

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 U.K., 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 · Exceedance probabilities · 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 generalised 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 serial correlation (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

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

on maps over a spatial grid. The increasing sophistication of models to address issues such as serial correlation, 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.

Of course, 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 both uncertainty in parameter estimation and randomness in future observations. 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 choose to perform analyses within a Bayesian setting.

[Coles and Powell \(1996\)](#) provided 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, and 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 \(2006; 2006a; 2008; 2016\)](#) 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 location in the southwest of the UK. This Section will include a primer in extreme value techniques and associated modeling 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 recommendations in Section 4.

2 Practical motivation and modeling

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

Figure 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 Figure 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

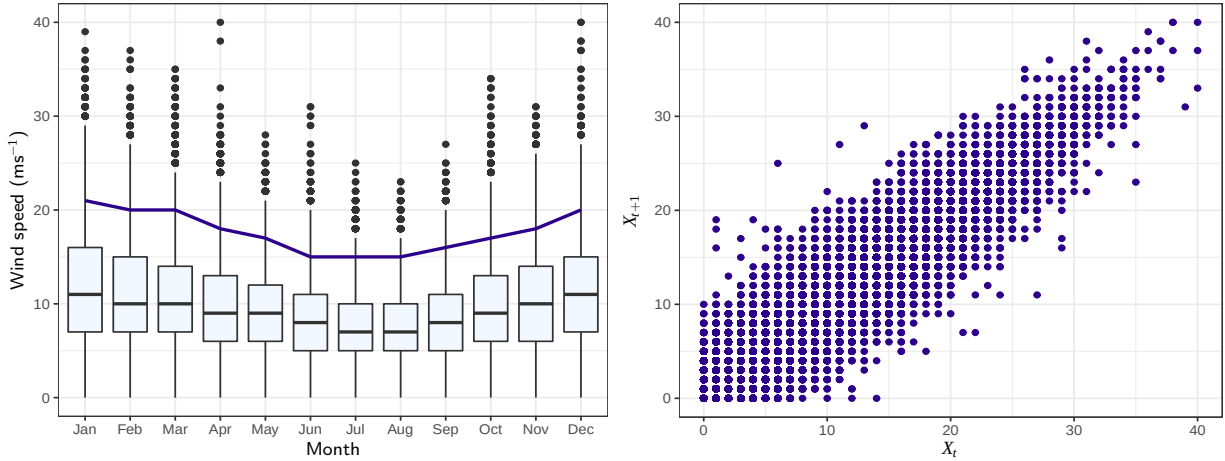


Fig. 1 Boxplots (by month, left) and each observation plotted against its lag 1 counterpart (right) for a series of hourly gust wind speed maxima observed at Yeovilton between January 1st 2003–December 31st 2012 (inclusive). The blue lines in the first plot correspond to high thresholds which have been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

ability to provide practitioners with a single point summary having all uncertainty due to parameter estimation *and* randomness in future observations ‘built in’, through the predictive return level.

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^θ}). 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 serial correlation 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 modeling 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 :

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

where ψ is a parameter vector containing marginal and dependence parameter(s). Appealing to bivariate extreme value theory, transformation from GP to standard Fréchet margins gives a range of models to use for the serial correlation between 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; 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. In this case, $\theta = 1$, 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 .

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). We also note the availability of graphical diagnostics to assess the assumption of asymptotic dependence / independence (e.g. $\chi/\bar{\chi}$ plots; see [Coles, 2001, Ch. 8](#)).

(ii) Direct estimation of the extremal index

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 first 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

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

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 Figure 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. For example, assuming either (i) or (ii) above to capture temporal dependence, estimates of $(\sigma_m, \xi_m, \theta_m)$, $m = 1, \dots$ might be obtained for each season m , 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 Figure 1, long term 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 $x = z_r$:

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

where H_m is the GP distribution function in season m with parameter set $(\lambda_{u_m}, \sigma_m, \xi_m)$, θ_m / n_m are the extremal index / number of observations in season m , and M is the number of seasonal units over which we partition our data. 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) and summarised in the Appendix, 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) imply 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 serial correlation. Given the seasonal variability observed in the wind speed extremes, we adopt a seasonal piecewise approach to modelling as discussed in Section 2.2.2. Specifically, following discussions about the UK wind climate in Walshaw (1994), we use a seasonal unit of one month, assuming stationarity of wind speed extremes *within* each month. As Walshaw (1994) argues, for wind speed extremes in the UK such an approach can provide a suitable means to capture this variation, the assumption of stationarity *within* months usually being reasonable. 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 Equation (9). Details on MCMC techniques are now extensively published (e.g. Smith and Roberts, 1993), 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 $(\eta_m, \xi_m, \alpha_m)^\top$ based on observations on wind speed extremes made at a nearby location. Specifically, we use:

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

$m = 1, \dots, 12$. The components of the mean vector $\boldsymbol{\mu}$ 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.

