

Stochastic Environmental Research and Risk Assessment

Investigating the properties of Bayesian posterior predictive return levels for environmental extremes

--Manuscript Draft--

Manuscript Number:	SERR-D-17-00163R1
Full Title:	Investigating the properties of Bayesian posterior predictive return levels for environmental extremes
Article Type:	Original research
Keywords:	Bayesian inference; Extreme value theory; Predictive return levels; Wind speeds
Corresponding Author:	Lee Fawcett, Ph. D. Newcastle University Newcastle upon Tyne , UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Newcastle University
Corresponding Author's Secondary Institution:	
First Author:	Lee Fawcett, Ph. D.
First Author Secondary Information:	
Order of Authors:	Lee Fawcett, Ph. D. Amy Charlotte Green
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	A key aim of most extreme value analyses is the estimation of the r-year return level; the wind speed, or sea-surge, or rainfall level (for example), we might expect to see once (on average) every r years. There are compelling arguments for working within the Bayesian setting here, not least the natural extension to prediction via the posterior predictive distribution. Indeed, for practitioners the posterior predictive return level has been cited as perhaps the most useful point summary from a Bayesian analysis of extremes, and yet little is known of the properties of this statistic. In this paper, we attempt to assess the performance of predictive return levels relative to their estimative counterparts obtained directly from the return level posterior distribution; in particular, we make comparisons with the return level posterior mean, mode and 95% credible upper bound. Differences between the predictive return level and standard summaries from the return level posterior distribution, for wind speed extremes observed in the UK, motivates this work. A large scale simulation study then reveals the superiority of the predictive return level over the other posterior summaries in many cases of practical interest.
Response to Reviewers:	Please refer to the file "refresponse.pdf"

We would like to thank the Editor, and the reviewers, for their excellent comments. We believe their input has resulted in a much improved manuscript. Please see the file "refresponse.pdf" for our detailed point-by-point responses to the reviewers' original comments / queries. In the revised version of the paper, new text has been highlighted in magenta.

On publication, all code will be made available via a dedicated repository.

Stoch Environ Res Risk Assess manuscript No.
(will be inserted by the editor)

Investigating the properties of Bayesian posterior predictive return levels for environmental extremes

Lee Fawcett · Amy C. Green

Received: Apr 2017 / Accepted: date

Abstract A key aim of most extreme value analyses is the estimation of the r -year *return level*; the wind speed, or sea-surge, or rainfall level (for example), we might expect to see once (on average) every r years. There are compelling arguments for working within the Bayesian setting here, not least the natural extension to prediction via the posterior predictive distribution. Indeed, for practitioners the *posterior predictive return level* has been cited as perhaps the most useful point summary from a Bayesian analysis of extremes, and yet little is known of the properties of this statistic. In this paper, we attempt to assess the performance of predictive return levels relative to their *estimative* counterparts obtained directly from the return level posterior distribution; in particular, we make comparisons with the return level posterior mean, mode and 95% credible upper bound. Differences between the predictive return level and standard summaries from the return level posterior distribution, for wind speed extremes observed in the UK, motivates this work. A large scale simulation study then reveals the superiority of the predictive return level over the other posterior summaries in many cases of practical interest.

Keywords Bayesian inference · Extreme value theory · Predictive return levels · Wind speeds

1 Introduction

Estimating extremes of environmental phenomena such as wind speed, sea-surge and rainfall plays an important role in structural design. Models from classical extreme value theory, such as the generalised extreme value (GEV) or gen-

eralised Pareto (GP) distributions (for block maxima or excesses over a high threshold respectively), give us limiting models for the tail behaviour of such variables and provide a general template for modelling and extrapolation. The aim of most practical applications is the estimation of the event we might expect to see, on average, once every r years: the so-called r -year *return level*, commonly notated as z_r (with estimate $\hat{z}_r$). Over the last three decades or so, pragmatic solutions to circumvent departures from the ideal of independent and identically distributed observations on extremes have been developed; see, for example, the seminal paper by [Davison and Smith \(1990\)](#). Some of the most commonly-used solutions result in sample size reduction; for example, the use of filtering schemes to avoid issues of **temporal dependence** (e.g. peaks over thresholds, or POT), or using only those extremes from within a particular calendar unit to avoid problems associated with seasonal variability. Such a reduction can result in extremely wide confidence intervals for z_r – sometimes giving confidence bounds which are implausible, at a practical level, for the variable being studied. The aim of some recent work, then, has been to investigate the use of methods which maximise the number of extremes pressed into use; see, for example, [Eastoe and Tawn \(2012\)](#) and [Fawcett and Walshaw \(2012, 2016\)](#), the latter illustrating methods which can substantially reduce return level estimation uncertainty relative to methods such as POT.

More recently, much focus has been given to the extension, and practical application, of the theory of multivariate extremes, often motivated by the need to account for spatial dependence between extremes. [Davison et al. \(2012\)](#) provide a comprehensive coverage of the development of models for extremes occurring spatially, including max-stable models: again, the aim is for the estimation of return levels, albeit on maps over a spatial grid. The increasing sophistication of models to address issues such as **temporal dependence**,

L. Fawcett
School of Mathematics, Statistics & Physics, Newcastle University,
Newcastle upon Tyne NE1 7RU, UK
Tel.: +44-191-2087228
E-mail: lee.fawcett@ncl.ac.uk

seasonal variability and trend, coupled with an increase in dimensionality required of an analysis which is – for example – spatial in flavour, often makes inference within a maximum likelihood setting difficult (if not altogether infeasible) and thus researchers might appeal to simulation-based procedures within the Bayesian paradigm to make inferences on model parameters.

The Bayesian paradigm does not just provide a more convenient and accessible inferential framework; it also allows the incorporation of external sources of information through the prior distribution, an obvious element of appeal for any analyst of extremes working with scarce data. Also appealing is the natural extension within the Bayesian framework to prediction; as Fawcett and Walshaw (2016) discuss, an estimate of the r -year posterior predictive return level, $\hat{z}_{r,\text{pred}}$, provides practitioners with a single point summary capturing estimation uncertainty and natural variability. Given the virtues of a Bayesian analysis of extremes, it is then surprising that more publications in the field do not work within this framework. Certainly, it is our experience that few practitioners will perform analyses within a Bayesian setting.

Coles and Powell (1996) provide a solid review of Bayesian inference for extremes up to that date. Since then, Coles and Tawn (1996) and Smith and Walshaw (2003) have investigated the merits of expertly-elicited priors whereas Beirlant *et al.* (2004) and Eugenia Castellanos and Cabras (2007) have considered objective priors for the GEV and GP models. Various authors have used the Bayesian paradigm to exploit meteorological structure in their data via hierarchical models for extremes – for example, Fawcett and Walshaw (2006), Sang and Gelfand (2009, 2010) and Davison *et al.* (2012). Smith (1999) compares predictive inference under the Bayesian and frequentist paradigms and Coles and Tawn (1996) give some informal comparisons between predictive return levels and estimates based solely on the posterior distribution for z_r . Fawcett and Walshaw (2006; 2008; 2016) demonstrate predictive inference for return levels of wind speed and sea-surge extremes, recommending $\hat{z}_{r,\text{pred}}$ as the most convenient, and useful, representation of a return level for practitioners. However (as far as we are aware) no published work supports this through a formal investigation into the performance of the predictive return level. The main contribution of this paper, then, is to explore the properties of $\hat{z}_{r,\text{pred}}$. In particular, we focus on a comparison of the exceedance probabilities of $\hat{z}_{r,\text{pred}}$ to their intended values r^{-1} , and the general performance of $\hat{z}_{r,\text{pred}}$ relative to other estimative summaries obtained directly from the posterior distribution for z_r .

This paper is organised as follows. In Section 2 we give some practical motivation for this work, including some results from an analysis of wind speed extremes at a loca-

tion in the southwest of the UK. This Section will include a primer in extreme value techniques and associated modelling procedures for readers who might be unfamiliar with this area, with a particular focus on recently-proposed methods for handling temporal dependence; the use of Bayesian methods for extremes will be discussed by illustration. In Section 3 we discuss the aims and design of our simulation study for investigating the performance of the predictive return level, followed by a detailed discussion of our findings from this study. We conclude with some general comments and areas for future work in Section 4.

2 Practical motivation and modelling

In this Section, we introduce the wind speed data we will refer to throughout the paper. We then give a brief overview of the basic methods for modelling extremes on such a process, including some general background on Bayesian sampling and specific details relating to the posterior predictive return level. Some results are then presented comparing estimative return levels (i.e. means and credible intervals from the posterior distributions for z_r) to their predictive counterparts for our wind speed series.

2.1 Data

Fig. 1 shows boxplots, and a plot of each observation against its lag 1 counterpart, for a series of hourly gust wind speed maxima observed at Yeovilton in southwest England between January 1st 2003–December 31st 2012 (inclusive). The boxplots reveal clear seasonal variability in the wind speed extremes when using the calendar month as a seasonal unit, and there is also significant first-order autocorrelation (which persists above monthly-varying thresholds which we might use to classify observations as extreme). Estimates of z_r or $z_{r,\text{pred}}$ based on fitting an appropriate model to the wind speed extremes might be used to inform the design of a new structure. For example, the British Standards Institute (BSI) use estimates of the 50-year wind speed, from fits to wind speed extremes such as those shown in Fig. 1, to produce contour maps displaying the strength requirements for new structures. Similarly, the Office for Nuclear Regulation (ONR) recommends that structures at nuclear sites in the UK are built to protect against the 10,000-year return level associated with variables to which these structures might be vulnerable. We argue that a Bayesian approach to return level inference can improve the estimation procedure, with the potential to reduce estimation uncertainty (through the incorporation of prior knowledge) and, in practical terms, the ability to provide practitioners with a single point summary having all uncertainty due to natural variability and parameter estimation ‘built in’, through the predictive return level.

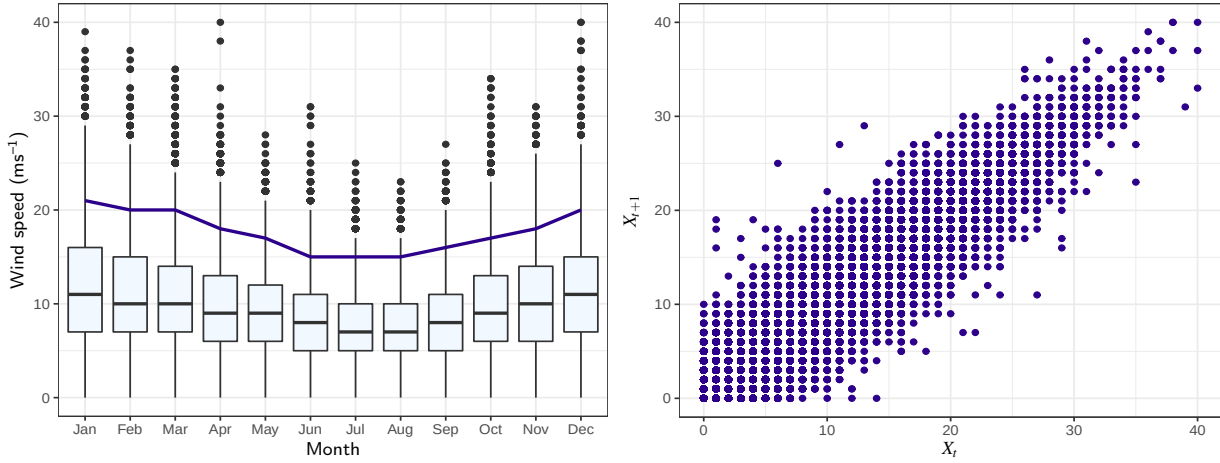


Fig. 1 Boxplots (by month, left) and each observation plotted against its lag 1 counterpart (right) for a series of hourly gust wind speed maxima observed at Yeovilton between January 1st 2003–December 31st 2012 (inclusive). The blue lines in the first plot correspond to high thresholds which have been chosen to identify observations as extreme. Units are ms^{-1} .

2.2 Statistical modelling

2.2.1 The basics

Let $\{X_n\}$ denote a stationary sequence of random variables with common distribution function F , and let $M_n = \max\{X_n\}$. It is typically the case that, as $n \rightarrow \infty$,

$$\Pr(M_n \leq x) \approx F^{n\theta}(x), \quad (1)$$

where $\theta \in [0, 1]$ is known as the *extremal index*; e.g. Leadbetter and Rootzén (1988). As $\theta \rightarrow 0$ there is increasing dependence in the extremes of the process; for an independent process, $\theta = 1$. In practice, F is unknown, and very small discrepancies in estimates of F obtained from observed data can lead to rather substantial discrepancies for F^n (and thus $F^{n\theta}$). Initially concerned with the independent case (i.e. $\theta = 1$), classical extreme value theory sought families of limiting models for F^n for large n . This leads to the GEV distribution (e.g. Jenkinson, 1955), with distribution function

$$\mathcal{G}(y) = \begin{cases} \exp\left[-(1 + \xi(y - \mu)/\varsigma)^{-1/\xi}\right], & \xi \neq 0 \\ \exp[-\exp(-(y - \mu)/\varsigma)] & \xi = 0, \end{cases} \quad (2)$$

defined on $\{y : 1 + \xi(y - \mu)/\varsigma > 0\}$, where $-\infty < \mu < \infty$, $\varsigma > 0$ and $-\infty < \xi < \infty$ are location, scale and shape parameters (respectively). The GEV can be used to model a set of block maxima $\{M_\tau\}$ with block length τ ; the calendar year is often used for τ , giving rise to an analysis of *annual maxima*.

Pickands (1975) showed that for large enough u the distribution of $(X - u)$, conditional on $X > u$, is approximately GP with distribution function

$$\mathcal{H}(y) = \begin{cases} 1 - (1 + \xi y/\sigma)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp[-y/\sigma] & \xi = 0, \end{cases} \quad (3)$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\sigma) > 0\}$, where $\sigma = \varsigma + \xi(u - \mu)$ and ξ are the GP scale and shape parameters (respectively). The GP distribution, being the limiting distribution for excesses over a high threshold u , provides a natural way of modelling extremes of time series such as our wind speed data discussed in Section 2.1. Modelling extremes in this way can be less wasteful than a block maxima approach using the GEV, since more extremes are usually pressed into use. Thus, in this paper we will focus on the use of the GP distribution as a model for excesses over a high threshold rather than the (often) more wasteful approach of using the GEV for block maxima.

2.2.2 Practicalities

Using the GP distribution to model threshold excesses, the linearity of $E[X - u | X > u]$ in u can be exploited in a *mean residual life plot* (MRL plot; see Coles, 2001, Ch. 4) to help find a suitably high threshold u for the classification of extremes. To maximise estimation precision, Fawcett and Walshaw (2016) suggest making use of *all* excesses over u , despite the obvious *temporal dependence* often present (as opposed to a standard POT analysis which fits to a filtered set of independent excesses). Specifically, they propose fitting (3) by adopting one of the following strategies:

(i) Parametric modelling of dependence

As in Smith et al. (1997) and Fawcett and Walshaw (2006), where appropriate assume a first-order Markov structure for the temporal evolution of extremes over u ; that is, assume the following likelihood for ψ :

$$L(\psi) = \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \psi) / \prod_{i=2}^{n-1} f(x_i; \psi), \quad (4)$$

where ψ is a parameter vector containing marginal and dependence parameter(s) and f , as appropriate, denotes a joint or marginal density function. Appealing to bivariate extreme value theory, transformation from GP to standard Fréchet margins gives a range of models to use for the dependence of consecutive extremes, the most commonly-used being the logistic family with dependence parameter α and distribution function:

$$G(x_i, x_{i+1}) = \exp \left\{ - \left(x_i^{-1/\alpha} + x_{i+1}^{-1/\alpha} \right)^\alpha \right\}. \quad (5)$$

Here, independence and complete dependence are attained when $\alpha = 1$ and $\alpha \rightarrow 0$ respectively. Differentiation of (5), with careful censoring when either one or both of (x_i, x_{i+1}) lies sub-threshold, gives pairwise contributions to the numerator in (4); univariate contributions to the denominator are given through (3). The polynomial relationship:

$$\theta \approx 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3, \quad (6)$$

as constructed in Fawcett and Walshaw (2012) and discussed in the Appendix, can then be used – after estimation of $\psi = (\sigma, \xi, \alpha)^\top$ (through maximum likelihood or Bayesian inference, for example) – to provide the fitted distribution for the right-hand-side of (1), using (3) as a model for F^n . Within this class of models for asymptotic dependence, other models can be used – for example, the bilogistic model, which allows for asymmetry in the dependence structure between (x_i, x_{i+1}) through the inclusion of an additional dependence parameter β ($0 < \alpha, \beta < 1$); see Coles (2001, Ch. 8) for more details. Indeed, Fawcett and Walshaw (2012) suggest polynomial expressions for the extremal index based on the fitted values of the dependence parameters here, too.

