_1, \ldots, \beta_m) \in [0, 1)^m$.

Different types of extremal dependence lead to different values of the extremal dependence parameters $\boldsymbol{\alpha}_{1:m}$ and $\boldsymbol{\beta}_{1:m}$. For $1 \leq j \leq m$, when $\alpha_j = 1$ and $\beta_j = 0$ the variables (X_t, X_{t+j}) are asymptotically dependent and are asymptotically independent when $\alpha_j < 1$. Within the asymptotic independence case a further resolution of the dependence structure is possible with $0 < \alpha_j < 1$ or $\alpha_j = 0$ and $\beta_j > 0$ corresponding to positive dependence; independence when $\alpha_j = \beta_j = 0$ and G_j is the Laplace distribution function; and negative dependence when $-1 \leq \alpha_j < 0$. For more information see Keef et al (2013).

Stationarity of $\{X_t\}$ requires condition (5). However these conditions do not

appear to impose any further constraints on the $\boldsymbol{\alpha}_{1:k}$, $\boldsymbol{\beta}_{1:k}$ and $G_{1:k}$, when k is the order of the Markov process in expression (8), for $k \leq m$. We have explored a range of examples that seem to support this. The reason for this freedom appears to be that we are not looking at the whole copula but at a slice with large X_t. Based on these empirical findings, we conjecture that there is no relationship between these features and so in our model these features are unconstrained. In contrast the values of $\boldsymbol{\alpha}_{k+1:m}$, and $\boldsymbol{\beta}_{k+1:m}$ and joint distribution function $G_{k+1:m}$, for any $m \geq k+1$, do have structure imposed by the stationary Markov behaviour and are determined entirely by $\boldsymbol{\alpha}_{1:k}$, $\boldsymbol{\beta}_{1:k}$ and $G_{1:k}$.

For stationary kth-order Markov chains with $k > 1$ no theoretical results are published other than for asymptotically dependent processes (Yun 2000). However for asymptotically dependent and asymptotically independent stationary first-order Markov processes, $k = 1$, Papastathopoulos et al (2017) derive two possible forms for $\boldsymbol{\alpha}_{2:m}$, $\boldsymbol{\beta}_{2:m}$ and $G_{1:m}$ for all $m \geq 2$ depending on the value of α_1. Specifically, when if $0 < \alpha_1 \leq 1$ then, for $2 \leq \tau \leq m$, $\alpha_\tau = \alpha_1^\tau$, $\beta_\tau = \beta_1$ and where $G_{1:\tau}$ is the joint distribution of $\mathbf{Z}_{1:\tau}$ where

$$Z_j = \sum_{i=1}^{j} \alpha_1^{\tau - i + i\beta_1} Z_{1,i} \text{ for } j = 1, \ldots, \tau,$$

where $Z_{1,i}$ are independent and identically distributed over i, with distribution function G_1. Note that the situation where $\alpha_1 = 1$ and $\beta_1 = 0$, i.e., asymptotic dependence at lag one leads to asymptotic dependence at all lags, and that $\{Z_\tau\}$ is a random walk, thus giving the results of Smith (1992). Alternatively, when $\alpha_1 = 0$ and $0 \leq \beta_1 < 1$, they find that, for $2 \leq \tau \leq m$, $\alpha_\tau = 0$, $\beta_\tau = \beta_1^\tau$ with $G_{1:\tau}$ the joint distribution of

$$Z_j = \prod_{i=1}^{j} (Z_{1,i})^{\beta_1^{i-1}} \text{ for } j = 1, \ldots, \tau,$$

with $Z_{1,i} \geq 0$ being independent and identically distributed over i, with distribution function G_1. These results prove helpful in Section 6 for developing tests for the process being a first, or higher, order extremal Markov chain.

Finally note that a kth-order Markov process can behave as a Markov process with order k_E, with $k_E \leq k$, in its extremes. In this case the information in expression (8) for $m = k_E$ is only required to determine $\boldsymbol{\alpha}_{k_E+1:m}$, and $\boldsymbol{\beta}_{k_E+1:m}$ and joint distribution function $G_{k_E+1:m}$ for any $m > k_E$. The situation $k_E < k$ arises for a kth-order Markov process when $(\boldsymbol{\alpha}_{k_E+1:k}, \boldsymbol{\beta}_{k_E+1:k})$ are determined by $(\boldsymbol{\alpha}_{1:k_E}, \boldsymbol{\beta}_{1:k_E})$ and $G_{1:k}$ factorises into $G_{1:k_E}$ and $G_{k_E:k}$, with the latter being the product of Laplace distribution functions. From a statistical perspective we are only interested in estimating k_E. For notational simplification though, throughout the rest of the paper we do not distinguish between k_E and k, and use k to denote k_E.

4 Models for conditional extremes

4.1 Marginal modelling

As $\{Y_t\}$ is a stationary series the marginal distributions F_Y are identical. As discussed in Section 1, there is an asymptotic justification for modelling the marginal excesses of u_Y by Y_t as following a GPD with distribution function (1). But we do not specify a parametric form for the distribution of Y_t below u_Y. Following Coles and Tawn (1991) we model the marginal distribution of Y_t as

$$F_Y(y) = \begin{cases} 1 - [1 - \tilde{F}(u_Y)]\left(1 + \xi\frac{y-u_Y}{\sigma_{u_Y}}\right)_+^{-1/\xi}, & y > u_Y, \\ \tilde{F}(y), & y \leq u_Y, \end{cases}$$

where $\tilde{F}(y)$ is the empirical marginal distribution function of $\mathbf{Y}_{1:n}$.

4.2 Temporal dependence modelling

The limiting form of the conditional distribution (8) motivates our asymptotically justified model for the conditional distribution of $\mathbf{X}_{t+1:t+k}$ given $X_t > u$ for a large fixed value u, and k the order of the extremal Markov process. Specifically, we assume that the limiting form (8) holds exactly for all values of $X_t > u$ with $m = k$, that the normalising functions a and b can be given by forms (9), and that $\mathbf{X}_{t:t+k}$ has a density. It follows that we have

$$\mathbf{X}_{t+1:t+k}|(X_t > u) = \boldsymbol{\alpha}_{1:k}X_t + X_t^{\boldsymbol{\beta}_{1:k}}\mathbf{Z}_{1:k}, \tag{10}$$

for $\boldsymbol{\alpha}_{1:k} \in [-1, 1]^k$, $\boldsymbol{\beta}_{1:k} \in [0, 1)^k$ and $\mathbf{Z}_{1:k}$ is a random variable, independent of t and X_t, with distribution function $G_{1:k}$ and joint density $g_{1:k}$. Trivially, model (10) satisfies the limiting property (8) as $\mathbf{Z}_{1:k}$ and X_t are independent and exceedances of $u_Y > 0$ are unit exponential. The recurrence relationship (10) cannot be interpreted as holding for all X_t, as it only applies for $X_t > u$. As such, series generated under this process (like tail chains in Smith (1992) and Papastathopoulos et al (2017)) have negative drifts that ensure the process returns from an extreme state to the body of the distribution.