Equation (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 (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 uncertainty due to parameter estimation and randomness in future observations. Although (10) is analytically intractable, it can be approximated since we have estimated the posterior distribution using MCMC. Regarding our sample $\psi^{(1)}, \dots, \psi^{(S)}$ as realisations from the stationary distribution $\pi(\psi | \mathbf{x})$, 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 $\hat{z}_{r,\text{pred}}$. Figure 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 Figure 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 (i) uncertainty in parameter estimation, and (ii) randomness in future observations. However, open questions remain about 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} for some return periods r , 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.

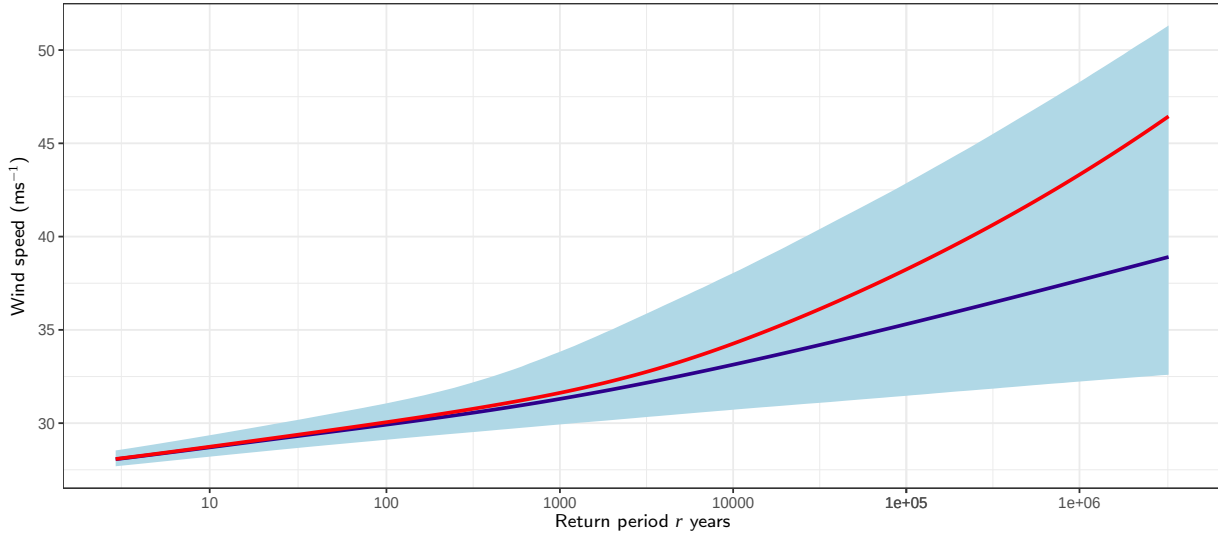


Fig. 2 Means (blue) taken from the MCMC samples of the posterior distributions for return levels z_r across a range of return periods r , with 95% credible intervals (outer light blue shaded area). Also shown, in red, are corresponding estimates of the predictive return levels $z_{r,\text{pred}}$. Units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

			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
		95% CI	(28.49, 29.85)	(28.85, 30.53)	(29.36, 31.68)	(29.92, 33.83)	(30.72, 38.05)
	$\hat{z}_{r,\text{pred}}$		29.13	29.65	30.47	31.62	35.26
Bradfield (after Fawcett & Walshaw, 2016)	$\hat{z}_r$	Post. mean	96.89	103.46	112.16	128.13	—
		Post. mode	94.82	99.91	108.22	123.31	—
		95% CI	(94.96, 98.85)	(94.11, 116.06)	(95.22, 125.55)	(117.62, 140.31)	—
	$\hat{z}_{r,\text{pred}}$		104.39	113.09	119.96	147.34	—

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 knots.

3 Simulation study

Given the results discussed in Section 2.3.4, we believe there is a need to investigate the properties of $z_{r,\text{pred}}$ relative to estimative return levels obtained as point summaries from the posterior distribution for z_r . In particular, of interest to practitioners might be the actual return probabilities associated with $z_{r,\text{pred}}$ and other estimative return levels, and how these compare to r^{-1} .

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 serial correlations 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 AR(1) process with lag 1 autocorrelation A , again covering a range of serial correlations through specific choices for A . Marginally, our reference series are 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 \(2001, Ch. 4\)](#).