Of course it might be that, for the dependence structure, asymptotic independence is more appropriate; that is, the measure

$$\chi = \lim_{z \rightarrow z^*} \Pr(X_{i+1} > z | X_i > z),$$

where z^* is the upper limit of the support of the marginal distribution, takes the value zero (in the case of asymptotic dependence, $\chi > 0$; informally, χ is the probability that one variable is extreme given that the other is extreme). Here, a standard time series model such as a Gaussian AR(1) process can be used in place of the models we have outlined for consecutive variables which are asymptotically dependent (a marginal transformation being used to convert to GP form for observations exceeding the threshold). Here, $\theta = 1$, although Ancona-Navarrete and Tawn (2000) derive penultimate approximations for $\theta(u_p)$, a threshold-dependent extremal index with threshold u_p set at the p -% quantile; for example, for an AR(1) process, $\theta(u_{0.95}) \approx 0.711$ and $\theta(u_{0.99}) \approx$

0.855. As with the approximation in (6), in the Appendix we also construct approximations for $\theta(u_p)$ for an AR(1) process with dependence parameter (and hence lag 1 autocorrelation) A , and threshold u_p . The crucial censoring device, explained above in the context of the models used for asymptotic dependence, is also used in the application of an AR(1) process for threshold excesses.

We note here that, within our description of models for asymptotic dependence, formal tests and / or comparisons are available for selecting the most suitable model for first-order dependence; see, for example, Coles (2001, Ch. 8). We also note that in both the asymptotic dependent / independent cases it is straightforward to investigate the merits of a higher-order dependence (e.g. the trivariate extension to the logistic model, or an AR(2) process, to model dependence between consecutive triples), although there are obvious restrictions in the case of asymptotic dependence given the fixed number of dependence parameters in the logistic / bilogistic models (1 / 2 respectively). Referring specifically to the assumption of asymptotic dependence / independence, graphical diagnostics are available to aid this choice (e.g. $\chi / \bar{\chi}$ plots; see Coles, 2001, Ch. 8).

(ii) Direct estimation of the extremal index

Here, initially ignore dependence and proceed by fitting the GP distribution to all excesses over u to approximate F^n in (1). Then estimate the extremal index directly, without considering a parametric model for the dependence structure (as in approach (i)), to adjust for extremal dependence and hence complete the right-hand-side in (1). Fawcett and Walshaw (2016) make various recommendations for the extremal index estimator that should be used under this approach, but a simulation study shows that the estimator of Ferro and Segers (2003), given by

$$\bar{\theta} = \min \left(1, \frac{2 \left\{ \sum_{i=1}^{K-1} (T_i - a) \right\}^2}{(K-1) \sum_{i=1}^{K-1} (T_i - b)(T_i - c)} \right), \quad (7)$$

where T_i are the $K-1$ inter-arrival times between our K threshold excesses ($a = b = c = 0$ if the largest inter-arrival time is no greater than 2, and $a = b = 1$ and $c = 2$ if the largest inter-arrival time is greater than 2), strikes a good balance between optimising accuracy and precision and providing the practitioner with an estimator which is easy to use.

In the presence of seasonal variability (as shown in Fig. 1 for our wind speed extremes), Fawcett and Walshaw (2016) recommend a seasonal piecewise approach to modelling, with a unique GP model for extremes within each season. Of course, this should only be attempted where there is con-

confidence that it is the same physical mechanism generating extremes at different times of the year, seasonal variability in the extremes arising as a result of just a change in the scale of this mechanism. This assumption seems reasonable for wind speeds in temperate climates (e.g. the UK), where it is usually the same alternating sequence of anticyclones and depressions leading to most of the storms which occur throughout the year; the seasonal variability comes from the severity of these systems. For example, assuming either (i) or (ii) above to capture temporal dependence, estimates of $(\sigma_m, \xi_m, \theta_m), m = 1, \dots, M$, might be obtained for each season m (where M is the number of seasonal units over which we partition our data), the analysis perhaps being simplified by assuming a common shape parameter ξ or extremal index θ across all seasons (where appropriate). Anecdotal evidence in, for example, [Walshaw \(1994\)](#) and [Fawcett and Walshaw \(2006\)](#) indicates there are no real gains to be made, in terms of return level inference, by allowing the GP parameters to vary smoothly through time (e.g. by assuming Fourier forms for the parameters).

Though not a feature of our data in Fig. 1, trends in extremes can be modelled by imposing a time dependence on the GP scale parameter, i.e. $\sigma = \exp\{\beta_0 + \beta_1 t\}$, $t = 1, 2, \dots$, where t is an indicator of time. More generally, a dependence on covariates can be induced by writing the parameters in the form $h(\mathbf{X}^\top \boldsymbol{\beta})$, where h is a specified function, $\boldsymbol{\beta}$ is a vector of parameters and $\mathbf{X}$ is a model vector.

2.2.3 Return levels

Inversion of the right-hand-side of (1), assuming a GP model for threshold excesses, gives the following expression for the r -year return level:

$$z_r = \begin{cases} u + \sigma \xi^{-1} \left[(\lambda_u^{-1} w_r)^{-\xi} - 1 \right] & \xi \neq 0 \\ u - \sigma \log(\lambda_u^{-1} w_r) & \xi = 0, \end{cases} \quad (8)$$

where $w_r = 1 - [1 - (rn_y)^{-1}]^{1/\theta}$, λ_u is the rate of threshold excess and n_y is the (average) number of observations per year. Under approach (i) for dependence, as outlined in Section 2.2.2, (σ, ξ, θ) in Eq. (8) can be replaced with their maximum likelihood estimates / Bayesian estimates to obtain estimates of z_r , with an assessment of uncertainty being made through standard errors / posterior standard deviations and (profile-likelihood) confidence intervals / credible intervals, respectively, in the usual way. Under approach (ii), (σ, ξ) can be replaced with their maximum likelihood or Bayesian estimates and θ replaced with an estimate obtained via Eq. (7), with a bootstrap procedure as proposed in [Fawcett and Walshaw \(2012\)](#) enabling the incorporation of uncertainty in estimates of θ into estimates of z_r .

To recombine seasonally-varying parameters when estimating return levels, assuming independence between seasons we can solve the following for $x = z_r$:

$$\prod_{m=1}^M H_m(x)^{n_m \theta_m} = 1 - r^{-1}, \quad (9)$$

where H_m is the GP distribution function in season m with parameter set $(\lambda_{u_m}, \sigma_m, \xi_m)$, and θ_m / n_m are the extremal index / number of observations in season m . Of course, as discussed earlier, inference can be simplified if we assume a constant shape or dependence across all seasons.

2.3 Bayesian inference for wind speed extremes

After performing investigations into the dependence structure of our wind speed extremes, such as those described in [Fawcett and Walshaw \(2006a\)](#), we conclude that a first-order Markov structure, assuming asymptotic dependence according to the logistic model (Eq. 5), is appropriate (specifically, diagnostics such as the $\chi / \bar{\chi}$ plots as discussed in [Coles \(2001, Ch. 8\)](#) implied asymptotic dependence; comparisons between the bivariate logistic model and other models, and model-orders, did not improve over a fit of the former to consecutive pairs of extremes). Thus, we adopt approach (i) as outlined in Section 2.2.2 for handling **dependence of consecutive observations**. Given the seasonal variability observed in the wind speed extremes, and our earlier discussion in Section 2.2.2, we adopt a seasonal piecewise approach to modelling. Specifically, following discussions about the UK wind climate in [Walshaw \(1994\)](#), we use the calendar month as our seasonal unit (giving $M = 12$), assuming stationarity of wind speed extremes *within* each month. Here, MRL plots have been used for the selection of monthly-varying thresholds.

Following the recommendations of [Fawcett and Walshaw \(2016\)](#) we adopt a fully Bayesian approach to inference, using Markov chain Monte Carlo (MCMC) techniques to draw approximate samples from the marginal posteriors for $(\sigma_m, \xi_m, \theta_m)$, $m = 1, \dots, 12$, and hence z_r through Eq. (9). Details on MCMC techniques are now extensively published (e.g. [Gamerman and Lopes, 2006](#)), so we omit such detail here; suffice to say, all inferences are performed using a Metropolis-Hastings procedure and interested readers can re-produce the results shown in this Section by downloading our R code, and the wind speed data, at:

www.mas.ncl.ac.uk/~nlf8/predict

2.3.1 Prior specification

Generally, we work with a re-parameterised GP scale parameter:

$$\eta = \log(\sigma - \xi u).$$

This re-parameterisation gives a scale which is threshold-independent (unlike σ); working with the natural logarithm retains the positivity of this parameter in our MCMC sampling scheme. Based on work in Fawcett and Walshaw (2016), we adopt an informative prior specification for the parameter vector $(\eta_m, \xi_m, \alpha_m)^T$ based on observations on wind speed extremes made at a nearby location. Specifically, we use:

$$(\eta_m, \xi_m) \sim N_{24}(\mu, \Sigma) \quad \text{and} \quad \alpha_m \sim \text{Beta}(10, 19),$$

$m = 1, \dots, 12$. The components of the mean vector μ are chosen to closely match our beliefs about what are the most likely values of (η_m, ξ_m) based on our study of wind speeds at the nearby location; we specify values for $\text{cov}(\eta_m, \xi_m)$ according to our beliefs about the covariances between these parameters at the nearby location, scaled (albeit rather crudely) to reflect our uncertainties about differences between monthly wind speeds at the two locations. We choose our priors for α_m for similar reasons and, as is often the case, we assume independence between the marginal and dependence parameters. Of course, given the re-parameterisation to a threshold-independent scale parameter, more objective priors – perhaps independent Normal priors with large variances for the marginal components and $U(0, 1)$ priors for the dependence parameters – could be used in the absence of any such external information.

2.3.2 MCMC sampling

We set initial values for all parameters to their prior means, updating using a simple Metropolis step to give successive draws

$$(\eta_m^{[j]}, \xi_m^{[j]}, \alpha_m^{[j]}), \quad j = 1, \dots, 10^5.$$

Within each Metropolis step we use a random walk update to generate candidate values for each of the parameters, tuning the innovation variances to optimise the efficiency of our sampler. Convergence is assessed by starting each chain at multiple new initial values and observing the trace plots.

Eq. (6) is used to obtain a sample from the posterior for the extremal index θ_m , after which a posterior sample for z_r is obtained on substitution of successive draws for the GP parameters and the extremal index into Eq. (9).

2.3.3 Prediction

As discussed earlier, one of the advantages of a Bayesian analysis of extremes is the natural extension to prediction via the posterior predictive distribution. If Y denotes a future extreme of our wind speed series, then we can write

$$\Pr\{Y \leq y | \mathbf{x}\} = \int_{\Psi} \Pr\{Y \leq y | \psi\} \pi(\psi | \mathbf{x}) d\psi \quad (10)$$

for the *predictive distribution* of our extremes, where $\mathbf{x}$ represents past observations, ψ is a generic parameter vector and $\pi(\psi | \mathbf{x})$ is the posterior density for ψ . Solving

$$\Pr\{Y \leq z_{r,\text{pred}} | \mathbf{x}\} = 1 - r^{-1} \quad (11)$$

for $z_{r,\text{pred}}$ therefore gives an estimate of the r -year return level that incorporates natural variability *and* uncertainty in parameter estimation. Although (10) is analytically intractable, it can be approximated since we have estimated the posterior distribution using MCMC.

Regarding the sample $\psi^{(1)}, \dots, \psi^{(S)}$ as realisations from the stationary distribution $\pi(\psi | \mathbf{x})$ (i.e. after a suitable burn-in discard), we have

$$\Pr\{Y \leq z_{r,\text{pred}} | \mathbf{x}\} \approx \frac{1}{S} \sum_{j=1}^S \Pr\{Y \leq z_{r,\text{pred}} | \psi^{[j]}\}, \quad (12)$$

which we can set equal to $1 - r^{-1}$ and solve for $z_{r,\text{pred}}$ using a numerical solver. In our analysis of wind speed extremes, we have $\psi = (\eta_m, \xi_m, \theta_m)^T$, $m = 1, \dots, 12$.

2.3.4 Some results

Table 1 shows some estimative return levels for the wind speed extremes; that is, some point summaries from the posterior distributions for z_r , for $r = 10, 50, 200, 1000$ and 10000 years. Also shown are summaries of the spread of these posteriors via 95% credible intervals. Accompanying these estimative return level summaries are their predictive counterparts $\hat{z}_{r,\text{pred}}$. Fig. 2 shows plots of both $\hat{z}_r$ and $\hat{z}_{r,\text{pred}}$ over a range of values for r (on the usual logarithmic scale for these plots to magnify results for long-range return periods; posterior means are shown for $\hat{z}_r$, along with the 95% credible intervals). It is clear from both Table 1 and Fig. 2 that designing a structure to withstand the extremes of wind-speed as suggested by the estimative return levels could result in under-protection (especially when using the posterior mode), relative to the predictive estimates. This is more apparent for long return periods – recall from Section 2.1 that the ONR in the UK currently recommends that nuclear structures are protected against the 10,000 year event. Indeed, although not the case here, studies often report the predictive return level lying beyond even the 95% credible upper bound for z_r ; as an example, see the second block of results in Table 1 for another wind speed location in the Peak District of Central England (after Fawcett and Walshaw, 2016).

