Stochastic Environmental Research and Risk Assessment (2018) 32:2233–2252
https://doi.org/10.1007/s00477-018-1561-x

ORIGINAL PAPER

# Bayesian posterior predictive return levels for environmental extremes

Lee Fawcett[1] · Amy C. Green[1]

Published online: 31 May 2018

**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 generalised Pareto (GP) distributions, 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, 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 that are implausible for the variable being studied. The aim of some recent work, then, has been to investigate the use of methods that 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 that 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: 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, seasonal variability and trend, coupled with an increase in dimensionality required of an analysis which is—for example—

✉ Lee Fawcett
lee.fawcett@ncl.ac.uk

[1] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

spatial in flavour, often makes inference within a maximum likelihood setting difficult. Moreover, the complexities that such modelling issues bring are often more naturally handled within a Bayesian framework (e.g. the specification of prior distributions for seasonal effects within a random effects model; see for example, Fawcett and Walshaw 2006a).

The incorporation of external sources of information through the prior distribution is 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, $z_{r,\mathrm{pred}}$, provides practitioners with a point summary capturing estimation uncertainty. It is surprising, then, that there are not more examples of Bayesian inference for extremes in the literature. 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 (2006a), 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 (2006a, 2008, 2016) demonstrate predictive inference for return levels of wind speed and sea-surge extremes, recommending $\hat{z}_{r,\mathrm{pred}}$ as the most convenient, and useful, representation of a return level for practitioners. However, 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 $\hat{z}_{r,\mathrm{pred}}$. In particular, we focus on a comparison of the exceedance probabilities of $\hat{z}_{r,\mathrm{pred}}$ to their intended values $r^{-1}$, and the general performance of $\hat{z}_{r,\mathrm{pred}}$ relative to other estimative summaries obtained directly from the posterior distribution for $z_r$.

This paper is organised as follows. In Sect. 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 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 Sect. 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 Sect. 4.

## 2 Practical motivation and modelling

In this section, we introduce the wind speed data we use throughout the paper. We then give a brief overview of the basic methods for modelling extremes on such processes, including some general background on Bayesian sampling and specific details relating to the posterior predictive return level. Some results are then presented comparing estimative and predictive return levels 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 and December 31st 2012 (inclusive). The boxplots reveal clear seasonal variability in the wind speed extremes, and there is also significant first-order autocorrelation (persisting above monthly-varying high thresholds). Estimates of $z_r$ or $z_{r,\mathrm{pred}}$ 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 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 that incorporates uncertainty due to model estimation through the predictive return level.

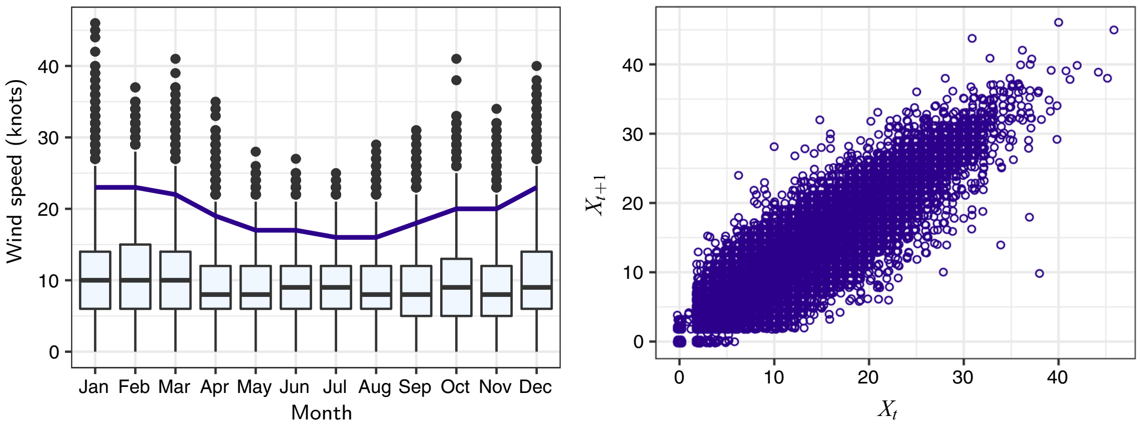

**Fig. 1** Boxplots (by month, left) and each observation plotted against its lag 1 counterpart (right) for a series of hourly gust wind speed maxima observed at Yeovilton between January 1st 2003 and December 31st 2012 (inclusive). The blue lines in the first plot correspond to high thresholds that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

## 2.2 Statistical modelling

### 2.2.1 The basics

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

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

where $\theta \in (0, 1]$ is known as the *extremal index*; e.g. Leadbetter and Rootzén (1988). As $\theta \to 0$ 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 $F^{n\theta}$. 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 d.f.

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

defined on $\{y\colon 1 + \xi(y-\mu)/\varsigma > 0\}$, 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 $\tau$; the calendar year is often used for $\tau$, giving rise to an *annual maxima analysis*.

Pickands (1975) showed that for large $u$ the distribution of $(X-u)|X > u$ is approximately GP with d.f.