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 GP marginal structure as before. For each series $\mathbf{y}^{(\ell)}$ we perform a full MCMC procedure, using objective (and, where available, conjugate) priors. 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

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

and

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

where $\bar{y}$ 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% confidence 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 and independence, but also the effects of model mis-specification on return level inference when asymptotic dependence / independence is incorrectly assumed. For information, we also 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 to analysis.

3.3.1 Asymptotic dependence

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

Figure 3 shows sampling distribution means, and 95% confidence intervals (taken as the sampling distribution 2.5% and 97.5% quantiles), for $p_{\bar{z}_r}$, $p_{\hat{z}_r}$, $p_{z_{r, \text{upper}}}$ and $p_{z_{r, \text{pred}}}$ (for some fixed values of r); Table 2 (top half) reports some numerical summaries. 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, Figure 2 shows

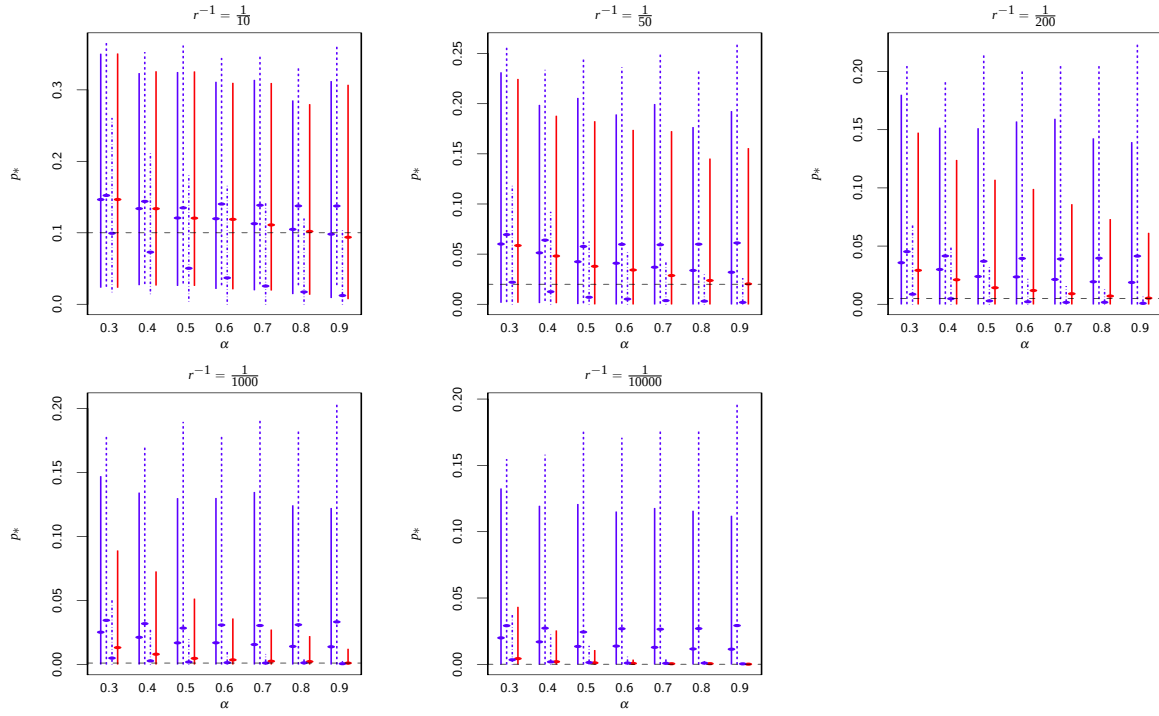


Fig. 3 Sampling distribution means (dots) and 95% confidence intervals (sampling distribution 2.5% and 97.5% quantiles, lines) for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (ii) the posterior predictive return level (red). Here, the simulated data are constructed with asymptotic dependence according to a bivariate logistic model with dependence α . The horizontal dashed line represents the target exceedance probability r^{-1} . Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