As discussed in Fawcett and Walshaw (2016), the predictive return level estimate might be preferred since it provides the practitioner with a single point summary which encapsulates both natural variability and uncertainty in parameter estimation. However, open questions remain about

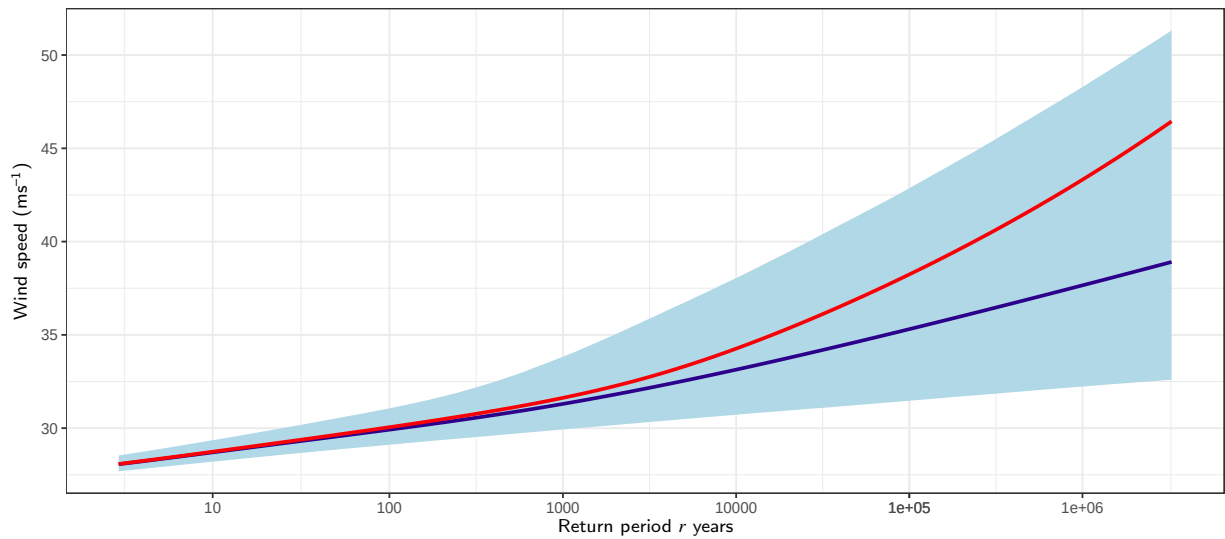


Fig. 2 Means (blue) taken from the MCMC samples of the posterior distributions for return levels z_r across a range of return periods r , with 95% credible intervals (outer light blue shaded area). Also shown, in red, are corresponding estimates of the predictive return levels $z_{r,\text{pred}}$. Units are ms^{-1} .

			Return period (r years)				
			10	50	200	1000	10000
Yeovilton (see Fig. 2)	$\hat{z}_r$	Post. mean	29.07	29.56	30.30	31.30	33.14
		Post. mode	28.02	28.06	28.11	28.15	28.20
	$\hat{z}_{r,\text{pred}}$	95% CI	(28.49, 29.85)	(28.85, 30.53)	(29.36, 31.68)	(29.92, 33.83)	(30.72, 38.05)
			29.13	29.65	30.47	31.62	35.26
Bradfield (after Fawcett & Walshaw, 2016)	$\hat{z}_r$	Post. mean	49.84	53.22	57.70	65.92	—
		Post. mode	49.29	51.40	55.67	63.44	—
	$\hat{z}_{r,\text{pred}}$	95% CI	(48.85, 50.85)	(48.41, 59.70)	(48.99, 64.59)	(60.51, 72.18)	—
			53.70	58.18	61.71	75.80	—

Table 1 Posterior means, modes and 95% credible intervals for return levels z_r , with estimates of the corresponding predictive return levels $z_{r,\text{pred}}$. Units are ms^{-1} .

the quantity $z_{r,\text{pred}}$. For example, how do exceedance probabilities of $z_{r,\text{pred}}$ compare to the intended values r^{-1} (on an annual scale)? How do these probabilities compare to those under an estimative approach for z_r ? Given results in, for example, Coles and Tawn (1996) and Fawcett and Walshaw (2016), we might expect $z_{r,\text{pred}}$ to give exceedance probabilities considerably lower than r^{-1} ; **implicit in the predictive return level is the allowance for uncertainty in parameter estimation, resulting in higher estimates of z_r and correspondingly lower estimates of r^{-1} .** But is this really the case, and if so, can these discrepancies be quantified and could they result in substantial *over-protection*? At the very least, practitioners should be aware of these discrepancies, should they choose to work with $z_{r,\text{pred}}$ (and we recommend that they do). Are there any advantages of using $z_{r,\text{pred}}$ as opposed to using some other point summary from the posterior for z_r , such as the upper end-point of the 95% credible interval or perhaps some other quantile? We aim to answer these questions, and more, in the simulation study in the next Section.

3 Simulation study

3.1 Study design

The first stage of our simulation study requires the simulation of a stationary reference series $\mathbf{y}^{(\text{Ref})}$ of length N years with N_y observations per year, where N is very large. We assume a first-order Markov structure, and we consider a range of models for the dependence between neighbouring extremes in $\mathbf{y}^{(\text{Ref})}$. Specifically, under the assumption of asymptotic dependence we use the logistic and bilogistic models as discussed in Section 2.2.2, covering a range of **temporal dependencies** through specific choices for $\alpha / (\alpha, \beta)$. To account for scenarios in which asymptotic independence might be a more plausible assumption, we also obtain $\mathbf{y}^{(\text{Ref})}$ from an $\text{AR}(1)$ process with lag 1 autocorrelation A , again covering a range of **temporal dependencies** through specific choices for A . Marginally, our reference series are **primarily** GP-distributed with scale and shape (σ, ξ) giving scale and shape $(\sigma^* = \sigma + \xi u, \xi^* = \xi)$ for excesses over some threshold u ; see Coles (2001, Ch. 4). **However,**

we also consider chains $\mathbf{y}^{(\text{Ref})}$ from distributions in one of the domains of attraction of the GP distribution (see Section 3.3, and in particular Section 3.3.6, for a discussion of our intentions in the simulation study relating to our marginal assumptions).

Now, at each replication ℓ in our simulation study, $\ell = 1, \dots, L$, we simulate a stationary series $\mathbf{y}^{(\ell)}$ of length n years, with n_y observations per year, n perhaps being typical of what we might usually work with in terms of environmental extremes. As with the reference dataset, we assume a first-order Markov structure according to models for asymptotic dependence and asymptotic independence, with the same marginal assumptions as before. For each series $\mathbf{y}^{(\ell)}$ we perform a full MCMC procedure, primarily using objective (and, where available, conjugate) priors (see Section 3.3.4 for a discussion of our intentions in the simulation study relating specifically to the choice of prior distributions). For example, for the AR(1) process with lag 1 autocorrelation A , we have that $\mathbf{y}^{(\ell)} | \mu, \tau \sim N(\mu = 0, \tau^{-1} = (1 - A^2)^{-1})$. We thus assume the conjugate prior specification

$$\mu | \tau \sim N\left(0, \frac{1}{c\tau}\right) \quad \text{and} \quad \tau \sim Ga(g, h),$$

yielding

$$\mu | \tau, \mathbf{y}^{(\ell)} \sim N\left(\frac{n\bar{\mathbf{y}}^\ell}{c + nn_y}, \frac{1}{(c + nn_y)\tau}\right)$$

and

$$\tau | \mathbf{y}^{(\ell)} \sim Ga\left(g + \frac{nn_y}{2}, h + \frac{cnn_y(\bar{\mathbf{y}}^\ell)^2}{2(c + nn_y)} + \frac{nn_y s^2}{2}\right),$$

where $\bar{\mathbf{y}}^\ell$ and s^2 are the mean and variance (respectively) of the simulated series $\mathbf{y}^{(\ell)}$, $c = 10^{-1}$ and $g = h = 10^{-3}$. This enables posterior inferences to be made on the autoregressive parameter A and hence the extremal index θ via the polynomial approximation constructed in the Appendix. An example in the case of asymptotic dependence is where we use the logistic model with dependence parameter α ; assuming the prior $\alpha \sim U(0, 1)$, we then perform Metropolis-Hastings sampling, as outlined in Section 2.3.2, to obtain draws from the posterior for α using the likelihood in Eq. (4), and hence for the extremal index θ via Eq. (6). In the case of asymptotic dependence / independence we then transform the margins from standard Fréchet / Normal, respectively, to GP with scale and shape (σ, ξ) , before performing a full MCMC procedure on excesses over a range of thresholds u .

The procedures outlined above yield S iterations after burn-in to obtain approximate samples $\sigma^{*(\ell)}$ and $\xi^{*(\ell)}$, $\ell = 1, \dots, L$, of length S from the posterior distributions of the GP scale

and shape σ^* and ξ^* , as well as approximate samples from the posteriors of the dependence parameters (i.e. α or A) and hence samples $\theta^{(\ell)}$ from the posterior of the extremal index; see Section 2.2.2. At each replication ℓ , via Eq. (8) we also obtain posterior samples $z_r^{(\ell)}$ from the r -year return levels for a range of return periods r . From these draws we can obtain the posterior mean $\bar{z}_r^{(\ell)}$, the posterior mode $\hat{z}_r^{(\ell)}$ and the posterior 95% credible interval upper bound $z_{r,\text{upper}}^{(\ell)}$; we also obtain the predictive return level $z_{r,\text{pred}}^{(\ell)}$ via Eq. (12), essentially giving sampling distributions of length L for each of these return level summaries. Defining p to be the proportion of annual maxima in $\mathbf{y}^{(\text{Ref})}$ exceeding each of $\bar{z}_r^{(\ell)}$, $\hat{z}_r^{(\ell)}$, $z_{r,\text{upper}}^{(\ell)}$ and $z_{r,\text{pred}}^{(\ell)}$, $\ell = 1, \dots, L$, gives sampling distributions for $p_{\bar{z}_r}$, $p_{\hat{z}_r}$, $p_{z_{r,\text{upper}}}$ and $p_{z_{r,\text{pred}}}$, respectively. Other than the sampling distributions for each of the return level summaries themselves, of particular interest might be comparisons between each of the proportions p_* and the intended exceedance probabilities r^{-1} (the $*$ subscript used here to denote generically any one of our estimators for r^{-1}).

3.2 Parameters in our study

In our study, we use $N = 10^5$ and $N_y = 365.25 \times 24$, in line with having hourly measurements on our variable. We use $\alpha = \{0.3, \dots, 0.9\}$ for the logistic model; for the bilogistic model we fix α at 0.5 and use $\beta = \{0.3, \dots, 0.9\}$; for the AR(1) process we use $A = \{0.2, \dots, 0.8\}$.

Marginally, we hold σ unit constant but consider a range of tail behaviours through the GP shape parameter ξ , where $\xi = \{-0.4, -0.1, 0, 0.1, 0.3\}$, yielding $\sigma^* = \xi u + 1$ and $\xi^* = \xi$, where we use $u = \{u_{0.9}, u_{0.95}, u_{0.99}\}$. We perform $L = 1000$ replications, within which we simulate chains $\mathbf{y}^{(\ell)}$ of length $n = 50$ years with $n_y = N_y$; for each chain, we perform MCMC with $S = 10,000$ iterations after an appropriate burn-in discard. For each combination of $(\sigma^*, \xi^*, \theta)$ across all dependence models considered, we perform small MCMC pilot runs in a bid to select suitable values for the MCMC tuning parameters before running the full simulation study, aiming for acceptance rates of between 20-30%.

3.3 Results

In this Section we present the findings of our simulation study, focusing on comparisons between the predictive return level and the other summaries obtained directly from the posterior distribution for z_r . Specifically, we give attention to the return level exceedance probabilities associated with the different Bayesian estimators for z_r , and we look at how these compare to the intended values r^{-1} . We consider

the cases of asymptotic dependence (Section 3.3.1) and independence (Section 3.3.2), but also the effects of model mis-specification on return level inference when asymptotic dependence / independence is incorrectly assumed (Section 3.3.3). We investigate the effects of prior specification on estimates of return level exceedance probabilities (Section 3.3.4), relative to those discussed in Sections 3.3.1–3.3.2; specifically, we look at the effects of using informative priors on marginal and dependence parameters, and the effects of mis-chosen informative priors (unless otherwise stated, all results in other sections lean on objective prior specifications). For information, we investigate the performance of the predictive return level under our approach, which uses information on *all* threshold excesses, to that obtained under the commonly-used POT approach (Section 3.3.5). We also assess the effects of using chains which are drawn marginally from distributions within the domain of attraction of the GP distribution, rather than directly from the GP distribution itself (Section 3.3.6). At the end of this Section we give some general comments on the sensitivity of our comparisons between the different estimators of r^{-1} to the marginal structure of the simulated chains (Section 3.3.7).

3.3.1 Asymptotic dependence

One arm of the study: Logistic dependence structure with $\xi = -0.4$ and $u = u_{0.95}$

Fig. 3 shows sampling distribution means, and 95% confidence intervals (taken as the interval between the sampling distribution 2.5% and 97.5% quantiles), for $\log(1 + rp_*)$. We present results for the ratio of p_* to r^{-1} (and on a logarithmic scale) to give the plots greater clarity – especially for large values of r , for which results summarising p_* directly would be very difficult to read. The horizontal dotted lines are at $\log 2 = \log [\mathbb{E}(rp_*) + 1] \geq \mathbb{E}[\log(1 + rp_*)]$, according to Jensen's inequality, in effect giving a theoretical upper bound to the means of our sampling distributions for $\log(1 + rp_*)$. The target of each of our estimators is a probability close to zero; also, over-estimation of these probabilities would arise from under-estimation of the corresponding return levels, perhaps resulting in under-protection from a safety point-of-view if such estimates were to be used as design parameters (e.g. as strength requirements for new buildings or structures). Thus, over-estimation of r^{-1} by p_* might be seen as more costly than under-estimation, but the root mean squared error (RMSE), given by

$$\sqrt{L^{-1} \sum_{\ell=1}^L (p_*^{(\ell)} - r^{-1})^2},$$

punishes under- and over-estimation equally. Thus, linear-exponential errors – or *linex* errors (e.g. Zellner, 1986) –

given by

$$\exp\{d(p_* - r^{-1})\} - d(p_* - r^{-1}) - 1, \quad d > 0,$$

can be used to impose an asymmetric error favouring under-estimation of r^{-1} . Table 2 therefore reports the mean *linex* error (MLE) in each component of our simulation study, along with the standard RMSE for comparison. Both of these error measures in Table 2 accompany the estimated bias for each estimator p_* . For these particular results the simulated chains display asymptotic dependence according to the bivariate logistic model for consecutive pairs in the process; marginally, $\xi = -0.4$ and $u = u_{0.95}$.

The superiority of the predictive return level over the most commonly-used posterior summary – the posterior mean – is obvious, especially for longer return periods. For example, Table 2 and Fig. 3 show that the predictive return level yields exceedance probabilities $p_{z_r, \text{pred}}$ which are increasingly more accurate and precise (i.e. narrower sampling distribution confidence intervals and smaller RMSE / MLE) as the return period r increases, especially as the extremal dependence in the series weakens (i.e. as $\alpha \rightarrow 1$). In comparison, the return level posterior mean is (at best) on a par with the predictive return level, in terms of its associated exceedance probabilities, when $r = 10$; for longer return periods the bias of these estimated exceedance probabilities is noticeably larger than those produced by the predictive return level (and increasingly so as r increases), as is our uncertainty in these estimates (e.g. wider sampling distribution confidence intervals and larger RMSE / MLE values). Where both the predictive return level and the return level posterior mean lead to exceedance probabilities which over-estimate r^{-1} (e.g. positive biases in the results reported in Table 2), there is usually a smaller bias in $p_{z_r, \text{pred}}$ than in $p_{\bar{z}_r}$. For most strengths of dependence, and for larger return periods, the sampling distribution means for $\log(1 + rp_*)$ are within their range (i.e. $\leq \log 2$) when $p_* = p_{z_r, \text{pred}}$, and certainly on more occasions than when $p_* = p_{\bar{z}_r}$. As we might expect, r^{-1} is often under-estimated when using the return level posterior 95% credible upper bound; especially for shorter-range return periods, we might expect $z_{r, \text{upper}}$ to over-estimate z_r , leading to too-small values for $p_{z_{r, \text{upper}}}$. The return level posterior mode consistently produces estimates $p_{\hat{z}_r}$ which are too large. Indeed, these are always substantially larger than those produced by the other three estimators, and with the largest uncertainty, casting doubt on the value of the posterior mode as a useful summary of the return level posterior distribution.