We need an asymptotically motivated model for $X_{t+k} \mid \mathbf{X}_{t:t+k-1}$ when $X_t > u$. As no formal limiting results exist our approach provides a heuristic approach which provides a flexible modelling framework. By assuming that model (10) holds exactly for $X_t = x_t > u$ it follows that

$$X_{t+k}|(\mathbf{X}_{t:t+k-1} = \mathbf{x}_{t:t+k-1}) = \alpha_k X_t + X_t^{\beta_k} Z_{k|1:k-1}, \tag{11}$$

where $Z_{k|1:k-1}$ is a random variable with the same distribution as the conditional distribution of Z_k given that

$$\mathbf{Z}_{1:k-1} = \frac{\mathbf{x}_{1:k-1} - \boldsymbol{\alpha}_{1:k}x_t}{x_t^{\boldsymbol{\beta}_{1:k}}} := \mathbf{z}_{1:k-1}.$$

This conditional variable $Z_k \mid \mathbf{Z}_{1:k-1} = \mathbf{z}_{1:k-1}$ has distribution function

$$G_{k|1:k-1}(z \mid \mathbf{z}_{1:k-1}) = \int_{-\infty}^{z} g_{k|1:k-1}(s \mid \mathbf{z}_{1:k-1})ds,$$

where $g_{k|1:k-1}(\cdot \mid \mathbf{z}_{1:k-1})$ is the associated conditional density function, given by $g_{1:k}(\mathbf{z}_{1:k-1}, \cdot)/g_{1:k-1}(\mathbf{z}_{1:k-1})$. It follows iteratively that for $j = 1, \ldots$ that $X_{t+k+j}|(\mathbf{X}_{t:t+k+j-1} = \mathbf{x}_{t:t+k+j-1})$ is also given by expression (11). Consequently we can simulate the values of X_{t+k+j}, for $1 \leq j \leq m$, without explicitly evaluating $\boldsymbol{\alpha}_{k+1:m}, \boldsymbol{\beta}_{k+1:m}$ and $G_{k+1:m}$.

As this is a statistical model developed based on heuristic arguments it may not give a structure that is consistent with the limiting tail chain of the kth-order process, but what is critical here is that it gives a flexible and parsimonious statistical model for capturing the dependence of the process over a high threshold. Section 7 presents evidence that supports our statistical model through the realistic realisations of extreme events that it generates.

5 Inference

5.1 Inference for model parameters

Here we assume that the process is a stationary kth-order extremal Markov process, with k known. The estimation of k is discussed in Section 6. Under this assumption we estimate the extremal marginal parameters (σ_{u_Y}, ξ), the extremal dependence parameters $(\boldsymbol{\alpha}_{1:k}, \boldsymbol{\beta}_{1:k})$ and the distribution $G_{k|1:k-1}$. Our approach is a pseudo maximum likelihood inference scheme with block bootstrap methods for obtaining confidence intervals. Specifically, we use stepwise inference, with marginal parameters estimated first, then the dependence parameters, and then $G_{k|1:k-1}$ is estimated non-parametrically. This approach to separate inference for marginal and dependence structure is standard in copula modelling and has been shown to not lose much efficiency; see Genest et al (1995) and Liang and Self (1996).

From likelihood (7) our approach corresponds to standard maximum likelihood estimation for (σ_{u_Y}, ξ) using all the threshold exceedances of u_Y by $\{Y_t\}$. Likelihood (7) then simplifies down to a product over the density contributions for $X_{t+k} \mid \mathbf{X}_{t:t+k-1}$ when $X_t > u$. Since $G_{1:k}$ and its marginals do not take any finite parametric form, we make a temporary working assumption that $\mathbf{Z}_{1:k}$ are independent Gaussian variables with $Z_j \sim \mathrm{N}(\mu_j, \gamma_j^2)$ for $j = 1, \ldots, k$ (Keef et al 2013). Under this assumption

$$X_{t+j} \mid \{X_t = x\} \sim N\left(\alpha_j x + \mu_j x^{\beta_j}, \gamma_j^2 x^{2\beta_j}\right) \qquad \text{for} \quad x > u, \qquad (12)$$

where $j = 1, \ldots, k$ for all t with $X_t > u$. If we denote the corresponding Gaussian likelihood by L_j, for $j = 1, \ldots, k$, then the overall likelihood is

$$L(\boldsymbol{\alpha}_{1:k}, \boldsymbol{\beta}_{1:k}, \boldsymbol{\mu}_{1:k}, \boldsymbol{\gamma}_{1:k}) = \prod_{j=1}^{k} L_j(\alpha_j, \beta_j, \mu_j, \gamma_j), \tag{13}$$

due to the independence assumption. Maximisation of likelihood L gives estimates $(\hat{\boldsymbol{\alpha}}_{1:k}, \hat{\boldsymbol{\beta}}_{1:k}, \hat{\boldsymbol{\mu}}_{1:k}, \hat{\boldsymbol{\gamma}}_{1:k})$. Unless there are constraints on the parameters across lags then in practice these values can be most easily be obtained through maximising L_j separately for each j. Although it may appear that ignoring the dependence of $\mathbf{Z}_{1:k}$ would bias the inference for $\boldsymbol{\alpha}_{1:k}$ and $\boldsymbol{\beta}_{1:k}$, standard regression results show this is not the case. Furthermore, Lugrin et al (2016) find that estimating $(\hat{\boldsymbol{\alpha}}_{1:k}, \hat{\boldsymbol{\beta}}_{1:k})$ while accounting for the dependence in $\mathbf{Z}_{1:k}$, through a Bayesian non-parametric estimate of $G_{1:k}$, gives only small improvements relative to our much simpler approach.