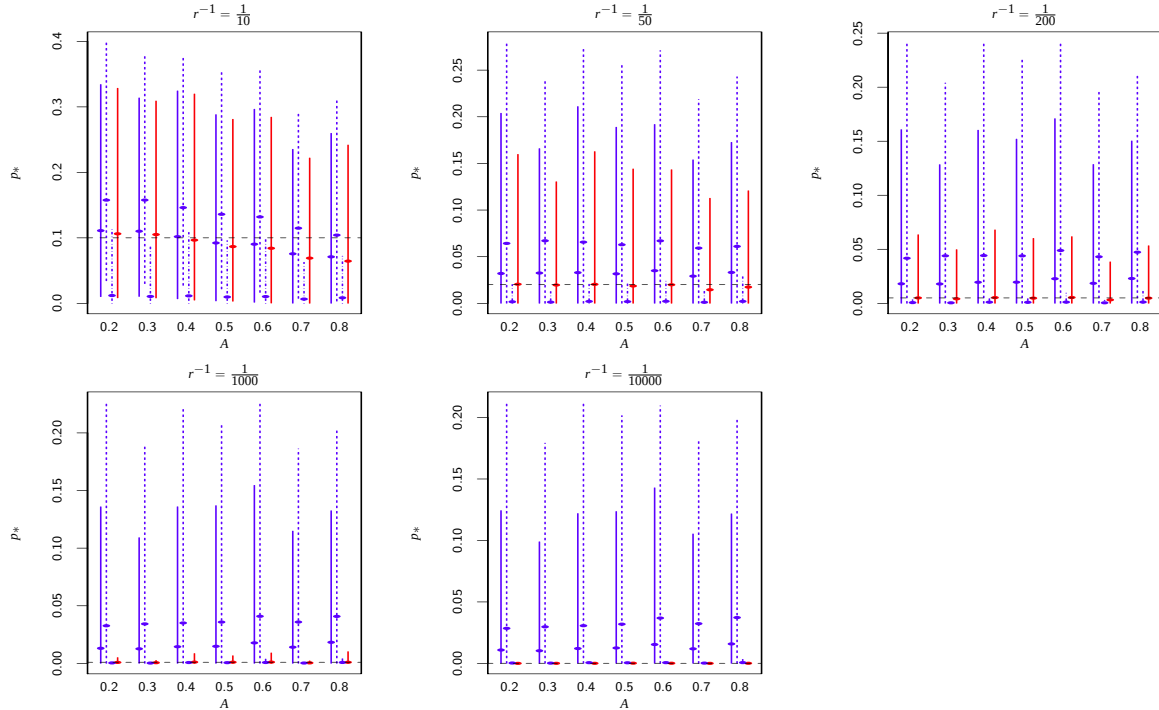


Fig. 4 Sampling distribution means (dots) and 95% confidence intervals (sampling distribution 2.5% and 97.5% quantiles, lines) for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (ii) the posterior predictive return level (red). Here, the simulated data are constructed with asymptotic independence according to an AR(1) process with lag 1 autocorrelation A . The horizontal dashed line represents the target exceedance probability r^{-1} . Marginally, $\xi = -0.4$ and $u = u_{0.95}$.