As with Fig. 3, the top row of Fig. 5 shows sampling distribution means for our four estimates of r^{-1} , but now across a smooth range of values for r for some fixed values of the logistic dependence parameter. Here, we choose $\alpha = 0.3$, $\alpha = 0.5$ and $\alpha = 0.9$, representing fairly strong extremal depen-

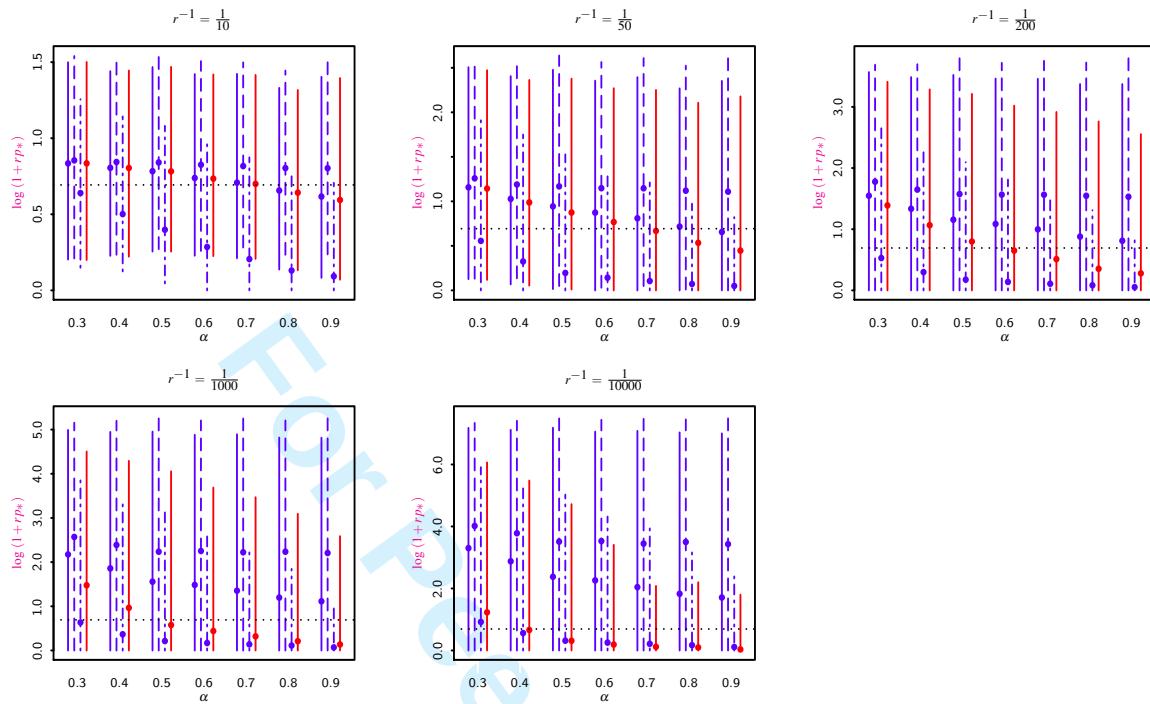


Fig. 3 Sampling distribution means (bullets) and 95% confidence intervals (vertical lines, running between the sampling distribution 2.5% and 97.5% quantiles) for $\log(1+rp_*)$, using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (ii) the posterior predictive return level (red). Here, the simulated data are constructed with asymptotic dependence according to a bivariate logistic model with dependence α . The horizontal dotted line is at $\log 2$, representing the maximum of $\mathbb{E}[\log(1+rp_*)]$. Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

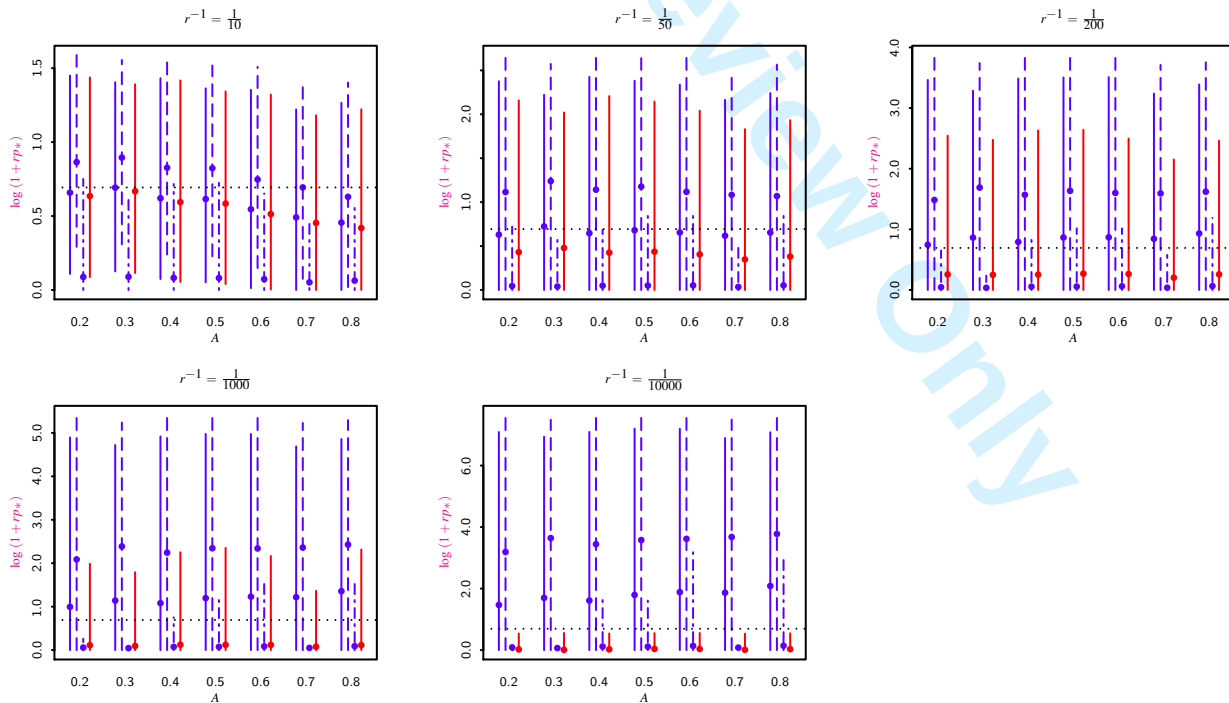


Fig. 4 Sampling distribution means (bullets) and 95% confidence intervals (vertical lines, running between the sampling distribution 2.5% and 97.5% quantiles) for $\log(1+rp_*)$, using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (ii) the posterior predictive return level (red). Here, the simulated data are constructed with asymptotic independence according to an $AR(1)$ process with lag 1 autocorrelation A . The horizontal dotted line is at $\log 2$, representing the maximum of $\mathbb{E}[\log(1+rp_*)]$. Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

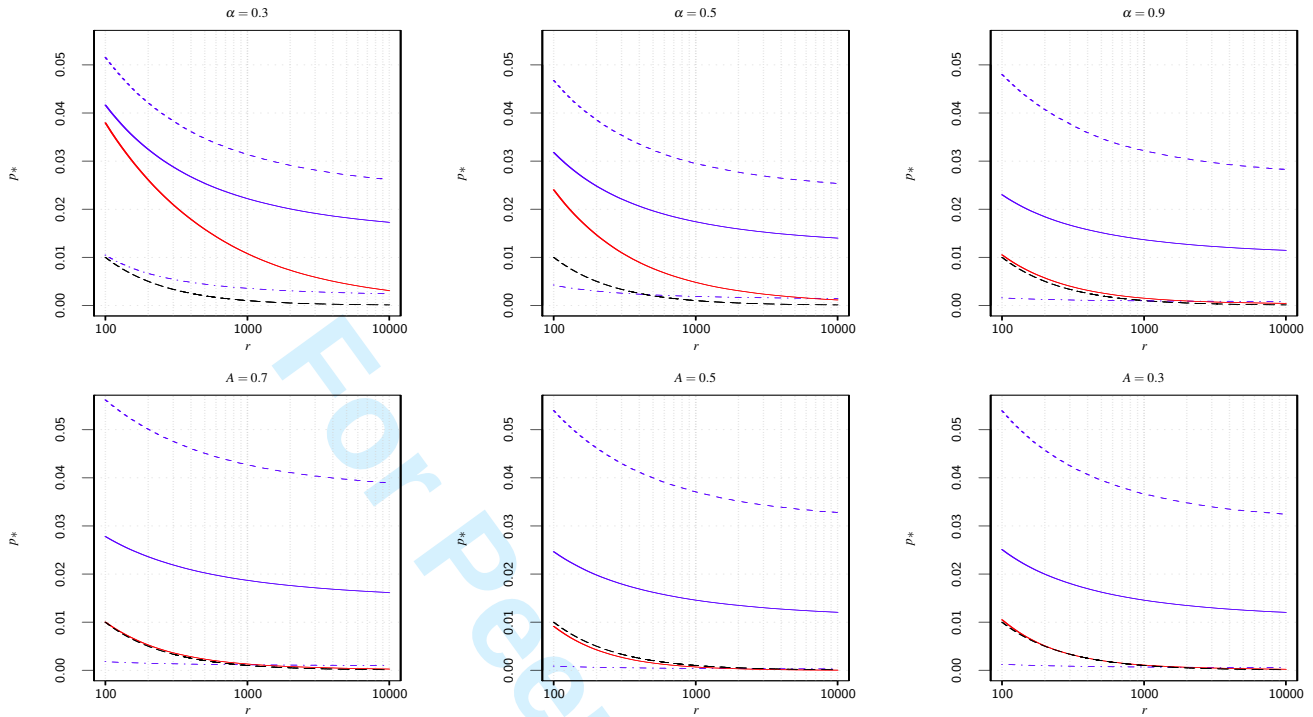


Fig. 5 Sampling distribution means for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior 95% credible upper bound), and (ii) the posterior predictive return level (red). Simulated data constructed with: asymptotic dependence according to a bivariate logistic model with dependence α (top row); asymptotic independence according to an AR(1) process with lag 1 autocorrelation A (bottom row). The black dashed line represents the target exceedance probability r^{-1} . Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

dence (similar to that observed in our wind speed extremes), moderate extremal dependence and near-independent extremes, respectively. We see that estimates based on the return level posterior mean and the predictive return level ($p_{\bar{z}_r}$ and $p_{z_{r,\text{pred}}}$ respectively) consistently over-estimate r^{-1} when $\alpha = 0.3$, but with estimates based on the predictive return level always being substantially less biased than those based on the posterior mean, especially for longer return periods. Interestingly, for this level of dependence estimates of r^{-1} based on the return level 95% credible upper bound are closest to the intended exceedance probability, and this observation is supported by the results in Table 2. As the extremal dependence weakens, we see an even closer agreement between our estimated exceedance probabilities based on the predictive return level and the intended exceedance probabilities r^{-1} , with estimates using the return level posterior mean consistently displaying a larger bias in our plots in Fig. 5. In agreement with the results shown in Fig. 3, the plots in Fig. 5 show that across all values for $r \geq 100$, estimated exceedance probabilities based on the return level posterior mode are always most biased, with substantial over-estimation of the intended exceedance probability.

To put these results into a practical context, recall from Section 2.1 that we discuss the use of the 10,000-year return

level estimate by the ONR in the UK, as a design requirement for structures at nuclear sites. For our wind speed data, monthly estimates (posterior means) of the logistic dependence parameter α are around 0.3. Focusing on the final plot in Fig. 3, and the first plot in Fig. 5, we see the much smaller bias in estimates of r^{-1} produced by the predictive return level than the return level posterior mean, and with greater precision in the predictive estimates; however here, for this long-range return period, the 95% credible upper bound for z_r produces the best estimate of r^{-1} (as discussed above). An over-estimate of r^{-1} could result in significant under-protection (as this arises from an under-estimate of z_r), and we note here that – relative to the estimates of r^{-1} based on the predictive return level and return level 95% credible upper bound – those based on the return level posterior mean are much over-estimated (and with more uncertainty).

Other arms of the study: main findings

Here, we report some findings from other arms of the study in which the simulated chains displayed asymptotic dependence. Largely, the general direction of the results already reported was replicated in other arms. For instance, sticking with the logistic model but changing the marginal shape parameter still resulted in return level exceedance probabil-

		Asymptotic dependence: logistic model for bivariate extremes with dependence α											
		$\alpha = 0.3$				$\alpha = 0.5$				$\alpha = 0.9$			
		$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$	$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$	$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$
$r^{-1} = \frac{1}{10}$	Bias	4.311	4.905	-0.350	4.325	2.802	4.295	-4.729	2.776	-0.291	3.555	-8.779	-0.715
	RMSE	9.290	9.990	5.857	9.327	8.637	9.915	6.870	8.626	7.688	9.386	9.365	7.596
	MLE	0.459	0.533	0.175	0.463	0.397	0.528	0.233	0.396	0.308	0.471	0.426	0.300
$r^{-1} = \frac{1}{50}$	Bias	3.692	4.655	-0.098	3.546	2.425	4.051	-1.284	1.943	1.131	4.005	-1.773	-0.004
	RMSE	6.641	7.765	2.846	6.473	6.116	7.829	2.530	5.505	5.413	8.044	2.408	4.243
	MLE	0.233	0.320	0.042	0.217	0.199	0.330	0.033	0.161	0.156	0.349	0.030	0.096
$r^{-1} = \frac{1}{200}$	Bias	2.745	3.711	0.164	2.110	1.981	3.359	-0.202	0.965	1.342	3.572	-0.376	0.070
	RMSE	5.167	6.372	1.732	4.243	4.919	6.671	1.483	3.376	4.517	7.113	1.414	2.366
	MLE	0.140	0.215	0.015	0.094	0.129	0.239	0.012	0.060	0.110	0.273	0.011	0.030
$r^{-1} = \frac{1}{1000}$	Bias	2.119	3.039	0.255	0.976	1.640	2.850	0.087	0.379	1.266	3.117	-0.006	0.050
	RMSE	4.254	5.523	1.265	2.470	4.231	5.958	1.183	1.871	4.037	6.488	1.265	1.378
	MLE	0.095	0.161	0.008	0.032	0.095	0.191	0.007	0.018	0.088	0.227	0.009	0.010
$r^{-1} = \frac{1}{10000}$	Bias	1.717	2.609	0.232	0.298	1.389	2.526	0.134	0.103	1.135	2.815	0.073	0.032
	RMSE	3.715	5.020	1.049	1.140	3.834	5.568	1.049	0.837	3.755	6.148	1.183	0.894
	MLE	0.072	0.133	0.006	0.007	0.078	0.166	0.006	0.004	0.076	0.204	0.008	0.004

		Asymptotic independence: AR(1) process with lag 1 autocorrelation A											
		A = 0.7				A = 0.5				A = 0.3			
		$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$	$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$	$P_{\bar{z}_r}$	$P_{\bar{z}_r}$	$P_{z_{r,upper}}$	$P_{z_{r,pred}}$
$r^{-1} = \frac{1}{10}$	Bias	-1.920	2.044	-9.065	-2.604	-0.242	4.242	-9.038	-0.823	0.732	5.217	-8.817	0.255
	RMSE	7.962	9.160	9.508	7.918	7.836	9.716	9.397	7.727	8.643	10.877	9.365	8.503
	MLE	0.324	0.444	0.439	0.319	0.316	0.502	0.428	0.305	0.392	0.637	0.426	0.378
$r^{-1} = \frac{1}{50}$	Bias	1.476	4.600	-1.758	-0.122	1.308	4.677	-1.859	-0.100	1.423	4.747	-1.815	0.135
	RMSE	5.727	8.706	2.408	4.050	5.320	8.450	2.074	3.808	5.788	9.028	2.236	4.382
	MLE	0.175	0.408	0.030	0.087	0.149	0.380	0.021	0.075	0.178	0.439	0.025	0.101
$r^{-1} = \frac{1}{200}$	Bias	1.860	4.512	-0.354	0.026	1.480	4.122	-0.439	-0.075	1.597	4.077	-0.405	0.007
	RMSE	5.050	8.068	1.414	2.258	4.301	7.423	0.707	1.673	4.648	7.836	1.049	2.121
	MLE	0.137	0.350	0.011	0.027	0.097	0.293	0.003	0.014	0.115	0.330	0.005	0.024
$r^{-1} = \frac{1}{1000}$	Bias	1.769	4.173	0.013	0.030	1.357	3.609	-0.062	-0.032	1.357	3.564	-0.033	0.007
	RMSE	4.615	7.570	1.225	1.183	3.728	6.723	0.447	0.548	4.050	7.113	0.775	1.000
	MLE	0.114	0.308	0.008	0.008	0.073	0.240	0.001	0.002	0.087	0.272	0.003	0.005
$r^{-1} = \frac{1}{10000}$	Bias	1.606	3.885	0.090	0.015	1.195	3.271	0.018	-0.008	1.194	3.231	0.045	0.011
	RMSE	4.336	7.246	1.183	0.447	3.376	6.317	0.316	0.000	3.701	6.708	0.707	0.316
	MLE	0.101	0.282	0.007	0.001	0.060	0.212	0.001	0.000	0.073	0.242	0.003	0.001