At this stage the Gaussian assumption is discarded and a non-parametric estimate of the conditional distribution function $G_{k|1:k-1}$ is formed. This step accounts for the dependence of $\mathbf{Z}_{1:k}$ in the subsequent inference. Our approach to estimating $G_{k|1:k-1}$ is to obtain an estimate of the joint density function $g_{1:k}$ and use this to derive the conditional distribution. For this purpose we use kernel density estimation, similar to Papastathopoulos and Tawn (2013). Under model (10) we have that $\mathbf{Z}_{t+1:t+k} \sim G_{1:k}$ for all t with $X_t > u$. We first derive an estimated identically distributed sample $\hat{\mathbf{z}}_{1:k}^{(i)}, i = 1, \ldots, n_u$ from $\mathbf{Z}_{1:k}$, where n_u is the number of exceedances by $\{Y_t\}$ of u_E. Specifically, let $t_1, \ldots, t_{n_u}$ be the indices of $t = 1, \ldots, n$ where $x_t > u$. By inverting equation (10) we have for $i = 1, \ldots, n_u$

$$\hat{\mathbf{z}}_{1:k}^{(i)} = \frac{\mathbf{x}_{t_i+1:t_i+k} - \hat{\boldsymbol{\alpha}}_{1:k} x_{t_i} - \hat{\boldsymbol{\mu}}_{1:k}(x_{t_i})^{\hat{\boldsymbol{\beta}}_{1:k}}}{\hat{\boldsymbol{\gamma}}_{1:k}(x_{t_i})^{\hat{\boldsymbol{\beta}}_{1:k}}}. \tag{14}$$

For each $j = 1, \ldots, k$, the sample $\hat{\mathbf{z}}_j = (\hat{\mathbf{z}}_j^{(i)}, i = 1, \ldots, n_u)$ has zero mean and unit variance.

Based on the $\hat{\mathbf{z}}_{1:k}$ data, we estimate the joint density $g_{1:k}$ using a multivariate kernel density estimation methods, i.e.,

$$\tilde{g}_{1:k}(\mathbf{z}) = \frac{1}{n_u} \sum_{i=1}^{n_u} K_{\mathbf{H}}\left(\mathbf{z} - \hat{\mathbf{z}}_{1:k}^{(i)}\right), \tag{15}$$

where $K_{\mathbf{H}}$ is the independent multivariate Gaussian kernel function, with $\mathbf{H} = (h_1, \ldots, h_k)$ the vector of the marginal bandwidths. It follows that our estimate of the conditional distribution function $G_{k|1:k-1}$ is

$$\hat{G}_{k|1:k-1}(z \mid \mathbf{z}_{1:k-1}) = \sum_{i=1}^{n_u} w_i \Phi\left(\frac{z - \hat{z}_k^{(i)}}{h_k}\right),$$

where the weights

$$w_i = \prod_{j=1}^{k-1} \phi\left(\frac{z_j - \hat{z}_j^{(i)}}{h_j}\right) \Bigg/ \sum_{r=1}^{n_u} \prod_{j=1}^{k-1} \phi\left(\frac{z_j - \hat{z}_j^{(r)}}{h_j}\right) \qquad i = 1, \ldots, n_u, \quad (16)$$

satisfy $0 \leq w_i \leq 1$, $\sum_{i=1}^{n_u} w_i = 1$, and where ϕ is the standard normal density.

5.2 Inference for cluster functionals

From the estimates of the model parameters we can derive estimates of interesting cluster functionals by using the estimated model to repeatedly simulate the within cluster behaviour of the process. Our strategy for generating within cluster behaviour is to use the properties of the tail chain, where a tail chain describes the nature of the Markov chain after an extreme observation, expressed in the limit as the observation tends to the upper endpoint of the marginal distribution; see Rootzén (1988), Smith (1992), Smith et al (1997), Yun (2000) and Drees et al (2015).

Tail chains were originally developed for asymptotically dependent processes, where for Laplace margins, they correspond to the limiting process of $\{X_t - u; t = 0, 1, \ldots\} | X_0 > u$ as $u \to \infty$. Normalisation of X_t by a subtraction of u leads to non-degeneracy for asymptotic dependence. Denote the limiting tail chain by $\{X_t^+; t = 0, 1, \ldots\}$. It follows that X_0^+ follows an standard Exponential distribution and subsequent values that exceed 0 contribute to the limiting cluster. This tail chain is used to approximate the behaviour of the asymptotically dependent processes during extreme events which exceed a high threshold v, giving the approximation for $\{X_t\}$, subject to $X_0 > v$, being provided $\{X_t^{(v)} := v + X_t^+; t = 0, 1, \ldots\}$. In essence the tail chain is assumed to hold exactly over v. The tail chain has a negative drift and so after sufficient steps no further exceedances of v are possible within the cluster.

Kulik and Soulier (2015) and Papastathopoulos et al (2017) consider tail chains for first-order asymptotically independent processes. Normalising using a subtraction of u leads to degeneracy in this case, and less powerful location-scale norming is required. In fact for kth-order Markov chains with Laplace margins, the required normalisations under the Heffernan and Tawn (2004) model formulation are given by the results in Section 4.2, with both location and scaling required.

To be precise we now set out how to simulate from the kth-order tail chain approximation above threshold v. The tail chain $\{X_t^{(v)}; t = 0, 1, \ldots\}$, is simulated in three steps. First the initial exceedance, $X_0^{(v)}$, of v is simulated as $X_0^{(v)} = v + E_0$ where E_0 is an Exponential(1) random variable. For the kth-order extremal Markov chain the next $k - 1$ values, $\mathbf{X}_{1:k-1}^{(v)}$, of the tail-chain

are generated jointly, conditional on $X_0^{(v)}$, as

$$\mathbf{X}_{1:k-1}^{(v)} = \hat{\boldsymbol{\alpha}}_{1:k-1} X_0^{(v)} + \left(X_0^{(v)}\right)^{\hat{\boldsymbol{\beta}}_{1:k-1}} \mathbf{Z}_{1:k-1}, \qquad (17)$$

where $\mathbf{Z}_{1:k-1}$ is sampled independently from $\hat{g}_{1:k-1}$, the marginal of $\hat{g}_{1:k}$ given by equation (15). At all subsequent time-steps $j \geq k$, the transition kernel of the tail chain is used, i.e.,

$$X_j^{(v)} = \alpha_k X_{j-k}^{(v)} + \left(X_{j-k}^{(v)}\right)^{\beta_k} Z_{j|j-k+1:j-1}, \qquad (18)$$

where the $Z_{j|j-k+1:j-1}$ values are sampled independently from $\hat{G}_{k|1:k-1}$. The final stage is to transform the simulated tail chain back to the original margins, i.e., $Y_t^{(v)} = F_Y^{-1}(F_X(X_t^{(v)}))$ for $t = 0, 1, \ldots$, and so when $X_t^{(v)} > u$ then $Y_t^{(v)} > u_Y$.

The justification for the use of the asymptotically motivated transition kernel (18) is that $X_{j-k}^{(v)} > u$. When $X_{j-k}^{(v)} < u$ the algorithm can still be used but the quality of the approximation is likely to become poor the further $X_{j-k}^{(v)}$ drops below u. When $X_{j-k}^{(v)} < 0$ the transition (18) cannot be used as $\beta_k < 1$ and so the tail chain is immediately terminated before generating $X_j^{(v)}$. This is not restrictive as $X_{j-k}^{(v)} < 0$ corresponds to the process falling below the median and so it is reasonable to treat the extreme event as having finished. The tail chain is run until $X_j^{(v)}$ is small enough that there is a negligible probability of obtaining a further exceedance of v. In practice it is most simple to run the chain for a fixed long time m, with $m = 40$ found to be sufficient for our examples in Section 7.