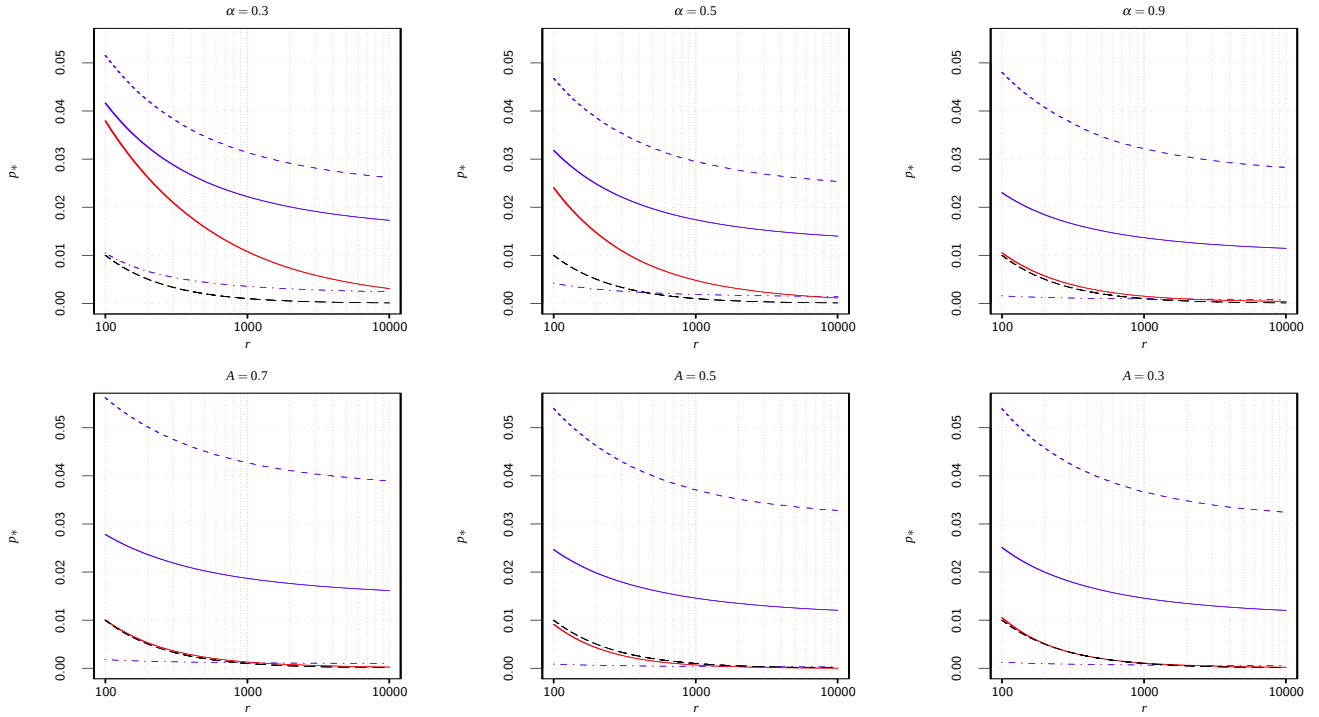


Fig. 5 Sampling distribution means for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (**blue**; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound), and (ii) the posterior 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 $u = u_{0.95}$.

		Asymptotic dependence: logistic model for bivariate extremes with dependence α											
		$\alpha = 0.3$				$\alpha = 0.5$				$\alpha = 0.9$			
		$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$	$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$	$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$
$r^{-1} = \frac{1}{10}$	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
	MSE	0.863	0.998	0.343	0.870	0.746	0.983	0.472	0.744	0.591	0.881	0.877	0.577
$r^{-1} = \frac{1}{50}$	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
	MSE	0.441	0.603	0.081	0.419	0.374	0.613	0.064	0.303	0.293	0.647	0.058	0.180
$r^{-1} = \frac{1}{200}$	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
	MSE	0.267	0.406	0.030	0.180	0.242	0.445	0.022	0.114	0.204	0.506	0.020	0.056
$r^{-1} = \frac{1}{1000}$	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
	MSE	0.181	0.305	0.016	0.061	0.179	0.355	0.014	0.035	0.163	0.421	0.016	0.019
$r^{-1} = \frac{1}{10000}$	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
	MSE	0.138	0.252	0.011	0.013	0.147	0.310	0.011	0.007	0.141	0.378	0.014	0.008

		Asymptotic independence: AR(1) process with lag 1 autocorrelation A											
		$A = 0.7$				$A = 0.5$				$A = 0.3$			
		$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$	$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$	$\bar{p}_{z_r}$	$\bar{p}_{z_r}$	$p_{z_r,upper}$	$p_{z_r,pred}$
$r^{-1} = \frac{1}{10}$	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
	MSE	0.634	0.839	0.904	0.627	0.614	0.944	0.883	0.597	0.747	1.183	0.877	0.723
$r^{-1} = \frac{1}{50}$	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
	MSE	0.328	0.758	0.058	0.164	0.283	0.714	0.043	0.145	0.335	0.815	0.050	0.192
$r^{-1} = \frac{1}{200}$	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
	MSE	0.255	0.651	0.020	0.051	0.185	0.551	0.005	0.028	0.216	0.614	0.011	0.045
$r^{-1} = \frac{1}{1000}$	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
	MSE	0.213	0.573	0.015	0.014	0.139	0.452	0.002	0.003	0.164	0.506	0.006	0.010
$r^{-1} = \frac{1}{10000}$	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
	MSE	0.188	0.525	0.014	0.002	0.114	0.399	0.001	0.000	0.137	0.450	0.005	0.001

Table 2 Estimated biases ($\times 100$), and mean squared errors (MSE, $\times 100$), for each of our four estimators of the return level exceedance probability r^{-1} . The top part of the table summarizes 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** are those which are the smallest in each component of the study.

	Asymptotic dependence: logistic model for bivariate extremes with dependence α											
	$\alpha = 0.3$				$\alpha = 0.5$				$\alpha = 0.9$			
	Dependence α		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$	Dependence α		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$	Dependence α		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$
Bias	3.132	3.127	-0.127	-3.598	1.654	2.269	-0.035	-0.767	0.617	0.825	-0.075	0.602
MSE	0.509	0.401	0.560	2.260	0.223	0.380	0.154	0.666	0.057	0.107	0.091	0.424

	Asymptotic independence: AR(1) process with lag 1 autocorrelation A											
	A = 0.7				A = 0.5				A = 0.3			
	Dependence A		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$	Dependence A		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$	Dependence A		Margins $\bar{\sigma}^*$	Margins $\bar{\xi}$
Bias	-0.097	0.054	0.307	-0.212	-0.231	0.093	0.236	0.167	-0.791	0.199	0.240	0.223
MSE	0.023	0.019	0.102	0.468	0.056	0.019	0.084	0.395	0.182	0.030	0.091	0.409

Table 3 Estimated biases ($\times 100$) and mean squared errors (MSE, $\times 100$), for posterior means of the marginal and dependence parameters when our simulated series exhibit asymptotic dependence according to the logistic model with dependence α (top part of the table); and when our series exhibit asymptotic independence according to an AR(1) process with lag 1 autocorrelation A. Marginally, we have $\xi = -0.4$ and $u = u_{0.95}$.