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

defined on $\{y\colon y > 0 \text{ and } (1+\xi y/\sigma) > 0\}$, where $\sigma = \varsigma + \xi(u-\mu)$ and $\xi$ 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. 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.

### 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. Specifically, they propose fitting (3) by adopting one of the following strategies:

1. *Parametric modelling of dependence*
   As in Smith et al. (1997) and Fawcett and Walshaw (2006a), where appropriate assume a first-order Markov structure for the temporal evolution of extremes over $u$; that is, assume the following likelihood for $\boldsymbol{\psi}$:

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

   where $\boldsymbol{\psi}$ 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 d.f.:

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

here, independence and complete dependence are attained when $\alpha = 1$ and $\alpha \to 0$ 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:

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

as constructed in Fawcett and Walshaw (2012) and discussed in the Appendix, can then be used—after estimation of $\boldsymbol{\psi} = (\sigma, \xi, \alpha)^{\mathsf{T}}$—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 $\beta$ $(0 < \alpha, \beta < 1)$; 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,

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

where $z^*$ is the upper limit of the support of the marginal distribution, takes the value zero (in the case of asymptotic dependence, $\chi > 0$).[1] 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 that 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 $A$, and threshold $u_p$. The crucial censoring device employed when either one or both of $(x_i, x_{i+1})$ lies sub-threshold (explained above in the context of the models used for asymptotic dependence) is also used in the application of an $AR(1)$ process.

We note here that, within our description of models for asymptotic dependence, formal tests 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. by invoking $d$-variate extreme value models (see for example, Coles and Tawn 1991), or an $AR(d)$ process, to model dependence between $d$ consecutive values in the process).

2. *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 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

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

   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 an easy-to-use estimator.

In the presence of seasonal variability, 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 that occur throughout the year. For example, assuming either (1) or (2) above to capture temporal dependence, estimates of $(\sigma_m, \xi_m, \theta_m), m =$

[1] In practice, the pair $(\chi, \bar{\chi})$ are often considered together, with $\chi$ summarising the strength of extremal dependence when $\bar{\chi} = 1$ (asymptotic dependence) and $\bar{\chi}$ measuring the strength of extremal dependence when $\bar{\chi} < 1$ (and so $\chi = 0$; asymptotic independence). See Coles (2001, Ch. 8) for more details.

$1, \ldots, M$, might be obtained for each season $m$, the analysis perhaps being simplified by assuming a common shape parameter $\xi$ or extremal index $\theta$ across all seasons (where appropriate). Anecdotal evidence in (for example) Walshaw (1994) and Fawcett and Walshaw (2006a) 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.

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. $\sigma = \exp\{\beta_0 + \beta_1 t\}$, $t = 1, 2, \ldots$, where $t$ is an indicator of time. More generally, a dependence on covariates can be induced by writing the parameters in the form $h(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta})$, where $h$ is a specified function, $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{X}$ is a model vector. For applications of General Additive Models (GAMs) to extremes, see (for example) Yee and Stephenson (2007) and Chavez-Demoulin and Davison (2005).

### 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:

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

where $w_r = 1 - \left[1 - (rn_y)^{-1}\right]^{1/\theta}$, $\lambda_u$ is the rate of threshold excess and $n_y$ is the (average) number of observations per year. Under approach (1), as outlined in Sect. 2.2.2, $(\sigma, \xi, \theta)$ 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 (2), $(\sigma, \xi)$ can be replaced with their maximum likelihood or Bayesian estimates and $\theta$ 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 $\theta$ into estimates of $z_r$.

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

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

where $H_m$ is the GP d.f. in season $m$ with parameter set $(\lambda_{u_m}, \sigma_m, \xi_m)$, and $\theta_m/n_m$ 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 (2006b), 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 (1) as outlined in Sect. 2.2.2 for handling dependence of consecutive observations. Given the seasonal variability observed in the wind speed extremes, and our earlier discussion in Sect. 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, 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 $(\sigma_m, \xi_m, \theta_m)$, $m = 1, \ldots, 12$, and hence $z_r$ through Eq. (9). Details on MCMC techniques are now extensively published (e.g. Gamerman and Lopes 2006); Sect. 2.3.2 gives more specific information about the algorithm we employ.

### 2.3.1 Prior specification

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

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

This re-parameterisation gives a scale that is threshold-independent (unlike $\sigma$); 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 $(\eta_m, \xi_m, \alpha_m)^{\mathsf{T}}$ based on observations on wind speed extremes made at a nearby location. Specifically, we use:

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

$m = 1, \ldots, 12$. The components of the mean vector $\boldsymbol{\mu}$ are chosen to closely match our beliefs about what are the most likely values of $(\eta_m, \xi_m)$ based on our study of wind speeds at the nearby location; we specify values for $\mathsf{cov}(\eta_m, \xi_m)$ 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 $\alpha_m$ 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 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, using a simple Metropolis update to give successive draws

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

after thinning to every tenth iteration. 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. No formal MCMC diagnostics are employed, although checks such as the Gelman–Rubin convergence diagnostic, and effective sample size computations (see for example, Gamerman and Lopes 2006), could be employed here.