For first-order processes, Winter and Tawn (2016) showed that it is equally easy to simulate the tail chain forwards and backwards from a cluster maximum M, given that $M > v$, or to simulate an arbitrary exceedance of v and simulate forwards only. For kth-order tail chains it is more computationally efficient to use the latter approach since it requires only forward simulation and does not require the initial simulated value to be the cluster maximum, which would lead to the simulation scheme requiring rejection methods to ensure this property. As all the cluster functionals that we are interested in can be evaluated using only forwards simulation we restrict attention to this approach.

There are a range of cluster functionals that we are interested in estimating. The extremal dependence measure χ_τ, defined by expression (2), is not helpful as for all processes that are asymptotically independent at lag τ then $\chi_\tau = 0$. However the sub-asymptotic extremal dependence measure $\chi_\tau(v)$ for $\tau = 1, 2, \ldots$, i.e.,

$$\chi_\tau(v) = \mathrm{P}(X_\tau > v \mid X_0 > v) = \mathrm{P}(Y_\tau > y_v \mid Y_0 > y_v), \qquad (19)$$

where $y_v = F_Y^{-1}(F_X(v))$, provides a helpful summary of the level of dependence in the tail. This can be evaluated as the proportion of tail chains which are above the threshold j steps after an exceedance. We are also interested in the distribution $\pi(i, v)$ of D_v, the number of exceedances of v, in a cluster. To be precise

$$D_v = \sum_{t \in C} \mathbb{I} \left(X_t - v \right)_+,$$

where $\mathbb{I}(.)$ is the indicator function and C is a set of values constituting a cluster and $\pi(i, v) = \mathrm{P}(D_v = i \mid D_v > 0)$. Furthermore we use the notation $\Pi(i, v) = \mathrm{P}(D_v \geq i \mid D_v > 0)$ to denote the probability of a cluster above v with at least i exceedances.

The mean of the cluster size distribution $\pi(\cdot, v)$ is the most widely used cluster functional of the extremes in a time-series at a level v, i.e.,

$$\sum_{i=1}^{\infty} i\pi(i, v) = \theta(v)^{-1}, \tag{20}$$

which is given by the reciprocal of the sub-asymptotic extremal index $\theta(v)$ (Ledford and Tawn 2003), with $\theta(v) \in [0, 1]$. See Winter and Tawn (2016) for details of how to use the forward tail chain for estimating these functionals.

When reporting results it is more instructive to give estimates of the probability of a cluster functional value occurring within a given time period rather than in a single event. To make such a conversion we need to account for there being a Poisson, mean ψ_v, number of independent and identically distributed clusters of the level v in the time period, where for $v \geq u$

$$\psi_v = \theta(v)[1 - \tilde{F}(u)]n_T \left[1 + \xi \left(\frac{v - u}{\sigma_u} \right) \right]_+^{-1/\xi},$$

where n_T is the number of observations in the period (Winter and Tawn 2016). The expression for ψ_v can be seen to be the expected number of exceedances of v if the series was independent multiplied by $\theta(v)$. As $\theta(v)$ is the ratio of the probability of a cluster divided by the probability of an exceedance, this converts ψ_v to be the expected number of clusters in the period. To illustrate the type of calculation involved, consider the probability of observing at least one cluster with at least i days above the critical level v occurring in the time period. By averaging over the Poisson number of clusters it follows that this probability is $1 - \exp\{-\psi_v \Pi(i, v)\}$.

6 Selection of the order of the extremal Markov process

We have presented an approach that enables inference for cluster functionals for a kth-order extremal Markov process when k is known. Here we present

methods for the selection of the order of the process when k is unknown. We denote the order of our modelled process by τ and the true order by k. If $\tau > k$ then the inclusion of unnecessary higher-order information introduces extra parameters than are required which leads to higher than necessary variability in our cluster functional inferences. If $\tau < k$ we may not adequately capture the extremal dependence structure which will lead to biases in our estimates of cluster functionals. Therefore, we are interested in developing diagnostic methods to select τ so that it is as close as possible to k. Inference methods for first order processes already exist, i.e., testing a null hypothesis of $k = 0$ against and alternative of $k = 1$ (Winter and Tawn 2016). We want to test whether incorporating higher order structure into the modelling leads to significantly different inferences, i.e., testing if $k = 1$ or $k > 1$.

A standard approach to estimate the order of a Markov chain is to identify the largest lag at which the partial auto-correlation function (PACF) is deemed to be significantly different from zero (Chatfield 2003), since this function, at lag j, gives the strength of the dependence between $(X_t, X_{t+j}) \mid \mathbf{X}_{t+1:t+j-1}$. However this is not necessarily appropriate for an extremal Markov process as the PACF inference is dominated by the data in the body of the distribution and extremal data may exhibit either more or less complex behaviour than data from the body of the distribution. The PACF is helpful as one of a set of diagnostics, but we require other diagnostics which focus more explicitly on the extreme events.

Our new diagnostic methods for the selection of the order τ of the extremal Markov process are motivated by standard univariate threshold selection diagnostics. Essentially these diagnostics are equivalent to using a threshold stability plot which assesses the stability of extremal parameters relative to a range of thresholds (Coles 2001). Here we compare the stability of estimates of the cluster functionals, discussed in Section 5.2, over a range of τ. For the inference of any cluster functional all the data are used from the τ-tuples $(X_t, \ldots, X_{t+\tau-1})$, with $X_t > u$, where u is the modelling threshold. We take our selected value, $\hat{\tau}$, to be the lowest value of τ above which the estimated cluster functionals are stable, other than for sampling variability. This exploits the changing bias in these cluster functionals as τ increases when $\tau < k$ and that there is no bias for all $\tau \geq k$ but simply increasing variance in the cluster functional estimates. This leads to a value of $\hat{\tau}$ that if decreased (increased) cause the estimates of cluster functional to change (not change) relative to the variability of the estimates.

The method can be applied to a wide range of cluster functionals and a selection for $\hat{\tau}$ made based on average values for the best values of τ for the different cluster functionals. Alternatively, the approach allows $\hat{\tau}$ to be selected for the most important cluster functional which depends on the context of the problem. For our analysis of heatwaves in Section 7, one such type of cluster functional is the probability of short, medium and long runs of ex-

ceedances of a high level.