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 values of the mean squared error (MSE)) 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 MSE 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_{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_{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 credibility of the posterior mode as a useful summary of the return level posterior distribution.

As with Figure 3, the top row of Figure 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 dependence (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_{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 Figure 5. In agreement with the results shown in Figure 3, the plots in Figure 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 Figure 3, and the first plot in Figure 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).

For information, results shown in the top part of Table 3 examine the accuracy and precision of the posterior mean as a summary for the dependence and marginal parameters. Although all values are very small (note that results are reported $\times 100$), we observe an increase in both bias and MSE for the dependence parameters as the dependence strengthens, with a corresponding increase in MSE for the marginal parameters also.

Other arms of the study: main highlights

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 probabilities 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 MSE. 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 $p_{z_{r,\text{pred}}}$ (around five times larger for $r = 10,000$), with consistently narrower 95% confidence intervals; when $\alpha = 0.9$, the outperformance of $z_{r,\text{pred}}$ 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 Figures 3 and 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 observed extremal dependence as given by the dependence parameter α in the (symmetric) 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 $u = u_{0.95}$

Figure 4 shows sampling distribution means, and 95% confidence intervals (taken as the sampling distribution 2.5%

and 97.5% quantiles), for $p_{\bar{z}_r}$, $p_{\bar{z}_r}$, $p_{z_{r,\text{upper}}}$ and $p_{z_{r,\text{pred}}}$ (for some fixed values of r); Table 2 (bottom half) reports some numerical summaries. For these particular results the simulated chains display asymptotic independence according to an AR(1) process with lag 1 autocorrelation A ; as in Section 3.3.1, marginally $\xi = -0.4$ and $u = u_{0.95}$. 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 plots in Figure 4 (certainly when $r > 10$) show sampling distribution means for $p_{z_{r,\text{pred}}}$ closer to the intended exceedance probabilities than those for $p_{\bar{z}_r}$, regardless of the strength of dependence A . 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 (especially as r increases). 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 MSE. 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 Figure 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).

For information, results shown in the bottom half of Table 3 reveal small biases and MSE values for the posterior means for both the marginal and dependence parameters.

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 MSE, 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 $p_{\bar{z}_r}$, $p_{\dot{z}_r}$, $p_{z_{r,\text{upper}}}$ 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 the next Section, we investigate our four estimators of r^{-1} under a standard POT analysis, wherein an attempt is made to completely filter out dependence).

Figure 6 (top row) shows plots similar to those in Figure 5: we have sampling distribution means for $p_{\bar{z}_r}$, $p_{\dot{z}_r}$, $p_{z_{r,\text{upper}}}$ 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 Figure 5, in which the correct form of dependence was assumed, we see a positive 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 MSE 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.

3.3.4 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.

Figure 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 ($\kappa = 5$ or $\kappa = 20$), 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 Figure 5 for a comparison. Similar findings were obtained when our series were $AR(1)$ processes.