Equation (6) is used to obtain a sample from the posterior for the extremal index $\theta_m$, 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

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

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

$$\Pr\left\{Y \leq z_{r,\text{pred}} | \boldsymbol{x}\right\} = 1 - r^{-1} \tag{11}$$

for $z_{r,\text{pred}}$ therefore gives an estimate of the $r$-year return level that captures 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 $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(S)}$ as realisations from the stationary distribution $\pi(\boldsymbol{\psi}|\boldsymbol{x})$, we have

$$\Pr\left\{Y \leq z_{r,\text{pred}} | \boldsymbol{x}\right\} \approx \frac{1}{S} \sum_{j=1}^{S} \Pr\left\{Y \leq z_{r,\text{pred}} | \boldsymbol{\psi}^{[j]}\right\}, \tag{12}$$

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 $\boldsymbol{\psi} = (\eta_m, \xi_m, \theta_m)^{\mathsf{T}}$, $m = 1, \ldots, 12$.

### 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 some specific $r$. 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 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 Sect. 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.

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 that encapsulates uncertainty in parameter estimation. 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 $r^{-1}$; 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 $z_{r,\text{pred}}$. Are there any advantages of using $z_{r,\text{pred}}$ as opposed to using some other point summary from

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

| | Return period ($r$ years) | | | | |
|---|---|---|---|---|---|
| | 10 | 50 | 200 | 1000 | 10,000 |
| Yeovilton (*see Fig.* 2) | | | | | |
| $\hat{z}_r$ (knots) | | | | | |
| Post. mean | 51.4 | 56.8 | 61.3 | 66.4 | 73.5 |
| Post. mode | 51.0 | 55.8 | 60.2 | 62.5 | 67.4 |
| 95% CI | (47.9, 56.4) | (51.4, 64.9) | (54.1, 72.7) | (56.9, 82.2) | (60.3,96.8) |
| $\hat{z}_{r,\text{pred}}$ (knots) | | | | | |
| | 54.2 | 57.6 | 63.3 | 70.5 | 82.7 |
| Bradfield (*after* Fawcett and Walshaw 2016) | | | | | |
| $\hat{z}_r$ (knots) | | | | | |
| Post. mean | 96.9 | 103.5 | 112.2 | 128.1 | |
| Post. mode | 94.8 | 99.9 | 108.2 | 123.3 | — |
| 95% CI | (95.0,98.9) | (94.1, 116.1) | (95.2, 125.6) | (117.6, 140.3) | |
| $\hat{z}_{r,\text{pred}}$ (knots) | | | | | |
| | 104.4 | 113.1 | 120.0 | 147.3 | |

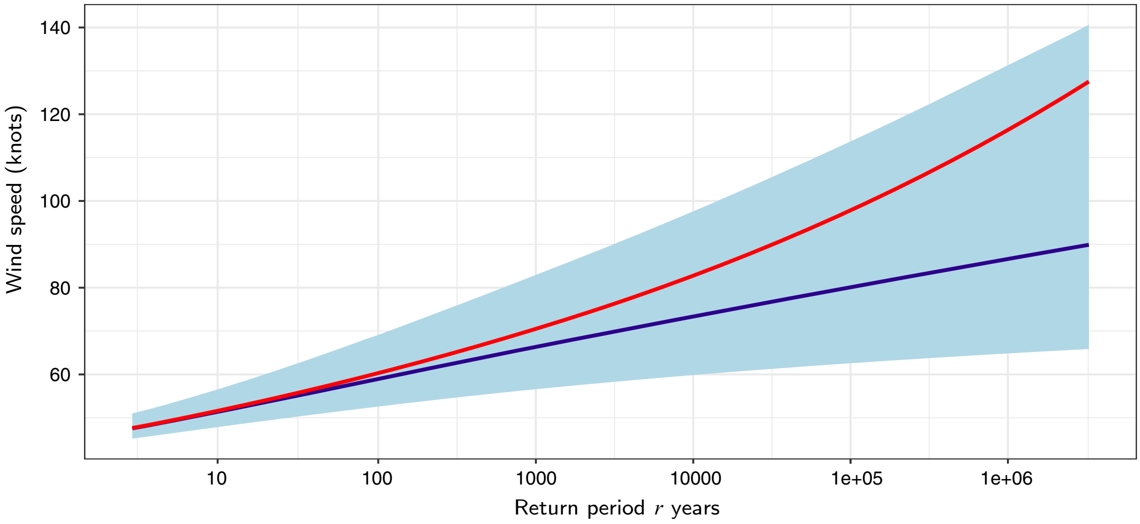

**Fig. 2** Means (blue) taken from the MCMC samples of the posterior distributions for return levels $z_r$ across a range of return periods $r$, with 95% credible intervals (outer light blue shaded area). Also shown, in red, are corresponding estimates of the predictive return levels $z_{r,\text{pred}}$

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 Sect. 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 $AR(1)$ 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 $(\sigma,\xi)$ 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.