If no particular cluster functional is required then we have found that the threshold dependent extremal dependence measure $\chi_j(v)$ and the sub-asymptotic extremal index $\theta(v)$, introduced in equations (19) and (20) respectively, are particularly helpful for use in the selection of τ, particularly when studied over a range of j and v. To be specific how we use the diagnostic methods, consider these two cluster functionals explicitly. Let $\tilde{\chi}_j(v)$ and $\tilde{\theta}(v)$ be the respective empirical estimates of $\chi_j(v)$ and $\theta(v)$. Here $\tilde{\chi}_j(v)$ is the proportion of pairs (X_t, X_{t+j}) with $X_t > v$ that also have $X_{t+j} > v$ and $\tilde{\theta}(v)$ is the runs estimator of Smith and Weissman (1994) using the optimal run length given by Ferro and Segers (2003). These empirical estimates are reliable only over a subset R of j and v, e.g., v and j not too large. Over R we pick n_J and n_V equally spaced values for j and v respectively. Under a fitted τth-order extremal Markov model we denote the respective estimates of these cluster functionals by $\hat{\chi}_j^{(\tau)}(v)$ and $\hat{\theta}^{(\tau)}(v)$, where these estimates are derived using the methods of Section 5.2. Then the best estimate of the order of the Markov process is $\hat{\tau}$, where

$$\hat{\tau} = \min \left\{ \tau > 0 : \sum_{v \in R} \{ |\hat{\theta}^{(\tau)}(v) - \tilde{\theta}(v)| + \frac{1}{n_J} \sum_{j \in R} |\hat{\chi}_j^{(\tau)}(v) - \tilde{\chi}_j(v)| \} < \epsilon \right\},$$

for a choice of $\epsilon > 0$.

When v is sufficiently large that $\tilde{\theta}(v)$ is unreliable, we also assess the performance of $\hat{\chi}_j^{(\tau)}(v)$ using the unconstrained pairwise conditional model of Heffernan and Tawn (2004) for $X_{t+j} \mid X_t$ when $X_t > u$. We denote the corresponding estimate by $\hat{\chi}_j(v)$, for $v \geq u$, and term it the unconstrained parametric estimate.

The diagnostic methods we propose for selecting the order of the extremal process do not check formally whether the inclusion of higher order structure leads to statistically significant differences. Winter and Tawn (2016) presented a likelihood ratio test for $k = 1$ against $k = 0$. So here we present a test for $k > 1$ against $k = 1$, by testing $k = \tau$ against $k = 1$, for range of $\tau > 1$. Reich et al (2014) perform such a test for their asymptotically dependent model. If the τth-order approach is found to obtain a significantly better fit than the first-order approach, a natural next step is to ask whether the τth-order result is a better fit than the jth-order result for all $j = 2, \ldots, \tau - 1$. Such a set of nested tests exists when modelling time-series using AR models (Brockwell and Davis 2006), but has not been developed yet in our context.

Details of our likelihood ratio test are as follows. If the extremal process is believed to be of order τ, we fit the model as described in Section 5.1 with $k = \tau$, maximising likelihood (13) over $(\boldsymbol{\alpha}_{1:\tau}, \boldsymbol{\beta}_{1:\tau}, \boldsymbol{\mu}_{1:\tau}, \boldsymbol{\gamma}_{1:\tau})$, and denoting

the maximised likelihood for model (12), by L_τ. If the extremal process is believed to be of order 1 then the parameters $(\boldsymbol{\alpha}_{1:\tau}, \boldsymbol{\beta}_{1:\tau})$ simplify under the results of Papastathopoulos et al (2017). As identified in Section 4.1, there are two possible forms for $(\boldsymbol{\alpha}_{1:\tau}, \boldsymbol{\beta}_{1:\tau})$ with either

$$\alpha_j = \alpha_1^j, \beta_\tau = \beta_1 \text{ when } 0 < \alpha_1 \leq 1 \tag{21}$$

or

$$\alpha_j = 0, \beta_\tau = \beta_1^j \text{ when } \alpha_1 = 0. \tag{22}$$

We maximise likelihood (13) over $(\boldsymbol{\alpha}_{1:\tau}, \boldsymbol{\beta}_{1:\tau}, \boldsymbol{\mu}_{1:\tau}, \boldsymbol{\gamma}_{1:\tau})$ with $(\boldsymbol{\alpha}_{1:\tau}, \boldsymbol{\beta}_{1:\tau})$ under the different sets of constraints given by expressions (21) and (22). We denote the respective maximised likelilhoods by $L_{1,a}$ and $L_{1,b}$. We then use the test statistics $D_a = 2\log(L_\tau/L_{1,a})$ and $D_b = 2\log(L_\tau/L_{1,b})$. Under the null hypothesis that $k = 1$, following standard likelihood methods these test statistics follow a chi-squared distribution on $2\tau - 2$ and $2\tau - 1$ degrees of freedom respectively. To counteract any problems associated with multiple testing, the Bonferroni correction is used (Dunn 1961).

7 Data analysis

Daily maximum temperature observations are taken at Orleans, in central France, for the period 1946-2012. Four missing values exist in the time-series and are omitted, none occur during the 2003 heatwave event that we focus aspects of our analysis on. Heatwaves are most likely to occur in summer months. The temperature data from the three month period of June-August are extracted from each year and exploratory analysis suggests that they form an approximately stationary time-series. Given that extreme hot temperature days are unlikely to occur outside this period the return levels that are estimated for the summer season correspond to the yearly return levels.

Throughout this section, unless stated otherwise, the critical level for extreme events is set at the one-year return level, denoted by v_1. In 2003 there were two events, of length 2 and 11, above the critical level v_1 during a four week period. Winter and Tawn (2016) used a first-order extremal Markov chain to estimate the probability of an event that lasts at least as long as each of these 2003 events. We are interested in how unlikely such events are when estimated using higher-order extremal Markov chains, and how sensitive these estimates are to the selected order of the chain.

First, a GPD is fitted to exceedances of the modelling threshold u_Y, where u_Y was chosen to be 29.7°C, based on standard diagnostics (Coles 2001). Diagnostic plots for this data set and justification of the GPD model and threshold choice are given in Winter and Tawn (2016). The estimated threshold exceedance probability is $1 - \tilde{F}(u_Y) = 0.099$ (0.007), with estimated GPD scale and shape parameters $\hat{\sigma}_u = 3.002$ (0.225) and $\hat{\xi} = -0.215$ (0.033); the standard errors are given in the parentheses. It follows that the estimated one-year

| j	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\chi}_j(v_1)$
1	0.713 (0.072)	0.524 (0.094)	0.508 (0.027)
2	0.576 (0.080)	0.538 (0.126)	0.276 (0.042)
3	0.440 (0.084)	0.514 (0.163)	0.186 (0.041)
4	0.342 (0.083)	0.400 (0.182)	0.144 (0.037)
5	0.395 (0.082)	0.301 (0.201)	0.117 (0.031)
6	0.288 (0.077)	0.286 (0.226)	0.095 (0.026)
7	0.313 (0.069)	0.253 (0.210)	0.076 (0.018)
8	0.259 (0.053)	0.280 (0.193)	0.067 (0.016)
9	0.198 (0.040)	0.091 (0.158)	0.036 (0.011)
10	0.162 (0.037)	-0.061 (0.143)	0.019 (0.008)

Table 1 Estimates for the extremal dependence parameters (α_j, β_j) and estimated extremal dependence measure $\hat{\chi}_j(v_1)$ for a set of different lag values $j = 1, \ldots, 10$ at the one year return level v_1. The estimates of $\chi_j(v)$ are obtained from the pairwise model for $X_{t+j} \mid X_t$. Standard errors are given in parentheses.

return level is $v_1 = 35^\text{o}\text{C}$.