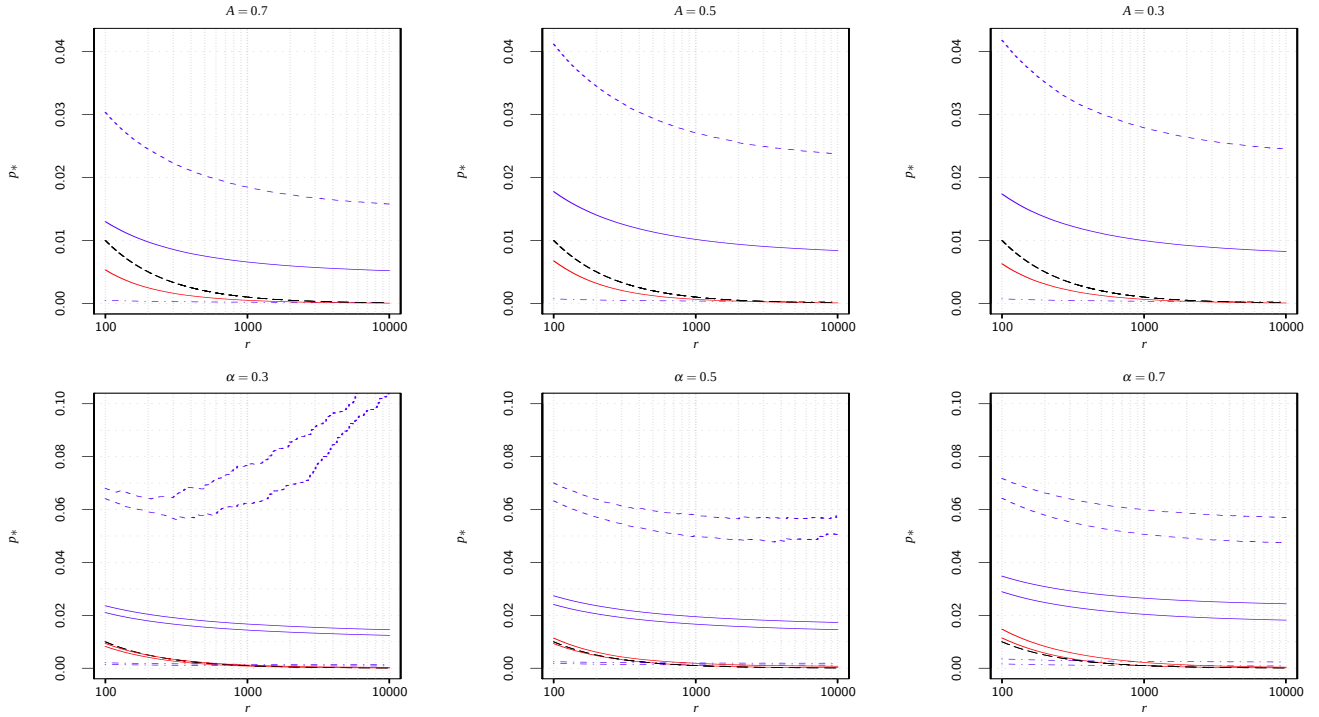


Fig. 6 Sampling distribution means for estimates of the return level exceedance probabilities r^{-1} , using (i) direct summaries from the return level posterior distribution (**blue**; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound), and (ii) the posterior 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 $u = u_{0.95}$.

4 Conclusions

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 estimation uncertainty – and from the viewpoint of prediction, randomness in future observations – 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.

In this paper, 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 high threshold with the GP distribution, properly accounting for serial correlation 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 peicewise seasonal approach to modelling, pressing threshold excesses from all seasons into use; other features, such as trends, can be simply captured through linear modeling 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 poste-

rior 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 modeling framework in [Fawcett and Walshaw \(2016\)](#), in which all excesses are modeled, but also within a more commonly-adopted POT modeling procedure. For a range of return periods r , on a fine scale, and across a range of serial correlations 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 serial correlation 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 modeling 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.

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] Coles, S.G. (2001). *An introduction to statistical modeling of extreme values*. Springer, London.
- [4] 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.
- [5] Coles, S.G. and Tawn, J.A. (1996). A Bayesian Analysis of Extreme Rainfall Data. *J. R. Statist. Soc., C*, **45**, pp. 463–478.
- [6] 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.
- [7] Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Stat. Sci.*, **27**, pp. 161–186.
- [8] 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.
- [9] 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.
- [10] Fawcett, L. and Walshaw, D. (2006). A hierarchical model for extreme wind speeds. *J. Roy. Statist. Soc., C*, **55**, **5**, pp. 631–646.
- [11] Fawcett, L. and Walshaw, D. (2006a). A hierarchical model for extreme wind speeds. *J. Roy. Statist. Soc., C*, **55**, **5**, pp. 631–646.
- [12] Fawcett L. and Walshaw D. (2008). Bayesian inference for clustered extremes. *Extremes*, **11**, pp. 217–233.
- [13] Fawcett, L. and Walshaw, D. (2012). Estimating return levels from serially dependent extremes. *Environmetrics*, **23**, **3**, pp. 272–283.
- [14] 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.
- [15] Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Statist. Soc., B*, **65**, pp. 545–556.
- [16] Leadbetter, M.R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**, pp. 431–476.
- [17] Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. *Annals of Statistics*, **3**, **1**, pp. 119–131.
- [18] 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.
- [19] 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.
- [20] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc., B*, **55**, pp. 3–23.
- [21] Smith, E.L. and Walshaw, D. (2003). Modelling Bivariate Extremes in a Region. *Bayesian Statistics*, **7**, pp. 681–690.
- [22] Smith, R.L. (1992). The Extremal Index for a Markov Chain. *J. Appl. Prob.*, **29**, pp. 37–45.
- [23] 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.
- [24] Smith, R.L., Tawn, J.A. and Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, pp. 249–268.
- [25] 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.