Table 2 Estimated biases ($\times 100$), root mean squared errors (RMSE, $\times 100$) and mean linear errors (MLE, $\times 100$), for each of our four estimators of the return level exceedance probability r^{-1} . The top part of the table summarises results when our simulated series exhibit asymptotic dependence according to the logistic model with dependence parameter α ; the bottom part when our simulated series exhibit asymptotic independence according to an AR(1) process with lag 1 autocorrelation A. Marginally, we have $\xi = -0.4$ and $u = u_{0.95}$. Values in **bold** or **bold italics** are those which are the smallest in each component of the study.

ities more in line with the intended values r^{-1} when using the predictive return level compared to the return level posterior mean, with substantially smaller biases for return periods of practical interest and **much smaller values of RMSE / MLE**. As an example, with $\xi = 0.1$ and $\alpha = 0.3$ or 0.5 , biases incurred by $p_{\bar{z}_r}$ were always larger than those incurred by $p_{z_{r,pred}}$ (around five times larger for $r = 10,000$), with consistently narrower 95% confidence intervals; when $\alpha = 0.9$, the outperformance of $z_{r,pred}$ relative to $\bar{z}_r$ was even more marked. One difference to note is in estimates of r^{-1} using the return level posterior mode: for simulated chains with increasingly heavy tails (e.g. when positive values for ξ were used), we observed smaller biases than those reported in Fig. 3, Fig. 5 and Table 2 (for which $\xi = -0.4$), perhaps indicating that this summary would be more useful for very positively skewed data (as might be expected).

Switching to the bilogistic model for consecutive extremes, allowing for asymmetry in the dependence structure, did not result in noticeable deviations from the results discussed so far, for combinations of dependence parameters α and β resulting in similar levels of dependent extremal dependence as given by the dependence parameter α in the logistic model. This might suggest a robustness of our findings across different dependence structures within an overall framework of asymptotic dependence. Similarly, our results were consistent across the other two threshold levels considered ($u_{0.9} / u_{0.99}$, the 90% / 99% marginal quantiles respectively).

3.3.2 Asymptotic independence

One arm of the study: AR(1) with $\xi = -0.4$ and $u = u_{0.95}$

Fig. 4 shows the same information as Fig. 3 (i.e. sampling distribution means and 95% confidence intervals for $\log(1 + rp_*)$), but now for the case of asymptotic independence where our simulated chains are AR(1) processes with lag 1 autocorrelation A ; as in Section 3.3.1, marginally $\xi = -0.4$ and $u = u_{0.95}$. The bottom half of Table 2 reports the estimated bias, RMSE and MLE for our four estimators of r^{-1} . The superiority of the predictive return level (in terms of estimating the intended exceedance probability r^{-1}) relative to estimates obtained using the return level posterior mean, is obvious, and more apparent than in the previous section when considering series with asymptotic dependence. For example, the sampling distribution means for $\log(1 + rp_*)$ are all within their range (i.e. $\leq \log 2$) when $p_* = p_{z_{r,\text{pred}}}$, regardless of the return period r and strength of dependence A ; this is not the case when $p_* = p_{\bar{z}_r}$. As in the case of asymptotic dependence, we also note greater precision in estimates based on the predictive return level, with narrower 95% confidence intervals (substantially so for larger return periods). The results reported in the bottom half of Table 2 confirm this, with smaller biases typically being observed for $p_{z_{r,\text{pred}}}$ than for $p_{\bar{z}_r}$ and much smaller values of RMSE / MLE. As in Section 3.3.1, estimates based on the return level posterior mode perform most poorly, with estimates based on the return level 95% credible upper bound seemingly performing well (though not as well as those based on the predictive return level) for some large values of r . These results are supported by the plots in the bottom row of Fig. 5, in which we see estimates of the intended exceedance probability r^{-1} based on the predictive return level being consistently less biased than all the others, for the three levels of dependence we focus upon ($A = 0.7$, $A = 0.5$ and $A = 0.3$, representing reasonably strong, moderate and weak dependence, respectively).

Other arms of the study: main findings

We report similar findings for other arms of the study in which the simulated chains display asymptotic independence according to an AR(1) process, but with different values of ξ or different thresholds being used. In all cases, estimates of the intended exceedance probability r^{-1} had smallest bias, and RMSE / MLE, when based on the predictive return level, and especially so for long return periods (e.g. $r = 1000$ and $r = 10000$). Estimates based on the return level posterior mean and posterior mode were consistently too large.

3.3.3 Mis-specification of dependence

We now investigate the effects of mis-specifying the dependence structure on our four estimates of the return level exceedance probability r^{-1} . Specifically, at each iteration ℓ , $\ell = 1, \dots, 1000$, we simulate $\mathbf{y}^\ell$ from an AR(1) process with lag 1 autocorrelation A ; inference then proceeds by assuming asymptotic dependence and fitting the logistic / bilogistic model to consecutive pairs in the process, and then obtaining estimates of r^{-1} using $p_{\bar{z}_r}$, p_{z_r} , $p_{z_{r,\text{upper}}}$ and $p_{z_{r,\text{pred}}}$ in the way we describe in Section 3.1. Conversely, we also simulate $\mathbf{y}^\ell$ with asymptotic dependence via the logistic / bilogistic models, inference then proceeding assuming asymptotic independence through the fitting of an AR(1) process. In reality, the precise form of dependence structure is unknown; diagnostic checks such as the $\chi/\bar{\chi}$ -plots discussed in Coles (2001, Ch. 8) can be used to help assess the nature of the dependence present (i.e. asymptotically dependent / independent), although their interpretation can be difficult. Thus, the aim of this part of the simulation study is to investigate our four estimators of the return level exceedance probability under an incorrect specification of dependence structure, something which could easily occur in an analysis of real data when we attempt to press all threshold excesses into use (in Section 3.3.5, we investigate our four estimators of r^{-1} under a standard POT analysis, wherein an attempt is made to completely filter out dependence).

Fig. 6 (top row) shows plots similar to those in Fig. 5: we have sampling distribution means for $p_{\bar{z}_r}$, p_{z_r} , $p_{z_{r,\text{upper}}}$ and $p_{z_{r,\text{pred}}}$ across a range of values for r . The simulated data exhibit asymptotic independence through an AR(1) structure with lag 1 autocorrelation A , but asymptotic dependence is incorrectly assumed with a dependence structure for consecutive pairs according to the logistic model with parameter α . Compared to the results shown in the bottom row of plots in Fig. 5, in which the correct form of dependence was assumed, we see a larger bias in estimates of r^{-1} with $p_{z_{r,\text{pred}}}$ across all values of lag 1 autocorrelation A ; however, the predictive return level still clearly outperforms both the posterior mean and mode in terms of the associated exceedance probabilities it yields and their proximity to the intended values r^{-1} . Although not reported here, values of the RMSE / MLE were consistently smaller for $p_{z_{r,\text{pred}}}$ than the other three estimates associated with the return level posterior distribution. For the opposite case of mis-specification in terms of the dependence structure – that is, when the data were simulated to exhibit asymptotic dependence but an AR(1) process was assumed – we observed an increase in the bias of estimated exceedance probabilities associated with the return level posterior mean, mode and 95% confidence upper bound, but with the predictive return level yielding estimates of r^{-1} close to the intended values.

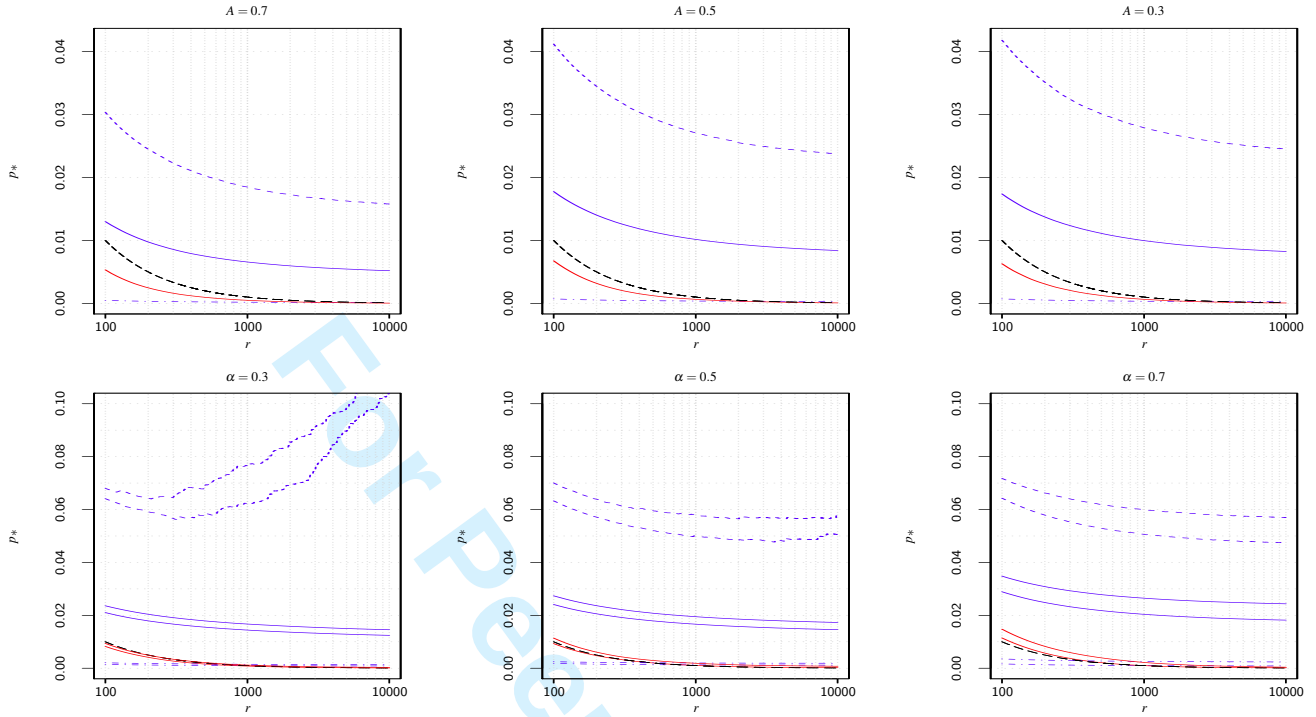


Fig. 6 Sampling distribution means for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound), and (ii) the posterior predictive return level (red). Simulated data constructed with: asymptotic independence according to an $AR(1)$ process with lag 1 autocorrelation A , but when fitting, asymptotic dependence assumed according to a bivariate logistic model (top row); asymptotic dependence according to a bivariate logistic model with dependence α , but dependence filtered using runs declustering with cluster termination interval $\kappa = 5$ and $\kappa = 20$ (bottom row; results using $\kappa = 20$ giving the higher curve each time). Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

3.3.4 Prior specification

All results discussed so far have assumed an objective (and, where possible, conjugate) prior specification for both dependence and marginal components of our simulated series. However, in keeping with our wind speed data analysis in Section 2.3, for some arms of the study we also adopt informative priors. For example, consider the case of asymptotic dependence under the logistic model with dependence parameter α . To emulate our approach to prior specification in Section 2.3.1, at each replication ℓ a pair of stationary series $(\mathbf{y}^\ell, \mathbf{y}^{\dagger\ell})$ is simulated, each series in this pair being drawn from the same GP distribution marginally and having the same dependence structure. Maximum likelihood estimates of the marginal and dependence parameters for $\mathbf{y}^{\dagger\ell}$, and the corresponding elements of their covariance matrix, are then used to inform the prior specification for the marginal and dependence parameters for $\mathbf{y}^\ell$. Specifically, we assume that $(\eta^* = \log(\sigma^* - \xi^*u), \xi^*) \sim N_2(\mu, \Sigma)$ and $\alpha \sim \text{Beta}(a, b)$, with sensible choices for μ , Σ , a and b based on our analysis of $\mathbf{y}^{\dagger\ell}$ (as opposed to an objective specification using independent $N(0, v)$ priors for η^* and ξ^* with large v , and a $U(0, 1)$ prior for α , as used in Section 3.3.1).

Using informative priors based on $\mathbf{y}^{\dagger\ell}$ usually resulted in no obvious change in the accuracy of the return level exceedance probabilities obtained using our four Bayesian estimates of return levels, relative to those obtained under an assumption of objective priors (although occasionally noticeable reductions in bias were observed under the informative prior specification). However, as expected, estimates were appreciably more precise, with much smaller values of $RMSE / MLE$ for estimates obtained using all four of our Bayesian posterior summaries (particularly so for those based on the predictive return level). As examples, when $\alpha = 0.5$ using the logistic model for series with asymptotic dependence, we see from Table 2 that: (i) the bias, $RMSE$ and $MLE (\times 100)$ for $r^{-1} = 1/1000$ are 0.379, 1.871 and 0.018, respectively, for estimates based on the predictive return level – assuming informative priors based on $\mathbf{y}^{\dagger}$ gives corresponding values of 0.371, 1.342 and 0.009; (ii) the bias, $RMSE$ and $MLE (\times 100)$ for $r^{-1} = 1/10000$ are 0.134, 1.049 and 0.006, respectively, for estimates based on the return level posterior mean – assuming informative priors based on $\mathbf{y}^{\dagger\ell}$ gives corresponding values of 0.037, 0.447 and 0.001.

To investigate the effects of a mis-chosen prior, for some arms of the study (‘strong’ dependence only – i.e. $\alpha = 0.3$

and $A = 0.9$ for the logistic model and $AR(1)$ processes respectively) we also allow μ , Σ , a and b to be informed by a maximum likelihood analysis of simulated chains $\mathbf{y}_i^{\pm\ell}$, $i = 1, \dots, 5$; unlike $\mathbf{y}^{\pm\ell}$, these chains have means / variances and temporal dependencies which are (increasingly) dissimilar to those in $\mathbf{y}^\ell$. Specifically, we set:

$$\bar{\mathbf{y}}_i^{\pm\ell} = (0.5i + 1) \times \bar{\mathbf{y}}^\ell \quad \text{and} \\ \text{s.d.}(\mathbf{y}_i^{\pm\ell}) = (-0.2i + 1.1) \times \text{s.d.}(\mathbf{y}^\ell)$$

and, depending on whether the simulated chains exhibit asymptotic dependence or asymptotic independence,

$$\alpha_{\mathbf{y}_i^{\pm\ell}} = 0.1i + 0.4 \quad \text{or} \\ A_{\mathbf{y}_i^{\pm\ell}} = -0.1i + 0.8$$

respectively, meaning that the series $\mathbf{y}_5^{\pm\ell}$ is the most dissimilar to $\mathbf{y}^\ell$ (hence leading to the most ill-informed prior specification). Informative priors based on $\mathbf{y}^{\pm\ell}$ resulted in some increases in the estimated bias, *RMSE* and *MLE* of our return level exceedance probabilities, relative to those based on $\mathbf{y}^{\pm\ell}$, especially when using the most dissimilar series on which to base our prior specifications ($\mathbf{y}_4^{\pm\ell}$ and $\mathbf{y}_5^{\pm\ell}$). Here, biases in estimates of r^{-1} were notably larger than those using the objective priors or the informative priors based on $\mathbf{y}^{\pm\ell}$ (but least so for estimates based on the predictive return level and for the larger return periods), and values of the *RMSE* / *MLE* were always larger than those using informative priors based on $\mathbf{y}^{\pm\ell}$ (again, least so for estimates based on the predictive return level, especially for return periods $r = 1000$ and $r = 10000$). Informative priors based on the least dissimilar series ($\mathbf{y}_1^{\pm\ell}$ and $\mathbf{y}_2^{\pm\ell}$) yielded very similar results to those based on $\mathbf{y}^{\pm\ell}$.