Before we examine the precise form of the extremal Markov process we investigate how the extremal dependence decays with lag. Figure 1 (left) shows that the auto-correlation function for the Orleans daily maximum temperature data decays monotonically in a near exponential form. However, this estimate is dominated by the data in the body of the series and so may not reflect the extremal dependence.

To assess if this feature is observed for the extreme events we examine the extremal dependence parameters $(\boldsymbol{\alpha}_{1:10}, \boldsymbol{\beta}_{1:10})$. Estimation uses a threshold corresponding to the 90% marginal quantile, with this level being selected based on the diagnostics proposed by Heffernan and Tawn (2004). Estimates (and standard errors in parentheses) of these parameters are given in Table 1. These estimates are obtained without making any extremal Markov process assumptions. The estimated values $\hat{\boldsymbol{\alpha}}_{1:10}$ are all statistically significantly different from one and zero, indicating the positive dependence form of asymptotic independence at lags $1 - 10$. Furthermore, there is a general pattern of the values decreasing with lag, though it is not entirely monotone. The $\hat{\boldsymbol{\beta}}_{1:10}$ exhibit a similar pattern. As $(\hat{\boldsymbol{\alpha}}_{1:10}, \hat{\boldsymbol{\beta}}_{1:10})$ are often correlated it is helpful to also consider a cluster functional estimate as the extremal dependence parameters combine to produce these. Here we examine the extremal dependence measure $\chi_j(v_1); j = 1, \ldots, 10$ estimated using the unconstrained parametric estimate, where $\hat{\chi}_j(v)$ is as described in Section 6. Table 1 shows that $\hat{\chi}_j(v_1)$ decreases monotonically with increasing lag, so the pattern of dependence decay is similar for both typical and extreme values.

Figure 1 (right) presents the PACF for the Orleans' daily maximum temperature data, which shows a large spike in the PACF at lag 1 with smaller values at all larger lags. This diagnostic was used by Winter and Tawn (2016) to
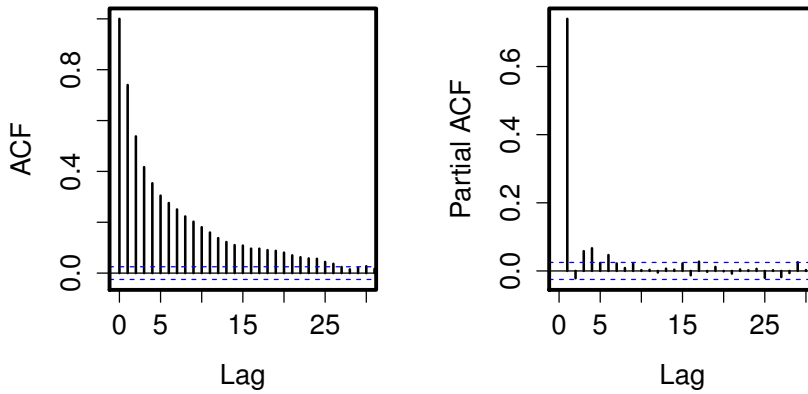
Fig. 1 Auto-correlation function (left) and partial auto-correlation function (right) for Orleans daily maximum temperature data. Dashed intervals represent a 95% tolerance intervals.

motivate their choice of a first-order Markov model. However, there are some values of the PACF that lie outside the 95% tolerance intervals up to lag 6. These values suggest that a first-order Markov model might omit some important higher-order structure and, as discussed in Section 6, this diagnostic may miss features of the extremal process.

We wish to examine whether there is statistically significant evidence for higher-order dependence than first-order for the process when in an extreme state, defined here to be when the process exceeds the 90% marginal quantile. A hypothesis test is constructed to test whether a τth-order dependence structure provides a significantly better fit than a first-order approach. However, if the null hypothesis is rejected this only suggests that the true order of the extremal Markov process is greater than or equal to 2. The test is explained in Section 6. Under a first-order model the parameters $(\boldsymbol{\alpha}_{2:10}, \boldsymbol{\beta}_{2:10})$ are constrained to satisfy either condition (21) or (22), whereas for the τth-order model $(\boldsymbol{\alpha}_{2:\tau}, \boldsymbol{\beta}_{2:\tau})$ are unconstrained. Tests are constructed for $\tau = 2, \ldots, 10$ and using Bonferroni bounds the significance level is set at $0.05/9$. All tests for which $\tau \geq 7$ were found to be significant at the 5% significance level.

Section 5 set out that our main diagnostic for the selection of the order of the extremal Markov process is a comparison of estimates of $\chi_j(v)$ for $j \geq 1$. We have two reference estimates to compare our extremal Markov models to: non-parametric estimates $\tilde{\chi}_j(v)$ when v is low enough that empirical estimates of $\chi_j(v)$ are reliable and the unconstrained parametric estimates $\hat{\chi}_j(v)$ for larger v. In each case we compare these estimates with τth-order extremal Markov model estimates $\hat{\chi}_j^{(\tau)}(v)$.

If the process is a kth-order extremal Markov process then we should find
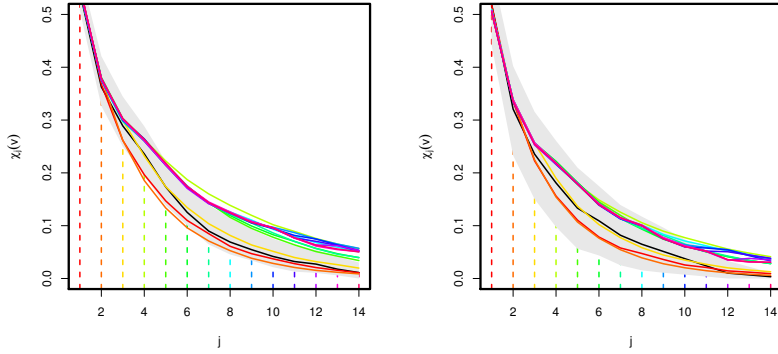
Fig. 2 Estimates of the threshold dependent extremal measure $\chi_j(v)$ using empirical approach (black) and different order extremal Markov chains (rainbow) with v set at 90% (left) and 95% (right) quantiles respectively. Grey shaded region corresponds to 95% confidence interval for empirical obtained via a block bootstrap approach.

that $\hat{\chi}_j^{(k)}(v)$ is close to $\tilde{\chi}_j(v)$ for all j for low v and to $\hat{\chi}_j(v)$ for high v. Furthermore as

$$\hat{\chi}_j^{(\tau_2)}(v) = \hat{\chi}_j^{(\tau_1)}(v) \text{ for all } j \leq \tau_1 < \tau_2,$$

our ability to distinguish between models of orders $\tau_1 < \tau_2$ is only through the values of $\hat{\chi}_j^{(\tau_2)}(v)$ and $\hat{\chi}_j^{(\tau_1)}(v)$ for $j > \tau_1$. Consequently in the plots of $\hat{\chi}_j^{(\tau)}(v)$ against j in Figure 2 we select a different colour when $j > \tau$.