Now, at each replication $\ell$ in our simulation study, $\ell = 1, \ldots, L$, 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. 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

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

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 $\theta$ 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 $\alpha$; assuming the prior $\alpha \sim U(0, 1)$, we then perform Metropolis–Hastings sampling, as outlined in Sect. 2.3.2, to obtain draws from the posterior for $\alpha$ using the likelihood in Eq. (4), and hence for the extremal index $\theta$ 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 $(\sigma, \xi)$, before performing a full MCMC procedure on excesses over a range of $u$.

The procedures outlined above yield $S$ iterations after burn-in to obtain approximate samples $\boldsymbol{\sigma}^{*(\ell)}$ and $\boldsymbol{\xi}^{*(\ell)}$, $\ell = 1, \ldots, L$, of length $S$ from the posterior distributions of the GP scale and shape $\sigma^*$ and $\xi^*$, as well as approximate samples from the posteriors of the dependence parameters (i.e. $\alpha$ or $A$) and hence samples $\boldsymbol{\theta}^{(\ell)}$ from the posterior of the extremal index; see Sect. 2.2.2. At each replication $\ell$, via Eq. (8) we also obtain posterior samples $\boldsymbol{z_r}^{(\ell)}$ 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 $\dot{z}_r^{(\ell)}$ and the posterior 95% credible interval upper bound $z_{r,\mathsf{upper}}^{(\ell)}$; we also obtain the predictive return level $z_{r,\mathsf{pred}}^{(\ell)}$ 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 $\mathbf{y}^{(\mathrm{Ref})}$ exceeding each of $\bar{z}_r^{(\ell)}$, $\dot{z}_r^{(\ell)}$, $z_{r,\mathsf{upper}}^{(\ell)}$ and $z_{r,\mathsf{pred}}^{(\ell)}$, $\ell = 1, \ldots, L$, gives sampling distributions for $p_{\bar{z}_r}, p_{\dot{z}_r}, p_{z_{r,\mathsf{upper}}}$ and $p_{z_{r,\mathsf{pred}}}$, 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 $\alpha = \{0.3, \ldots, 0.9\}$ for the logistic model; for the bilogistic model we fix $\alpha$ at 0.5 and use $\beta = \{0.3, \ldots, 0.9\}$; for the $AR(1)$ process we use $A = \{0.2, \ldots, 0.8\}$.

Marginally, we hold $\sigma$ unit constant (i.e. $\sigma = 1$) but consider a range of tail behaviours through the GP shape parameter $\xi$, 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 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 (Sect. 3.3.1) and independence (Sect. 3.3.2), but also the effects of model mis-specification on return level inference when asymptotic dependence/independence is incorrectly assumed (Sect. 3.3.3). We investigate the effects of prior specification on estimates of return level exceedance probabilities (Sect. 3.3.4), relative to those discussed in Sects. 3.3.1 and 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 an *all* threshold excesses, to that obtained under the commonly-used POT approach (Sect. 3.3.5). We also assess the effects of using chains that are drawn marginally from distributions within the domain of attraction of the GP distribution, rather than directly from the GP distribution itself (Sect. 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 (Sect. 3.3.7).

### 3.3.1 Asymptotic dependence

***One arm of the study: logistic dependence structure with $\xi = -0.4$ and $u = u_{0.95}$*** Figure 3 shows sampling distribution means, and 95% confidence intervals, for $\log(1 + rp_*)$. 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. 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

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

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 $p_{z_{r,\mathrm{pred}}}$ that are increasingly more accurate and precise as the return period $r$ increases, especially as the extremal dependence in the series weakens (i.e. as $\alpha \to 1$). 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 (increasingly so as $r$ increases), as is our uncertainty in these estimates. Where both the predictive return level and the return level posterior mean lead to exceedance probabilities that over-estimate $r^{-1}$, there is usually a smaller bias in $p_{z_{r,\mathrm{pred}}}$ 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 $p_* = p_{z_{r,\mathrm{pred}}}$, 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 $z_{r,\mathrm{upper}}$ to over-estimate $z_r$, leading to too-small values for $p_{z_{r,\mathrm{upper}}}$. The return level posterior mode consistently produces estimates $p_{\dot{z}_r}$ that are too large. 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 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 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