3.3.5 Comparisons with POT

In this part of the simulation study we investigate the performance of our four methods for estimating the return level exceedance probability r^{-1} when a standard declustering scheme is employed to filter out a set of independent threshold excesses. Under a POT procedure, a cluster of extremes over some high threshold u is deemed to have terminated once at least κ consecutive sub-threshold observations have been made; from each cluster identified in this way the maximum is then carried forward into the analysis, the GP distribution being used as a model for the set of cluster peak excesses. Although in practice this is a commonly-used procedure to circumvent the problems of temporal dependence, as Fawcett and Walshaw (2012, 2016) discuss, not only is it wasteful of data (often leading to infeasibly wide credible intervals for quantities such as return levels) but parameter and return level estimates can be extremely sensitive to the choice of κ . Thus, we do not recommend a POT analysis at all, and we favour an approach as detailed in Section 2.2.2 of

this paper and used so far in this simulation study. However, we include some results based on declustered data here for information and comparison purposes.

Fig. 6 (bottom row) shows sampling distribution means for our estimates of r^{-1} across a range of return periods r , having declustered our simulated series $\mathbf{y}^\ell$ at each iteration $\ell = 1, \dots, 1000$ using $\kappa = 5$ and $\kappa = 20$. As before, in separate arms of the study we simulate series exhibiting asymptotic dependence (using the logistic / bilogistic model for consecutive extremes in the process) and asymptotic independence (using an $AR(1)$ process). However, since the aim is to eliminate dependence between the extremes pressed into use, we assume the extremal index $\theta \approx 1$ for our cluster peak excesses, and we bypass the stage in our analysis where we estimate the dependence parameter(s). Thus, the aim here is to compare results based on declustered data to those from Sections 3.3.1–3.3.2, in which *all* threshold excesses were used and the dependence structure estimated; we can also investigate the sensitivity of our estimators of r^{-1} to the choice of declustering interval κ . Regardless of the declustering interval used, the predictive return level consistently yields exceedance probabilities closer to the intended r^{-1} across the full range of return periods considered, with both the posterior means and modes resulting in relatively over-estimated exceedance probabilities. When declustering, all posterior summaries – including the predictive return level – yield exceedance probabilities that are more biased than those obtained having pressed all extremes into use; see the top row of Fig. 5 for a comparison.

3.3.6 Marginal domain of attraction assumption

So far, our simulated chains have always been drawn from a GP distribution marginally, which is the limiting distribution for excesses over a high threshold. In practice, our threshold excesses will in fact arise from a distribution in one of the *domains of attraction* (DoA) of the GP distribution; see, for example, Coles (2001, Ch. 3). Thus, for completeness, for both asymptotically dependent and independent extremes we also simulate chains $\mathbf{y}^\ell$ with Weibull, Fréchet and Uniform margins (representing, respectively, models from the Gumbel, Fréchet and Weibull DoA). Switching from GP margins to distributions in one of the DoA of the GP distribution did not seem to have any real effect on our estimators for r^{-1} , relative to the results shown in Figs. 3–5 and Table 2. Similarly, switching between the three DoA did not reveal anything over-and-above the differences we observed when changing the value of the shape parameter under a GP marginal assumption (see the “Other arms of the study: main findings” discussions in Sections 3.3.1 and 3.3.2). For example, after what we might reasonably expect from sampling variability, the results shown in Table 2 were

in line with analogous results using chains which had been marginally transformed to Uniform (Table 2 shows results for $\xi = -0.4$, giving Weibull-type tails with a finite upper endpoint).

3.3.7 Marginal structure: General remarks

The results reported in Table 2 and Figs. 3–6 compare our four return level summaries for simulated chains with relatively short, bounded tails (the GP marginals here have $\xi = -0.4$). As we discuss throughout Section 3.3, similar findings were obtained across most other parameters in our study design (recall that $\xi = \{-0.4, -0.1, 0.0, 0.1, 0.3\}$ and $u = \{u_{0.9}, u_{0.95}, u_{0.99}\}$). However, comparisons in some arms of the simulation study were clearly being influenced by the marginal shape parameter ξ . As might be expected, for much heavier-tailed margins – e.g. when $\xi = 0.3$ – the resulting posterior distribution for z_r was substantially more right-skewed, resulting in larger biases for estimates of r^{-1} based on the return level posterior mean ($p_{\bar{z}_r}$) and the return level 95% credible upper bound ($p_{z_{r,upper}}$). For these arms of the study, estimates of r^{-1} based on the posterior mode (p_{z_r}) out-performed the other estimative summaries, although estimates based on the predictive return level ($p_{z_{r,pred}}$) seemed to be generally less biased and with smallest error. Generally, as the value of ξ increases – giving rise to heavier-tailed margins and greater right-skew in the posterior for z_r – the performance of $p_{\bar{z}_r}$ and $p_{z_{r,upper}}$ deteriorate in terms of estimated bias, RMSE and MLE, but $p_{z_{r,pred}}$ retains the accuracy and precision observed in Table 2 and Figs. 3–6. Comparisons between our estimators do not appear to be sensitive to the scale of the underlying GP distribution. Although the GP marginal scale σ is held unit constant, as discussed in Section 3.2 excesses over u have a threshold- and shape-dependent scale σ^* . For arms of the study in which ξ was constant but σ^* varied, we did not see any real departure from the general findings reported in Table 2 / Figs. 3–6. In short: comparisons between our estimators are more sensitive to shape than to scale of the underlying GP distribution, but estimates of r^{-1} produced by the predictive return level seem relatively robust to changes in scale and shape.

4 Conclusions

4.1 General summary

In this paper we have discussed the merits of a Bayesian approach to inference on environmental extremes, and the natural extension to prediction such an inferential framework offers. In our experience, practitioners often find the standard reporting of return level estimates – a point estimate with some measure of uncertainty (e.g. a maximum likelihood estimate with standard error / 95% confidence interval,

or, within a Bayesian setting, the posterior mean and standard deviation / 95% credible interval) – difficult to work with in practice. Certainly, as we discuss in this paper, standard approaches such as POT analyses can yield estimates of return levels with extremely and unrealistically wide confidence / credible intervals, sometimes giving bounds which lie beyond the physical constraints of the variable being studied. Although Bayesian credible intervals have a more intuitive interpretation than frequentist confidence intervals (i.e. providing the stated probability coverage), our experience suggests that practitioners would prefer to work with a single point summary in which natural variability and estimation uncertainty have been properly accounted for. For this reason, Fawcett and Walshaw (2016) recommend the posterior predictive return level estimate as the most appropriate posterior summary to feed back to practitioners.

We build on earlier work presented in Fawcett and Walshaw (2016) in which an estimation strategy which attempts to maximise precision is outlined. Our recommended approach is to model all excesses over a threshold with the GP distribution, accounting for temporal dependence through estimation of the extremal index (either via a parametric model for extremal dependence or through direct non-parametric estimation of the extremal index). Where extremes vary seasonally, we recommend a piecewise seasonal approach to modelling (where appropriate), pressing threshold excesses from all seasons into use; other features, such as trends, can be simply captured through linear modelling of the GP scale parameter. We advocate a Bayesian approach to analysis, in which precision can be further increased through the specification of informative prior distributions for the GP parameters and from which predictive inference is neatly handled.

The main contribution of our work in this paper is to assess the performance of the posterior predictive return level relative to what we refer to as estimative return levels – standard point estimates taken directly from the return level posterior distribution. We do this through a large scale simulation study, in which data with various dependence structures, and tail behaviours, are simulated. We compare posterior predictive inferences for return levels to their estimative counterparts within the recommended modelling framework in Fawcett and Walshaw (2016), in which all excesses are modelled, but also within a more commonly-adopted POT modelling procedure. For a range of return periods r , on a fine scale, and across a range of temporal dependencies in the simulated data, we compare return level exceedance probabilities for return level summaries – specifically, the return level posterior mean, posterior mode, and posterior 95% credible upper bound – to those obtained from the posterior predictive return level, and to their expected values r^{-1} . Our general findings are that, for most commonly-observed levels of temporal dependence and for both asymptotically

dependent and independent extremes, the posterior predictive return level has exceedance probabilities much more in line with what we would expect to see than do the standard estimative posterior summaries (e.g. posterior mean/mode). In our simulation study, the posterior predictive return level also yields estimates of exceedance probabilities with much higher precision than the corresponding exceedance probabilities obtained from the estimative summaries. We believe the findings presented throughout Section 3 of this paper lend firm justification for the adoption of the posterior predictive return level as the best return level summary for practitioners, whether the modelling framework of Fawcett and Walshaw (2016) is adopted or a simple POT analysis is used. Further, if all excesses are used as in Fawcett and Walshaw (2016), but an incorrect assumption regarding the dependence structure is made, the posterior predictive return level still yields exceedance probabilities more in-line with what we would expect, compared to the other estimative summaries. The superiority of the predictive return level also seems to hold under informative prior specification / misspecification, and across different marginal assumptions.

4.2 Further thoughts

One of the practical advantages of the posterior predictive return level, as we discuss throughout this paper, is the incorporation of estimation uncertainty and natural variability into a single point estimate, perhaps to be used to aid structural design. Although the results of our simulation study in Section 3.3 go some way to indicate the superiority of the predictive return level relative to more standard point summaries from the return level posterior distribution, it might be useful for such point estimates to take account of the *consequences* of error. Indeed, in a machine learning or Bayesian decision theoretic context (e.g. Berger, 2010), the aim is to choose the decision function $\delta(x)$ which minimises the *a posteriori expected loss* for some model parameter ψ :

$$\int_{\Psi} L(\psi, \delta(x)) \pi(\psi|x) d\psi. \quad (13)$$

In Eq. (13), L represent a loss function: $L(\psi, \delta(x)) = (\psi - \delta(x))^2$ gives squared errors, although as we discuss in Section 3.3.1, for estimates of r^{-1} we might rather use linex errors since over- and under-estimation might not be equally serious. In a predictive setting, it is necessary to have a predictive version of Eq. (13). Conditioning on the observed (x) and averaging over the unknowns (e.g. parameter(s) ψ and future observations y), gives

$$\int_Y L(y, \delta(x)) f_Y(y|x) dy, \quad (14)$$

where $f_Y(y|x)$ is the posterior predictive density for y . The optimal decision $\delta(x)$ can then be seen as the action that

minimises this *predictive a posteriori expected loss*. From an inference point-of-view, $\delta(x)$ is a function whose output $\hat{\psi}$ is an estimate of ψ .

Obviously, this sort of approach for formally taking account of the consequences of error in our estimators for r^{-1} will be highly sensitive to the choice of loss function. For example, it can be shown that the posterior mean minimises Eq. (13) when L returns squared errors, and the posterior median when L returns absolute errors. Although we outline a rationale for using linex errors for our problem, to penalise over-estimation of r^{-1} more heavily than under-estimation, we feel that more work is needed to determine the suitability of linex errors here, and more generally a linex loss function for use in minimising Eq. (13) / Eq. (14).

The contribution of parameter uncertainty to the predictive return level can be estimated by comparing $\hat{z}_{r,\text{pred}}$ to what we call naïve return level estimates. Fig. 7 shows, for one arm of our simulation study, sampling distribution means for the predictive return level alongside sampling distribution means for this naïve estimator. Here, at each replication in the simulation study, rather than account for parameter uncertainty via Eq. (12) we assume that each of our marginal and dependence parameters are fixed at their posterior means; we substitute these means directly into Eq. (8) (of course, we could fix the model parameters at some other posterior summary, or indeed their likelihood modes). Thus, the difference between between the solid and dashed lines in the plots in Fig. 7 can be seen as the average contribution to $\hat{z}_{r,\text{pred}}$ of the implicit allowance for uncertainty in parameter estimation. The results are shown for asymptotically dependent chains simulated according to a bivariate logistic model for consecutive pairs in the series, for three strengths of dependence; however, discrepancies of similar magnitude were observed for other arms of the study (although for heavier-tailed chains the naïve estimator was more sensitive to the choice of posterior summary used to fix the model parameters). In a real data context, plots of $\bar{z}_r$ against $\hat{z}_{r,\text{pred}}$ can be used to reveal such contributions to the predictive return level.

Acknowledgements We would like to thank two referees for making excellent suggestions, resulting in a significantly improved manuscript.

References

- [1] Ancona-Navarrete, M.A. and Tawn, J.A. (2000). A Comparison of Methods for Estimating the Extremal Index. *Extremes*, **3**, pp. 5–38.
- [2] Beirlant, J., Goegebeur, J., Teugels, J., Segers, J., De Waal, D. and Ferro, C. (2004). *Statistics of Extremes*. Wiley, New York.
- [3] Berger, J.O. (2010). *Statistical Decision Theory and Bayesian Analysis*. Springer, London.
- [4] Coles, S.G. (2001). *An introduction to statistical modeling of extreme values*. Springer, London.