Figure 2 plots these diagnostics for v corresponding to the marginal 90% and 95% quantiles (denoted $v_{0.9}$ and $v_{0.95}$). With $v_{0.9}$ it appears that the third-order scheme comes closest to the pattern observed in the empirical estimates. First- and second-order schemes seem to underestimate the strength of the dependence whereas higher-order estimates seem to lead to an overestimation, reflecting their greater variation. Similar patterns are found for $v_{0.95}$, although the higher-order schemes seem to be contained within the empirical confidence bands for higher values of j due to the increased uncertainty in the empirical estimate. Figure 3 shows the diagnostic for $v = v_1$, which suggests that lower-order schemes are picking up the general behaviour better, being contained with the confidence intervals at all values of j. However, the higher-order schemes do seem to pick up some higher-order structure that is present in the original data set that is missed by a lower-order scheme. Taking all the diagnostics into account, we conclude that the third-order scheme seems to provide the most reliable estimates of $\chi_j(v)$ at all levels.

The cluster functionals that are of most importance for heatwaves are $\theta(v_1)$, the reciprocal of the mean cluster length, and $\Pi(2, v_1), \Pi(6, v_1), \Pi(11, v_1)$, the probabilities of a cluster with at least 2, 6 and 11 exceedances of v_1. The prob-
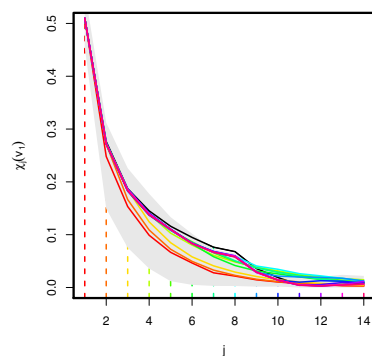
Fig. 3 Estimates of the threshold dependent extremal measure $\chi_j(v_1)$ using unconstrained parametric estimates (black) and different order extremal Markov chains (rainbow). Grey shaded region corresponds to 95% confidence interval for unconstrained parametric approach obtained via a block bootstrap approach.
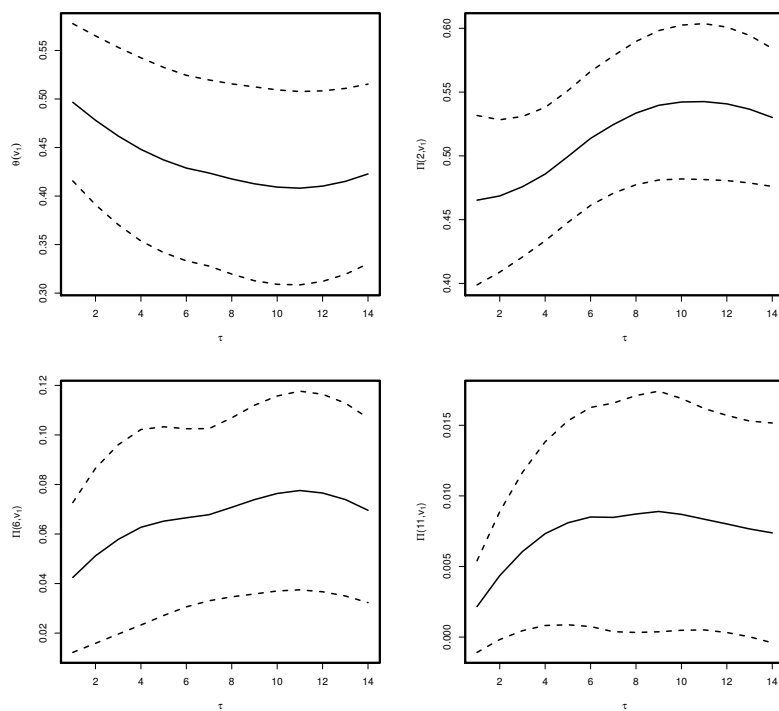
Fig. 4 Estimates of within cluster extremal quantities for different higher-order schemes with v set at the one-year return level v_1. Cluster functions are $\theta(v_1), \Pi(2, v_1), \Pi(6, v_1)$ and $\Pi(11, v_1)$. Modified bootstrapping approach used to obtain 95% confidence intervals (dotted). Estimates have been smoothed using loess method for clarity.

abilities of short and long events are of particular interest as these correspond the observed durations of the 2003 European heatwave at Orleans. We wish to assess the sensitivity of these cluster functional estimates to the choice of the order of the extremal process.

Figure 4 shows estimates of these cluster functionals for orders of $\tau = 1 \ldots, 14$ for the extremal Markov model. As explained in Section 6, we aim to identify the lowest order for which these cluster functionals remain constant at all higher orders, other than for sampling variability. The uncertainty bounds used in this figure are obtained via the block bootstrap. As the accurate evaluation of cluster functional is computationally intensive it is not feasible to run many bootstrap replications. Instead we run a reduced number of replications to approximate the standard error for the cluster functional sampling distribution and then construct symmetric confidence intervals around the point estimate using this standard error.

The estimates of the average duration of a heatwave and the probabilities of short, median and long clusters all increase when a higher-order extremal Markov chain is used. Typically the estimates increase rapidly until $\tau = 3$, continue to increase until $\tau = 7$ and then stabilise. However, this pattern is somewhat masked by the confidence intervals which broadly cover all estimates at all orders, so there is limited information about choice of the order. Of course we could have used a lower value of the critical level than v_1 which would have been better for diagnostic purposes, but would not have shown the sensitivity issues of the features most relevant in practice. But even at level v_1 we find that for probabilities $\Pi(2, v_1)$ our diagnostic suggests that we need $\tau \geq 3$.