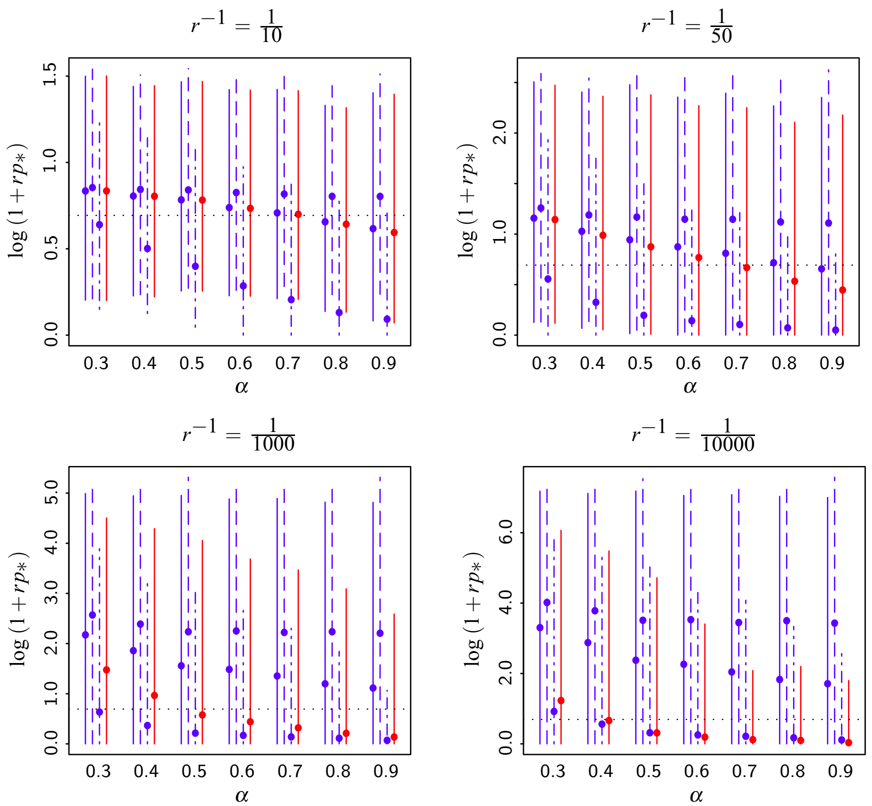

**Fig. 3** Sampling distribution means (bullets) and 95% confidence intervals (vertical lines, running between the sampling distribution 2.5 and 97.5% quantiles) for $\log(1 + rp_*)$, using (1) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (2) the posterior predictive return level (red). Here, the simulated data are constructed with asymptotic dependence according to a bivariate logistic model with dependence $\alpha$. The horizontal dotted line is at log2, representing the maximum of $\mathbb{E}[\log(1 + rp_*)]$. Marginally, $\xi = -0.4$ and $u = u_{0.95}$

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 exceedance probability.

To put these results into a practical context, recall from Sect. 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 $\alpha$ 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 $r^{-1}$. 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 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 *RMSE*/*MLE*. As an example, with $\xi = 0.1$ and $\alpha = 0.3$ or 0.5, biases incurred by $p_{\tilde{z}_r}$ were always larger than those

**Table 2** Estimated biases (×100), root mean squared errors (*RMSE*, ×100) and mean linex errors (*MLE*, ×100), for each of our four estimators of the return level exceedance probability $r^{-1}$

| | | *Asymptotic dependence: logistic model for bivariate extremes with dependence* $\alpha$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\alpha = 0.3$ | | | | $\alpha = 0.5$ | | | | $\alpha = 0.9$ | | | |
| | | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ |
| $r^{-1} = \frac{1}{10}$ | 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* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A = 0.7$ | | | | $A = 0.5$ | | | | $A = 0.3$ | | | |
| | | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ | $p_{\bar{z}_r}$ | $p_{\dot{z}_r}$ | $p_{z_{r,\mathrm{upper}}}$ | $p_{z_{r,\mathrm{pred}}}$ |
| $r^{-1} = \frac{1}{10}$ | 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*** |

The top part of the table summarises results when our simulated series exhibit asymptotic dependence according to the logistic model with dependence parameter $\alpha$; 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

incurred by $p_{z_{r,\mathrm{pred}}}$ (around five times larger for $r = 10,000$), with consistently narrower 95% confidence intervals; when $\alpha = 0.9$, the outperformance of $z_{r,\mathrm{pred}}$ 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 $\xi$ were used), we observed smaller biases than those reported in Figs. 3, 5 and Table 2 (for which $\xi = -0.4$), perhaps indicating that this summary would be more useful for very positively skewed data.

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 $\alpha$ and $\beta$ resulting in similar levels of observed extremal dependence as given by the dependence parameter $\alpha$ 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 $u = u_{0.95}$*** Figure 4 shows the same information as Fig. 3 but now for the case of asymptotic independence where our simulated chains are $AR(1)$ processes with lag 1