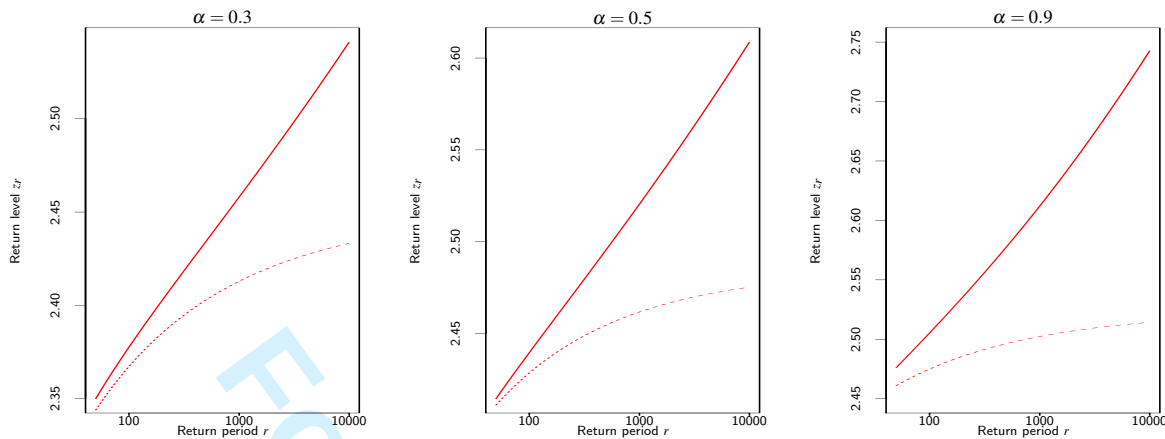


Fig. 7 Sampling distribution means for predictive return levels $z_{r,\text{pred}}$ (solid lines) and “naïve” return levels (dashed lines). Simulated data are constructed with asymptotic dependence according to the bivariate logistic model with dependence α . Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

- [5] Coles, S.G. and Powell, E.A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review*, **64**, 1, pp. 119–136.
- [6] Coles, S.G. and Tawn, J.A. (1996). A Bayesian Analysis of Extreme Rainfall Data. *J. R. Statist. Soc., C*, **45**, pp. 463–478.
- [7] Davison, A.C. and Smith, R.L. (1990). Models for Exceedances over High Thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, pp. 393–442.
- [8] Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Stat. Sci.*, **27**, pp. 161–186.
- [9] Eastoe, E. F. and Tawn, J.A. (2012). Modelling the distribution for the cluster maxima of exceedances of sub-asymptotic thresholds. *Biometrika*, **99**, **1**, pp. 43–55.
- [10] Eugenia Castellanos, M. and Cabras, S. (2007). A default Bayesian procedure for the generalized Pareto distribution. *J. Stat. Plan. Inf*, **137**, **2**, pp. 473–483.
- [11] Fawcett, L. and Walshaw, D. (2006). A hierarchical model for extreme wind speeds. *J. Roy. Statist. Soc., C*, **55**, **5**, pp. 631–646.
- [12] Fawcett, L. and Walshaw, D. (2006a). Markov chain models for extreme wind speeds. *Environmetrics*, **17**, **8**, pp. 795–809.
- [13] Fawcett L, Walshaw D. (2008). Bayesian inference for clustered extremes. *Extremes*, **11**, pp. 217–233.
- [14] Fawcett, L. and Walshaw, D. (2012). Estimating return levels from serially dependent extremes. *Environmetrics*, **23**, **3**, pp. 272–283.
- [15] Fawcett, L. and Walshaw, D. (2016). Sea-surge and wind speed extremes: Optimal estimation strategies for planners and engineers. *Stoch. Env. Res. Risk. Assess.*, **30**, pp. 463–480.
- [16] Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Statist. Soc., B*, **65**, pp. 545–556.
- [17] Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, New York.
- [18] Leadbetter, M.R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**, pp. 431–476.
- [19] Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. *Annals of Statistics*, **3**, **1**, pp. 119–131.
- [20] Sang, H. and Gelfand, A.E. (2009). Hierarchical Modeling for Extreme Values Observed over space and time. *Environmental and Ecological Statistics*, **16**, pp. 407–426.
- [21] Sang, H. and Gelfand, A.E. (2010). Continuous Spatial Process Models for Extreme Values. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, pp. 49–65.
- [22] Smith, E.L. and Walshaw, D. (2003). Modelling Bivariate Extremes in a Region. *Bayesian Statistics*, **7**, pp. 681–690.

- [23] Smith, R.L. (1992). The Extremal Index for a Markov Chain. *J. Appl. Prob.*, **29**, pp. 37–45.
- [24] Smith, R.L. (1999). Bayesian and frequentist approaches to parametric predictive inference (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) *Bayesian Statistics*, **6**, pp. 589–223.
- [25] Smith, R.L., Tawn, J.A. and Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, pp. 249–268.
- [26] Walshaw, D. (1994). Getting the Most From Your Extreme Wind Data: A Step by Step Guide. *J. Res. Natl. Inst. Stand. Technol.*, **99**, pp. 399–411.
- [27] Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Am. Stat. Ass.*, **81**, pp. 446–451.

Appendix

Here, we present some results from a simulation study to support the discussion in part (i) of Section 2.2.2 and the main simulation study in Section 3. The aim is to establish a simple polynomial approximation to the extremal index θ dependent on the parameter(s) in a model being used to capture first-order temporal dependence. For example, in the case of asymptotic dependence, we might assume a logistic model with dependence parameter α (see Eq. 5) for consecutive extremes in the process. Then, given an estimate of this dependence parameter, we require an associated estimate of the extremal index – along with estimated marginal parameters from the GP distribution – to estimate return levels via Eq. (8) or Eq. (9).

Define (arbitrarily) x_n such that $F^n(x_n) = 1/2$ in Eq. (1). Then, using Eq. (1), we can define

$$\theta_n = -\frac{\log \Pr(\max\{X_1, \dots, X_n\} \leq x_n)}{\log 2}, \quad (15)$$

and so $\theta_n \rightarrow \theta$ as $n \rightarrow \infty$. We can use Eq. (15) to investigate the relationship between θ and α in the logistic model. Specifically, we simulate N first-order Markov chains, each of length n , with logistic dependence α governing the strength

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

of **temporal dependence** present in the extremes of the process; then the probability in the numerator of Eq. (15) is estimated as the proportion of simulated chains whose maximum does not exceed x_n . The first plot in Fig. 8 shows the results of such simulations for $\alpha = \{0.05, 0.10, \dots, 1.00\}$, using $N = n = 10,000$; the other two plots show corresponding results when using the bilogistic model (see Section 2.2.2) and another model occasionally used for bivariate extremes (the negative logistic model with dependence parameter ρ ; see, for example, Coles (2001, Ch. 8)). The smooth line in each of these plots shows a fitted polynomial, giving that in Eq. (6) for the logistic model and as used in the simulation study in Section 3. In the case of the logistic model, and as a check, the simulated values and resulting fitted polynomial are compared to limiting values obtained via a computationally intensive Fourier transform method outlined in Smith (1992). We use polynomial relationships rather than smoothing splines (for example) because they are extremely simple, and as the comparisons to Smith's results show, are more than adequate.

For the asymptotically independent case we use a Gaussian $AR(1)$ process in our simulation study in Section 3, with lag 1 autocorrelation A . To establish a polynomial approximation for the extremal index in this setting, for each value $A = \{0.05, 0.10, \dots, 0.95\}$ we simulate $N = 10,000$ $AR(1)$ processes, each of length $n = 10,000$. Then, using the approach of Ferro and Segers (2003) (which we discuss in part (ii) of Section 2.2.2 of this paper), our approximation to the extremal index θ for each value of A is the sampling distribution mean of $\bar{\theta}$ obtained via Eq. (7). Of course, Gaussian $AR(1)$ processes exhibit asymptotic independence and thus $\theta = 1$ regardless of the value of A . However, as Ancona-Navarrete and Tawn (2000) discuss, such processes might exhibit dependence above thresholds of practical interest, resulting in estimators such as that in Eq. (7) estimating $\theta(u_p)$ (rather than θ itself), a threshold-based penultimate approximation to θ . Fig. 9 shows the results of this simulation study for three thresholds u_p , where we use $p = 0.9, 0.95$ and 0.99 . Again, simple polynomials are fitted to the points in each plot to obtain approximations to $\theta(u_p)$ depending on the value of A .

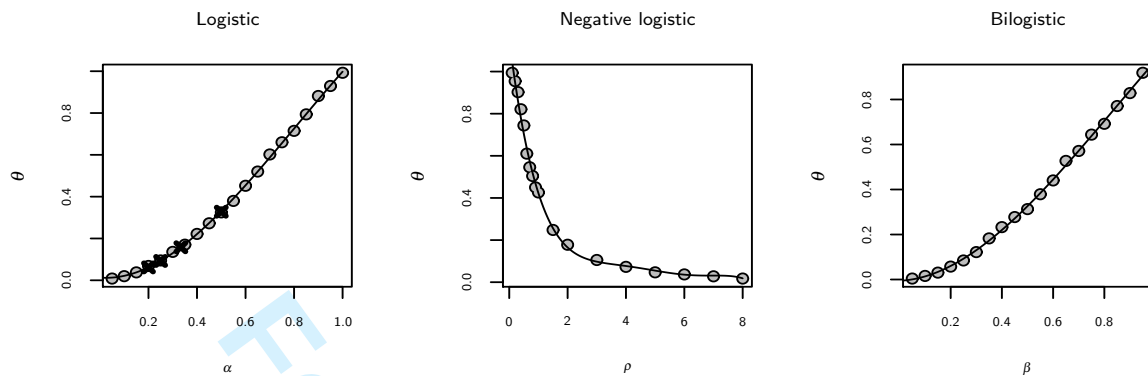


Fig. 8 Simulated values of the extremal index θ for α (logistic), ρ (negative logistic) and β (bilogistic with $\alpha = 0.6$). The solid lines correspond to fitted polynomials: $\theta = 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3$; $\theta = 1.153 - 1.107\rho + 0.463\rho^2 - 0.096\rho^3 + 0.010\rho^4 - 0.0004\rho^5$; $\theta = -0.005 + 0.045\beta + 1.539\beta^2 - 0.607\beta^3$. The crosses in the first plot show limiting values of θ for some values of α in the logistic model, as derived in [Smith \(1992\)](#).

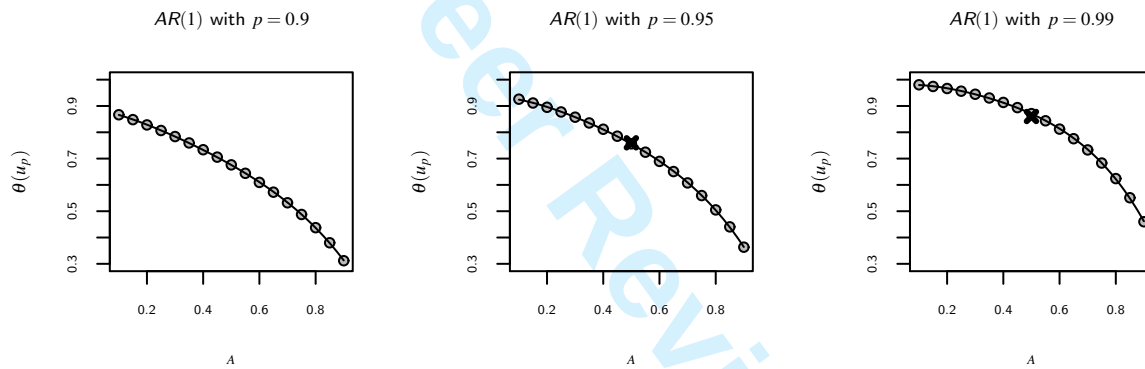


Fig. 9 Simulated values of the threshold-based extremal index $\theta(u_p)$ for an $AR(1)$ process with a threshold set at the 90% marginal quantile $u_{0.9}$ (left); 95% marginal quantile $u_{0.95}$ (middle); 99% marginal quantile $u_{0.99}$ (right). The solid lines correspond to fitted polynomials: $\theta(u_{0.9}) = 0.891 - 0.168A - 0.920A^2 + 1.234A^3 - 0.886A^4$ (left); $\theta(u_{0.95}) = 0.935 - 1.246A^2 + 1.629A^3 - 1.141A^4$ (middle); $\theta(u_{0.99}) = 0.976 + 0.146A - 1.223A^2 + 1.702A^3 - 1.364A^4$ (right). Where shown, crosses show values obtained in extensive simulations in [Ancona-Navarrete and Tawn \(2000\)](#).

Investigating the properties of Bayesian posterior predictive return levels for environmental extremes

We thank the reviewers and the Editor for taking the time to read our paper. We feel that the minor revisions they have suggested have greatly enhanced the paper and make for a much more complete manuscript. Although the recommendation was for minor revisions to the original submission, we have taken much time to carefully review each and every query raised by the reviewers. Some of the reviewers' comments required additional simulations, and so the revised manuscript has taken longer to prepare than we anticipated – we thank the journal for their patience here.

Any textual changes or additions to the paper are highlighted in magenta text in the revised manuscript. Our point-by-point response to the reviewers' reports is given below, where the reviewers' original comments are shown in *italics*.

Reviewer 1's report

We thank the reviewer for their kind words about our paper. We were pleased to learn the reviewer found our work interesting, well-organised and clearly written. We now give our responses to each of the reviewer's queries.

- (a) *The paper emphasizes an important selling point of the predictive return level: that it is a point estimate capturing both estimation uncertainty and natural variability. How can the user assess how much each of these sources contributes to it? Also, if a point estimate is needed for structural design, should it not take account – informally or formally through a loss function – of the consequences of error, and how can that be done in the predictive setting?*

First of all, untangling the contribution to the predictive return level of estimation uncertainty and natural variability. Implicit in any return level estimation procedure (e.g. a Bayesian analysis resulting in a posterior distribution for z_r or a predictive return level $z_{r,\text{pred}}$, or perhaps a frequentist estimation framework resulting in a maximum likelihood estimate for z_r) is the allowance for natural variability; we have a fitted distribution and we are estimating the return level as a high quantile from that fitted distribution. The added feature of a predictive return level is the allowance for uncertainty in parameter estimation since we integrate over our posterior distribution. It is this uncertainty in prediction, due to parameter uncertainty, which $z_{r,\text{pred}}$ offers over (for example) $\bar{z}_r$ or $\hat{z}_r$. In the Conclusions section of our paper, we have added some discussion (Section 4.2) which we believe goes some way to clarifying this. For example, in Figure 7 (new to the paper) we have produced plots of the predictive return level alongside what we term a “naïve” return level estimate, which does not take into account estimation uncertainty. Rather, this naïve estimator assumes the GP marginal parameters, and the extremal index, are fixed at their posterior means, the resulting return level estimates being obtained as a function of these posterior means. We discuss that the difference between the two curves (predictive and “naïve”) constitutes the contribution of estimation uncertainty to the predictive return level. We hope this addition to the paper is sufficient for the reviewer; of course, we are happy to add more text and perhaps results from more arms of the simulation study should the reviewer think this is necessary. Although these are results from our simulations, showing the sampling distribution means for our predictive and naïve return levels, in practice - with real data - the contribution to $z_{r,\text{pred}}$ of estimation uncertainty

could be observed as the difference between the predictive return level and perhaps the return level posterior mean.

Secondly, taking into account the consequences of error. We think the reviewer makes an excellent point here, and our response to this query overlaps slightly with our response to query (c) below. As the reviewer suggests, we think the best way to do this, formally, would be through the incorporation of a loss function. We now include a specific discussion in the paper on our thoughts here (please see the first paragraph of Section 4.2), and we hope the reviewer agrees this makes the manuscript much more complete. We do not directly implement any of the material discussed here as we feel this would be too much work for the current paper, although (as we discuss in Section 4.2) our intention is certainly to follow this up in future work. As we discuss, the choice of specific loss function might need careful consideration – in response to query (c) below we now examine “linex errors” for our four estimators, to enable over-estimation of r^{-1} to be punished more severely than under-estimation (as opposed to squared errors which punish over- and under-estimation symmetrically), and we give a practical rationale for doing so. Thus, linex errors might be incorporated more formally into a loss function for finding an estimator which minimises the Bayesian *a posteriori expected loss* (see Eq. (13) of the revised manuscript), or perhaps the *predictive a posteriori expected loss* (see Eq. (14)) as an equivalent in the predictive setting. These ideas will form part of the second author’s work in her new Ph.D position at our institution.

- (b) *Can any general features of the relationship between estimators be distilled from the simulation results? For example, what is the shape of the posterior distribution of z_r and how does it appear to affect properties of the various estimative return levels? Can anything be said about the effect of the posterior distribution of model parameters on the relationship between predictive and estimative return levels? Is the relationship more sensitive to shape than to scale of the underlying GPD? What is the likely effect of using an informative prior distribution and presumably therefore a more concentrated posterior distribution?*

We thank the reviewer for asking these interesting questions which we feel warrant the inclusion of some new text in the paper. We have added a new subsection to the results Section of the simulation study – Section 3.3.7 – which gives some general remarks about the shape of the posterior distribution for z_r and its effect on the properties of the estimative return levels. We also discuss the effects of posterior inferences for the marginal GP shape parameter ξ on our comparisons between the predictive and estimative return levels, and we discuss the sensitivity of such comparisons to changes in the marginal GP scale parameter σ^* also. In short, for chains simulated with the heaviest tails ($\xi = 0.3$), the resulting posterior for z_r was substantially more right-skewed than for other values of ξ (and, in particular, for the results shown throughout most of Section 3.3 in which $\xi = -0.4$). This resulted in larger biases for estimates of r^{-1} based on the posterior mean and 95% credible upper bound than the posterior mode and predictive return level. Estimators did not seem to be sensitive to changes in the GP scale σ^* . The reviewer is referred to Section 3.3.7 of the revised manuscript for more details, as well as other remarks made throughout the discussion in Sections 3.3.1 and 3.3.2.