Finally we focus on estimating the probability of particular cluster functionals occurring in a year. To take into account the uncertainty that we found in the selection of the order of the extremal Markov process, we compare estimates using $\tau = 1, 3$ and 7. For $\tau = 1$ the estimated probability of observing at least one event in a year that lasts at least 2 days as 0.208 (0.200, 0.216); for 11 days the equivalent probability is 0.001 (1×10^{-4}, 0.004), equivalent to the 1000 year return level. When $\tau = 3$ these estimates are 0.196 (0.171, 0.221) and 0.002 (0, 0.004) respectively. The equivalent probabilities for $\tau = 7$ are 0.201 (0.179, 0.224) and 0.003 (0, 0.007) respectively. Thus it appears that the inclusion of higher order structure does not greatly affect the probability of smaller events but can lead to a 3-fold increase in the point estimates of the probability of very long duration extreme events. As expected, uncertainty estimates are wider for the higher-order approaches, reflecting the increased number of parameters to be estimated.

8 Discussion and conclusion

This paper provides a new framework for incorporating higher-order Markov models for temporal dependence when modelling extreme events covering processes which can be either asymptotically dependent or asymptotically independent. For this purpose we have developed a kth-order extremal Markov model framework for incorporating higher-order information using the conditional extremes approach of Heffernan and Tawn (2004). Such an approach is motivated by an application to heatwave events, since all the existing time series extremes models, which have been developed under assumptions of either a first-order Markov model or that the variables are asymptotically dependent, do not adequately capture the properties we observe for heatwaves.

Our results show that using standard time series diagnostics to identify the order of an extremal Markov process can lead to errors when interest is restricted to the extremes of the process. This necessitated the development of a range of new diagnostics for choosing the 'best' order scheme to use for extreme events. Specifically, in our example the use of standard time-series diagnostics ignored structure in the extremes which leads to the underestimation in the probability of longer and potentially devastating heatwave events. One area for further work is to formalise and unify our range of heuristic diagnostic methods for estimating the order of the extremal Markov process. To help achieve this a systematic study of the performance of these methods in a simulation study is needed. This study should cover both asymptotically independent and asymptotically dependence kth-order Markov processes, each with varying strengths of dependence.

As in Winter and Tawn (2016), daily maximum temperatures have been analysed instead of looking at the joint distribution of daily maximum and minimum temperatures. An extremal Markov model would still be appropriate in such a situation but a different order scheme might be required and stationarity be imposed separately on the maximum and minimum temperature series. The effect of climate change and other large scale climatic phenomena have not been incorporated into this paper. Winter et al (2016) illustrate how the tail chain simulation approach with first-order dependence structure can be altered to take into account the effect of covariates. A similar extension could be applied to the methodology outlined here.

References

Bortot P, Tawn JA (1998) Models for the extremes of Markov chains. Biometrika 85(4):851–867

Brockwell PJ, Davis RA (2006) Introduction to Time Series and Forecasting. Springer

Chatfield C (2003) The Analysis Of Time Series: An Introduction. CRC Press

Coles SG (2001) An Introduction to Statistical Modeling of Extreme Values. Springer Verlag

Coles SG, Tawn JA (1991) Modelling Extreme Multivariate Events. Journal of the Royal Statistical Society: Series B 53(2):377–392

Davison AC, Smith RL (1990) Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society: Series B 52(3):393–442

Davison AC, Padoan SA, Ribatet M (2012) Statistical modeling of spatial extremes. Statistical Science 27(2):161–186

Drees H, Segers J, Warchol M (2015) Statistics for tail processes of Markov chains. Extremes 18:369–402

Dunn OJ (1961) Multiple comparisons among means. Journal of the American Statistical Association 56(293):52–64

Dupuis DJ (2012) Modeling waves of extreme temperature: the changing tails of four cities. Journal of the American Statistical Association 107(497):24–39

Fawcett L, Walshaw D (2006) Markov chain models for extreme wind speeds. Environmetrics 17(8):795–809

Ferro CAT, Segers J (2003) Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B 65(2):545–556

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

de Haan L, Ferreira A (2006) Extreme Value Theory: An Introduction. Springer, New York

Heffernan JE (2000) A directory of coefficients of tail dependence. Extremes 3:279–290

Heffernan JE, Resnick SI (2007) Limit laws for random vectors with an extreme component. The Annals of Applied Probability 17(2):537–571

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

Hitz A, Evans R (2015) One-component regular variation and graphical modeling of extremes. ArXiv e-prints 1506.03402

Joe H (1997) Multivariate Models and Dependence Concepts. Chapman and Hall/CRC

Keef C, Papastathopoulos I, Tawn JA (2013) Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. Journal of Multivariate Analysis 115:396–404

Kulik R, Soulier P (2015) Heavy tailed time series with extremal independence. Extremes 18:1–27

Ledford AW, Tawn JA (1997) Modelling dependence within joint tail regions. Journal of the Royal Statistical Society: Series B 59(2):475–499

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

Liang KL, Self SG (1996) On the asymptotic behaviour of the pseudolikelihood ratio test statistic. Journal of the Royal Statistical Society: Series B 58(4):785–796

Lugrin T, Davison A, Tawn JA (2016) Bayesian uncertainty management in temporal dependence of extremes. Extremes 19:491–515

Papastathopoulos I, Tawn JA (2013) Graphical structures in extreme multivariate events. In: Proceedings of 25th Panhellenic Statistics Conference, pp 315–323

Papastathopoulos I, Strokorb K, Tawn JA, Butler A (2017) Extreme events of Markov chains. Advances in Applied Probability 49

Pickands J (1971) The two–dimensional Poisson process and extremal processes. Journal of Applied Probability 8:745–756

Reich BJ, Shaby BA, 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 19:119–135

Resnick SI (1987) Extreme Values, Regular Variation, and Point Processes. Springer Verlag

Ribatet M, Ouarda TBMJ, Sauquet E, Gresillon JM (2009) Modeling all exceedances above a threshold using an extremal dependence structure: Inferences on several flood characteristics. Water Resources Research 45:W03,407

Rootzén H (1988) Maxima and exceedances of stationary Markov chains. Advances in Applied Probability 20:371–390

Smith RL (1992) The extremal index for a Markov chain. Journal of Applied Probability 29(1):37–45

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

Smith RL, Tawn JA, Coles SG (1997) Markov chain models for threshold exceedances. Biometrika 84(2):249–268

Wadsworth JL, Tawn JA (2013) A new representation for multivariate tail probabilities. Bernoulli 19(5B):2689–2714

Winter HC, Tawn JA (2016) Modelling heatwaves in central France: a case study in extremal dependence. Journal of the Royal Statistical Society: Series C 65(3):345–365

Winter HC, Tawn JA, Brown SJ (2016) Detecting changing behaviour of heatwaves with climate change. Submitted

Yun S (2000) The distributions of cluster functionals of extreme events in a dth-order Markov chain. Journal of Applied Probability 37(1):29–44