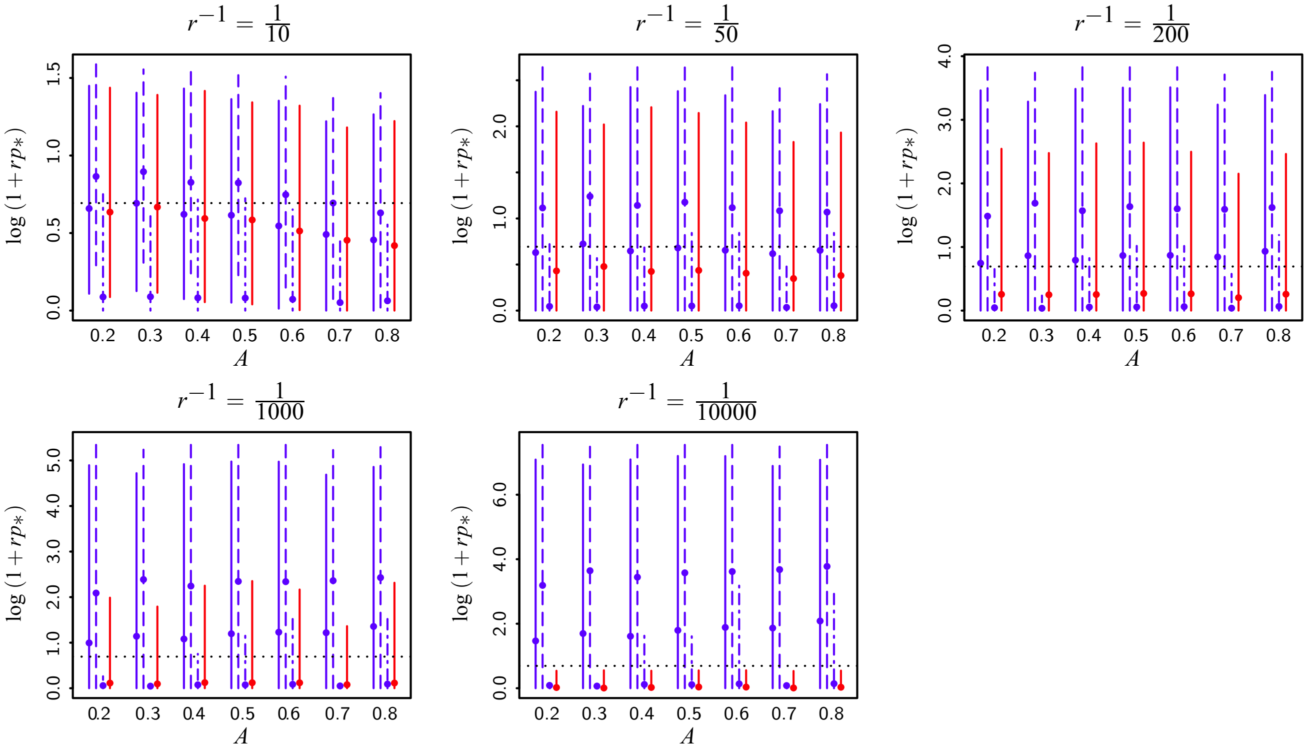

**Fig. 4** Sampling distribution means (bullets) and 95% confidence intervals (vertical lines, running between the sampling distribution 2.5% and 97.5% quantiles) for $\log(1 + rp_*)$, using (1) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound) and (2) 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 log2, representing the maximum of $\mathbb{E}[\log(1 + rp_*)]$. Marginally, $\xi = -0.4$ and $u = u_{0.95}$

autocorrelation $A$; as in Sect. 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 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 Sect. 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 $\xi$ 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. 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 $\ell$, $\ell = 1, \ldots, 1000$, we simulate $\mathbf{y}^\ell$ from an $AR(1)$ process with lag 1 autocorrelation $A$; inference then