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 serial correlation. 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 Equations (8) or (9).

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

$$\theta_n = - \frac{\log \Pr(\max\{X_1, \dots, X_n\} \leq x_n)}{\log 2}, \quad (13)$$

and so $\theta_n \rightarrow \theta$ as $n \rightarrow \infty$. We can use (13) 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 of serial correlation present in the extremes of the process; then the probability in the numerator of (13) is estimated as the proportion of simulated chains whose maximum does not exceed x_n . The first plot in Figure 7 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 Equation (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 Equation (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 Equation (7) estimating $\theta(u_p)$ (rather than θ itself), a threshold-based penultimate approximation to θ . Figure 8 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 .

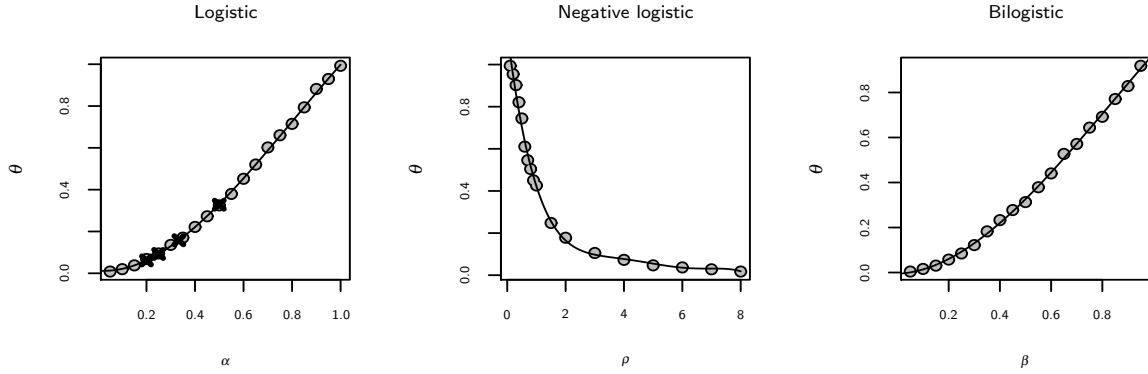


Fig. 7 Simulated values of the extremal index θ for α (logistic), ρ (negative logistic) and β (bilogistic with $\alpha = 0.6$). The solid lines correspond to fitted polynomials: $\theta = 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3$; $\theta = 1.153 - 1.107\rho + 0.463\rho^2 - 0.096\rho^3 + 0.010\rho^4 - 0.0004\rho^5$; $\theta = -0.005 + 0.045\beta + 1.539\beta^2 - 0.607\beta^3$. The crosses in the first plot show limiting values of θ for some values of α in the logistic model, as derived in [Smith \(1992\)](#).

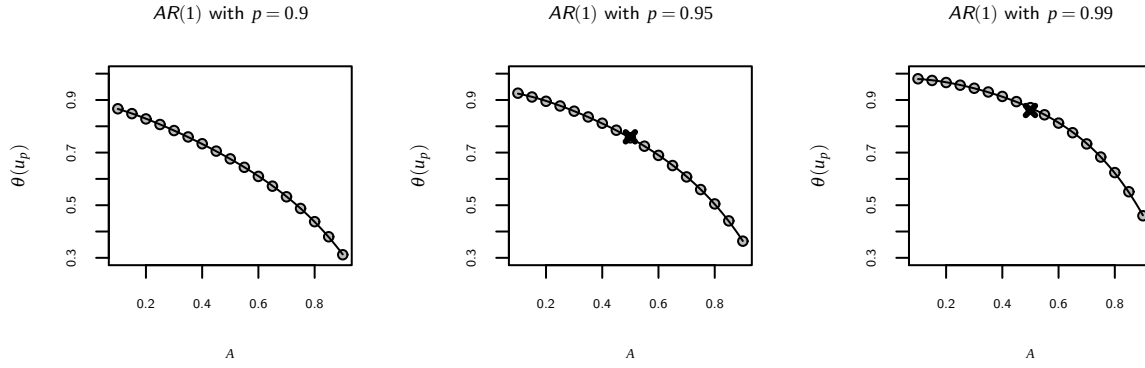


Fig. 8 Simulated values of the threshold-based extremal index $\theta(u_p)$ for an $AR(1)$ process with a threshold set at the 90% marginal quantile $u_{0.9}$ (left); 95% marginal quantile $u_{0.95}$ (middle); 99% marginal quantile $u_{0.99}$ (right). The solid lines correspond to fitted polynomials: $\theta(u_{0.9}) = 0.891 - 0.168A - 0.920A^2 + 1.234A^3 - 0.886A^4$ (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 \(2000\)](#).