The reviewer also asks about the effects of an informative prior specification. We thought this was a very good question, and an obvious oversight in our initial submission! Thus, we ran extra arms of the simulation study to examine the effects of more informative prior specifications on the estimators of r^{-1} . The results are now discussed extensively in Section 3.3.4 of the revised manuscript. To summarise, using an informative prior specification (obtained by making use of a ‘paired’ simulated dataset drawn from the same marginal distribution and with the same dependence structure), resulted in no apparent change in the bias of our estimators for r^{-1} , relative to those reported in Table 2 when flat, objective priors were used. However, there were some substantial gains in precision as a result of more concentrated posterior distributions; specific results

are discussed in Section 3.3.4. In response to comments made by reviewer 2, we also examined the effects on our estimates of r^{-1} of ‘incorrect’ prior specifications (in other words, priors which could be considered ill-informed) and although we noticed increased biases in estimates of r^{-1} , and a loss of precision, across the board, estimates based on the posterior predictive return level were always superior. Again, full details can now be found in Section 3.3.4 of the paper.

- (c) *Is mean squared error the best measure of accuracy of the various estimators p_* of r^{-1} ? The target is a probability close to zero, and over- and under-estimation may not be equally serious.*

We thank the reviewer for this comment. On the recommendation of reviewer 2, we now provide root mean squared error (*RMSE*), rather than mean squared error, as this measure is on the same scale as our estimated biases. However, alongside *RMSE* we now also present results for the mean linear-exponential error, or mean *linex* error (*MLE*). Full details are given in Section 3.3.1 of the revised manuscript, with a reference. As we discuss, unlike squared errors, linex errors can punish errors due to under- and over-estimation asymmetrically. In Section 3.3.1 of the revised manuscript we argue that – from a safety point-of-view – over-estimation of r^{-1} might be more serious than under-estimation, and so we choose a specific form of linex error which punishes over-estimation more severely. Table 2 reports the *RMSE* and *MLE* for some arms of our simulation study; when these error measures, and estimated biases, are inspected, the desired effects of our linex errors are clearly demonstrated. Our response here partially overlaps with our response to comment (a) above, and we discuss the potential role of linex errors in the context of loss functions and minimising Bayesian expected loss; see Section 4.2 of the revised manuscript.

- (d) *Further comments and suggestions are listed below.*

- (1) *Fig. 1 and later: knots or ms^{-1} ? How should the effects of seasonality and dependence in the right panel be disentangled?*

Thank you – we have checked, and the units are ms^{-1} . All plots, figures and discussions have now been corrected. As for the disentangling of seasonality and dependence in the right panel of Fig. 1: as a check, we produced a scatterplot of consecutive pairs of values after pre-whitening our wind speed series, where the extremes within each month were transformed to unit exponential using the monthly-varying fitted GP parameters. The autocorrelation in the plot after pre-whitening was not dissimilar to that shown in Fig. 1. If the reviewer thinks it important, we could replace the plot currently shown in Fig. 1 with the plot based on the pre-whitened series (we were reluctant to do this because this would require extra discussion to describe the pre-whitening process etc.; since the pre-whitened version of the plot did not reveal anything over the original plot, we didn’t think this was an important / necessary addition).

- (2) *p3, ll41-42: should $\theta \in [0, 1]$ rather than $\theta \in (0, 1)$?*

Yes, you are correct. Thank you for spotting this – we have now corrected this!

- (3) *p4, ll8-9 and later: say that f in eq(4) denotes, as appropriate, a joint or marginal density function?*

Yes, good point; we have now added text where appropriate to make this clear.

ll14-15: would ‘dependence of consecutive observations’ be better understood than ‘serial correlation’?

Yes, we agree. Thus, we have replaced the text at this point – and all other places in the paper – with the words suggested by the reviewer (or similar).

ll22-23 and later: the crucial censoring device makes it possible to apply a model built for consecutive observations when data are from exceedances that occur only intermittently. It

is described here in the context of the Markov Chain model for asymptotic dependence but presumably is used also for the AR model described later. If true, it seems important to make that clear (or if not, to say how else the problem is solved).

We thank the reviewer for pointing this out, and we have now added some text at this point in the paper (top of the second column of text on page 4 of the revised manuscript).

ll46-47: should asymptotic dependence and asymptotic independence be defined?

We think some description is probably needed, and so we have added some text at this point in the paper (now near the bottom of the first column of text on page 4 of the revised manuscript).

ll48-49: readers may be puzzled by the discrepancy between marginal normality of the AR process and the GP margins of the model apparently still in force at this point for threshold excesses. Some pages later the authors say that a marginal transformation will be used to convert to GPD form over the threshold, but it would be helpful to say so here.

Thank you – on re-reading our paper, we agree that this is confusing and so we have added some text at this point for clarification.

- (4) *p7, ll43-44: it might be helpful to say why results in Coles & Tawn (1996) and Fawcett and Walshaw (2016) suggest that exceedance probabilities of $z_{r,\text{pred}}$ might be much less than r^{-1} .*

We agree, and so we have added some text at this point. Specifically, we say: “we might expect $z_{r,\text{pred}}$ to give exceedance probabilities considerably lower than r^{-1} ; implicit in the predictive return level is the allowance for uncertainty in parameter estimation, resulting in higher estimates of z_r and correspondingly lower estimates of r^{-1} .”

- (5) *p8, §3.2: what is the basis for choice of the particular values used for N , L and S ?*

Short answer: computational feasibility! At each replication L in the simulation study we simulate a Markov chain and then perform a full MCMC procedure on that simulated chain. We actually initialise these chains at the maximum likelihood estimates of the GP marginal parameters and the dependence parameter, to help expedite convergence to the stationary distribution, and so after some extensive pilot runs we believed $S = 10,000$ was an adequate number of MCMC iterations here. Given the scale of the study ((i)asymptotic dependence / independence, and using two different models for the former; (ii) assessment of results across many different levels of dependence; (iii) objective / informative priors; (iv) examination of the sensitivity of our results to declustering; (v) sensitivity of our results to the marginal specification, including five different shape parameters ξ , the testing of different margins from models within the domains of attraction of the GP distribution and three different thresholds), we believed anything more than $L = 1000$ would be impractical and that for relative comparisons to be made between the different estimators of r^{-1} , 1000 replications would suffice. In fact, the number of replications used here is the same as that used in a previous study, published in Fawcett and Walshaw (2012), to which we wanted to compare some results on a like-for-like basis. As for N , the number of years in our reference series $\mathbf{y}^{(\text{Ref})}$: we wanted this to be as large as possible so our estimators p_* , based on a comparison of each estimate of z_r to the annual data within $\mathbf{y}^{(\text{Ref})}$, would be as accurate as possible. Although for each arm of our study the generation of $\mathbf{y}^{(\text{Ref})}$ is a one-off procedure, the comparison of each of our estimates of z_r^ℓ to $\mathbf{y}^{(\text{Ref})}$, $\ell = 1, \dots, 1000$, is a fairly expensive process across all arms of the study. And as we discuss above, the $L = 1000$ replications within the study are computationally expensive anyway! We admit that, ideally, L and N would be larger; however, we think for the purposes of the comparisons we are making in this paper these choices are at least adequate. To check this, the second author conducted an experiment for one arm of the study using $L = 10,000$ and $N = 100,000$; results (at least for this arm) did not reveal anything over-and-above the results shown in our original study.

Last line: Fig. 2? – do the authors mean Table 2 (and Fig. 3)?

Yes, thank you for spotting this! We have now corrected the text here.

- (6) *p9, Figs. 3 and 4: this reader found the captions of these Figures confusing. What do dots and lines in the first line refer to?*

The plots in Figs. 3 and 4 of the original submission were quite difficult to read, and on the recommendation of reviewer 2 these plots have been constructed on a different scale altogether (mainly because the plots were *impossible* to read for very large r). Also, in the original plots, scaling and re-sizing issues made the dots look more like short horizontal lines; in the new version of the paper, the size of these plots has been determined within R at the time of their construction, rather than re-sized / scaled within $\text{\LaTeX}$ afterwards, so hopefully things are much clearer now. In the first line of the captions for Figs. 3 and 4, we now refer to “bullets” rather than “dots”, and we also state clearly that the lines cover the range from the 2.5% and 97.5% quantiles in the sampling distributions for our estimators for r^{-1} . We hope this has made the captions less confusing for the reviewer.

- (7) *p11: Is Table 3 essential to the main argument?*

Table 3 is not essential to the main argument at all and, on reflection, we think it can distract attention away from the main argument. Thus, we have removed this table altogether from the revised manuscript.

- (8) *p14, §4: when extremes vary seasonally the authors recommend a piecewise seasonal analysis using threshold excesses from all seasons. Is that always wise? Extremes of some environmental variables may be generated by distinct physical mechanisms at different times of the year, so that the apparent transfer of information between seasons and the accompanying apparent gain in precision may actually be illusory and long extrapolation misleading.*

Of course, the reviewer is right here. We would argue that for our wind speed extremes such an approach is valid: in temperate climates (e.g. the U.K.) it is usually the same alternating sequence of anticyclones and depressions which causes storms, regardless of the time of the year, and it is just the scale of these systems which changes. Thus, we believe including observations which are extreme *given their time of year* is informative to the analysis. We are very careful not to over-generalise our approach here, and in the second paragraph of Section 2.2.2 we now say:

“In the presence of seasonal variability (as shown in Fig. 1 for our wind speed extremes), Fawcett and Walshaw (2016) recommend a seasonal piecewise approach to modelling, with a unique GP model for extremes within each season. Of course, this should only be attempted where there is confidence that it is the same physical mechanism generating extremes at different times of the year, seasonal variability in the extremes arising as a result of just a change in the scale of this mechanism. This assumption seems reasonable for wind speeds in temperate climates (e.g. the UK), where it is usually the same alternating sequence of anticyclones and depressions leading to most of the storms which occur throughout the year; the seasonal variability comes from the severity of these systems.”

Elsewhere in the text where we refer to our seasonal piecewise approach to modelling, we are cautious not to suggest the general applicability of this approach.

- (9) *p9, §4: the simulations of §3 were for stationary data and objective priors. Can anything be said about the likely effect of non-stationarity and informative priors on conclusions from them? (This partly overlaps with a point in (b) above.)*

The reviewer makes very good suggestions here. In our simulations we did not account for non-stationarity (e.g. trend or seasonal variation). Given the length of the manuscript already,

and the computation time required to run further simulations, we have also refrained from considering non-stationarity in the revised version of the paper. Of course, we could have generated data from the seasonal piecewise model we apply to our wind speed extremes in §2 of the paper; however, given the reviewer’s comments (and our response) in (8) above, the method chosen for handling such variability in practice is highly dependent on the variable being studied (and the location at which it was measured), perhaps making it difficult to generalise any findings from the simulation study. Trends could have been built into our generated sequences, perhaps by allowing the GP scale parameter σ^* to be time-dependent. However, as already discussed the computation time and manuscript space necessary to cover this were, we felt, prohibitive.

Much more important, and of practical interest (we felt), were the sensitivity of our results to different prior specifications. We therefore prioritised this over the effects of non-stationarity, and in the results Section of the simulation study we now include a subsection specifically devoted to prior specification. This as something reviewer 2 also picked up on, which also helped us prioritise this over investigations into the effects of non-stationarity. Please see §3.3.4 of the revised manuscript and our comments to query (b) above.

(10) *p15, References: references [10] and [11] look to be to the same paper.*

We thank the reviewer for spotting this. References [10] and [11] were indeed duplicates, and we have removed [11].

Reviewer 2's report

We thank the reviewer for taking the time to read our paper, and for their kind remarks about it! Below, we give our response to each of the reviewer's queries. Note that where the reviewer is referred directly to the revised manuscript for our response (e.g. Section 3.3.4 in our response to (d) below), the associated text in the paper is clearly highlighted in **magenta**.

- (a) *(minor) You should go through the entire paper and check and clean the writing. For example, in the first paragraph at the top right of page 4, you have 'We (also) note here ...' three times; this really isn't needed, and similar small tersenings could be made throughout. Also, the choice of words is sometimes odd, e.g., the use of 'credibility of the poster mode' in the first column of page 11, because credibility has a specific technical meaning in a Bayesian context, and it is not the meaning here, and anyway you really mean 'usefulness' or 'utility' or 'value' or 'suitability' rather than 'credibility. This means a careful reading to catch things like this. Also check that the references are appropriate: e.g., on page 5 you suggest that the reader go to a nearly 25-year old paper for information about MCMC, a topic on which there are now many books, so maybe some of these could be cited instead?*

We thank the reviewer for their comments here. We have now reviewed the entire manuscript carefully, making textual changes to clean and tighten our writing. Our MCMC reference in the original submission has now been replaced with a text book which is on the reading list for a Master's course in Modern Bayesian Inference at our University; we have been told this book is a commonly-used reference text for MCMC.

- (b) *(a bit less minor) In Figures 3 and 4, I think it would be better to present the ratio or p^* to $1/r$ (i.e., plot rp^*), since (i) presumably this is more important than the absolute difference between the values, and (ii) the graphs are impossible to read anyway for very large r . Maybe the vertical axes need to be logarithmic (and likewise for Figure 5)? Maybe clarity would be better if you showed the RMSE in the Tables (as it's on the same scale as the bias);*

We completely agree with everything the reviewer says here, and we have made the necessary changes. Our plots in Figs. 3 and 4 now show our estimates on a logarithmic scale, to make the plots legible for all values of return period r , and we have presented the ratio of p^* to $1/r$ as the reviewer suggests. We have also made sure the size of the plots is determined at the generation stage in R, as opposed to re-sizing within L^AT_EX, to ensure maximum clarity. Given the information being displayed in the plots, some readers may need to give the plots careful attention if they intend to understand fully what they reveal, but with the reviewer's recommendations (and with careful treatment being given to the production process of the plots) we believe the plots in the revised manuscript are a big improvement over those displayed in the original submission. In Table 2, results are now shown for *RMSE* (as opposed to *MSE*) for the reason the reviewer suggests. We also give results for the *mean linex error*, in response to a comment made by reviewer 1 (to penalise over-estimation of $r-1$ more heavily than under-estimation; see §3.3.1 for our rationale here).

- (c) *(less minor still) so far as I can tell, the simulations are always from the marginal GPD distribution, rather than from a marginal distribution in the domain of attraction (DA) of the GPD (e.g., the Weibull). I hesitate to ask you to add yet more simulations, but since data are always in fact going to come from the DA of the extremal model, it seems important to assess the effect of the DA assumption both for the margins and the joint distribution (the latter seems to be done, so far as I can understand) and both. Perhaps in Supplementary Material—or perhaps I've misunderstood and this is in fact what you do;*

The reviewer makes an excellent point here. In response, we have added more simulations to our study and summarised the results (as briefly as we can) in a new section of the paper. We refer the reviewer to Section 3.3.6 of the revised manuscript.

- (d) *Bayesian output depends on the prior. Do you assess how badly your approach can go wrong if the prior is mis-chosen? This could be informative for possible users of your methods.*

Again, an excellent suggestion, and something we completely overlooked in our original submission. We have added more simulations to our study to investigate the effects of informative prior specification - and we consider the case where such a specification may be ‘good’ or ‘bad’. Rather than repeat here what is already written in the revised manuscript, we refer the reviewer to Section 3.3.4 of our new submission.