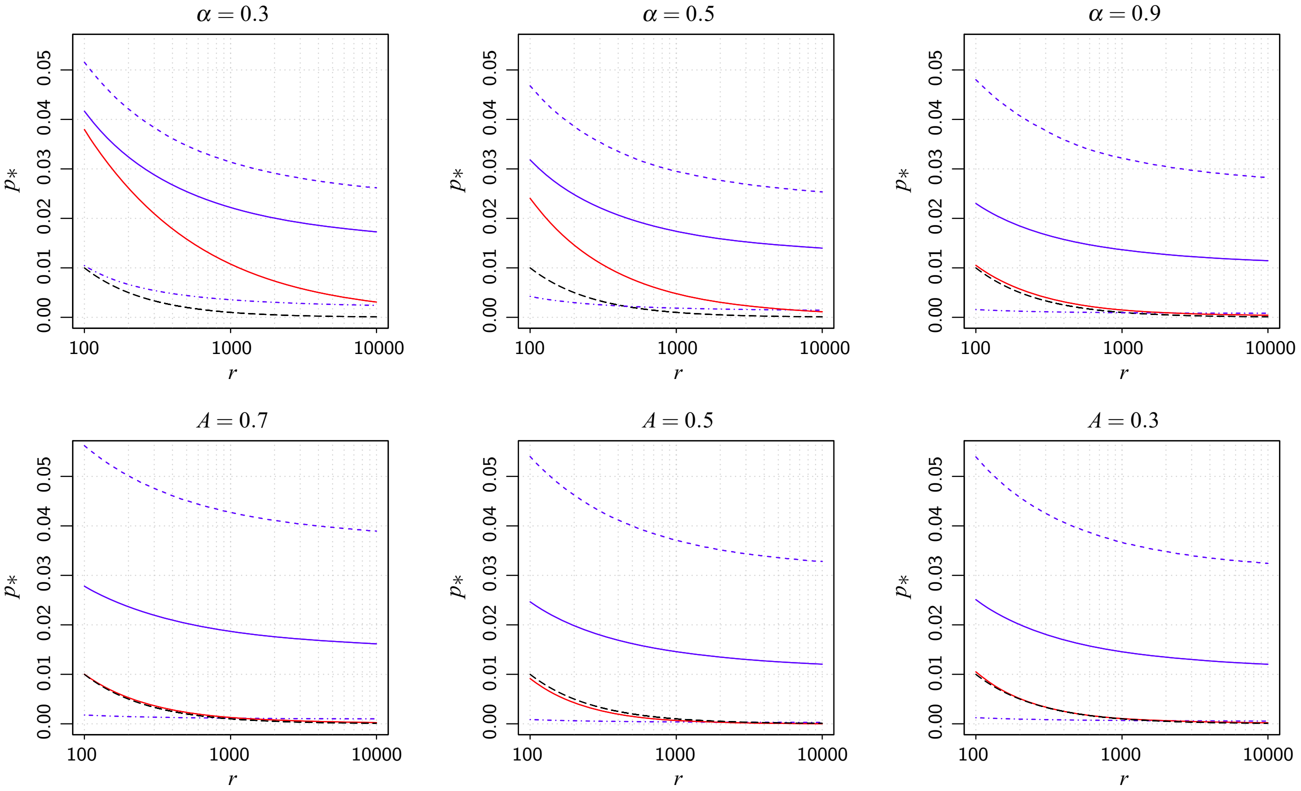

**Fig. 5** Sampling distribution means for estimates of the return level exceedance probabilities $r^{-1}$, using (1) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound), and (2) the posterior predictive return level (red). Simulated data constructed with: asymptotic dependence according to a bivariate logistic model with dependence $\alpha$ (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 $u = u_{0.95}$

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 Sect. 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, 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 that could easily occur in an analysis of real data when we attempt to press all threshold excesses into use.

Figure 6 (top row) shows plots similar to those in Fig. 5. The simulated data exhibit asymptotic independence through an $AR(1)$ structure with lag 1 autocorrelation $A$, but asymptotic *dependence* is incorrectly assumed with dependence structure for consecutive pairs according to the logistic model with parameter $\alpha$. 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.

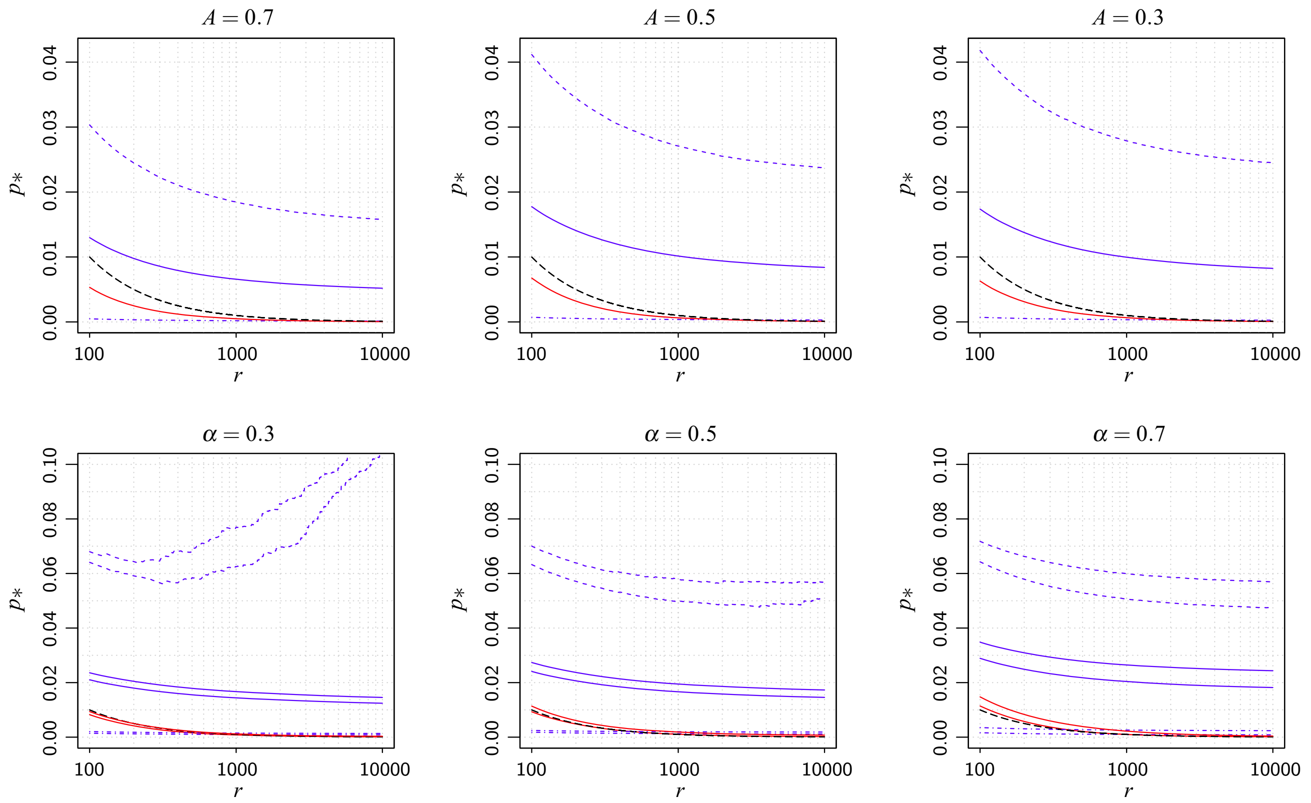

**Fig. 6** Sampling distribution means for estimates of the return level exceedance probabilities $r^{-1}$, using (1) direct summaries from the return level posterior distribution (blue; solid = posterior mean, dashed = posterior mode, dot-dashed = posterior 95% credible upper bound), and (2) the posterior 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 $\alpha$, 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 Sect. 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 $\alpha$. To emulate our approach to prior specification in Sect. 2.3.1, at each replication $\ell$ 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 $(\eta^* = \log(\sigma^* - \xi^* u), \xi^*) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\alpha \sim \text{Beta}(a, b)$, with sensible choices for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $a$ and $b$ based on our analysis of $\mathbf{y}^{\dagger\ell}$ (as opposed to an objective specification using independent $N(0,\, v)$ priors for $\eta^*$ and $\xi^*$ with large $v$, and a $U(0,\, 1)$ prior for $\alpha$, as used in Sect. 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: (1) 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; (2) the bias, *RMSE* and *MLE* ($\times 100$) for $r^{-1} =$

1/10,000 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 $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $a$ and $b$ to be informed by a maximum likelihood analysis of simulated chains $\mathbf{y}_i^{\ddagger\ell}$, $i = 1, \ldots, 5$; unlike $\mathbf{y}^{\dagger\ell}$, these chains have means/variances and temporal dependencies which are (increasingly) dissimilar to those in $\mathbf{y}^{\ell}$. Specifically, we set:

$$\bar{\mathbf{y}}_i^{\ddagger\ell} = (0.5i + 1) \times \bar{\mathbf{y}}^{\ell} \quad \text{and}$$

$$\text{s.d.}\,(\mathbf{y}_i^{\ddagger\ell}) = (-0.2i + 1.1) \times \text{s.d.}\,(\mathbf{y}^{\ell})$$

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

$$\alpha_{\mathbf{y}_i^{\ddagger\ell}} = 0.1i + 0.4 \quad \text{or}$$

$$A_{\mathbf{y}_i^{\ddagger\ell}} = -0.1i + 0.8$$

respectively, meaning that the series $\mathbf{y}_5^{\ddagger\ell}$ is the most dissimilar to $\mathbf{y}^{\ell}$ (hence leading to the most ill-informed prior specification). Informative priors based on $\mathbf{y}^{\ddagger\ell}$ resulted in some increases in the estimated bias, *RMSE* and *MLE* of our return level exceedance probabilities, relative to those based on $\mathbf{y}^{\dagger\ell}$, especially when using the most dissimilar series on which to base our prior specifications ($\mathbf{y}_4^{\ddagger\ell}$ and $\mathbf{y}_5^{\ddagger\ell}$). Here, biases in estimates of $r^{-1}$ were notably larger than those using the objective priors or the informative priors based on $\mathbf{y}^{\dagger\ell}$ (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 $\mathbf{y}^{\dagger\ell}$ (again, least so for estimates based on the predictive return level, especially for return periods $r = 1000$ and $r = 10{,}000$). Informative priors based on the least dissimilar series ($\mathbf{y}_1^{\ddagger\ell}$ and $\mathbf{y}_2^{\ddagger\ell}$) yielded very similar results to those based on $\mathbf{y}^{\dagger\ell}$.

### 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 $\kappa$ 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 $\kappa$. Thus, we do not recommend a POT analysis at all, and we favour an approach as detailed in Sect. 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 $\ell = 1, \ldots, 1000$ using $\kappa = 5$ and $\kappa = 20$. As before, in separate arms of the study we simulate series exhibiting asymptotic dependence and asymptotic independence. However, since the aim is to eliminate dependence between extremes, 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 Sects. 3.3.1 and 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 $\kappa$. 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 overestimated exceedance probabilities. When declustering, all posterior summaries 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 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, 4 and 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 Sects. 3.3.1, 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 that 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, 4, 5 and 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 Sect. 3.3, similar findings were obtained across most other parameters in our study design. However, comparisons in some arms of the simulation study were clearly being influenced by the marginal shape parameter $\xi$. As might be expected, for much heavier-tailed margins 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 and the return level 95% credible upper bound. For these arms of the study, estimates of $r^{-1}$ based on the posterior mode ($p_{\dot{z}_r}$) out-performed the other estimative summaries, although estimates based on the predictive return level seemed to be generally less biased and with smallest error. Generally, as the value of $\xi$ increases the performance of $p_{\bar{z}_r}$ and $p_{z_{r,\mathrm{upper}}}$ deteriorate in terms of estimated bias, *RMSE* and *MLE*, but $p_{z_{r,\mathrm{pred}}}$ retains the accuracy and precision observed in Table 2 and Figs. 3, 4, 5 and 6. Comparisons between our estimators do not appear to be sensitive to the scale of the underlying GP distribution. Although the GP marginal scale $\sigma$ is held unit constant, as discussed in Sect. 3.2 excesses over $u$ have a threshold- and shape-dependent scale $\sigma^*$. For arms of the study in which $\xi$ was constant but $\sigma^*$ varied, we did not see any real departure from the general findings reported in Table 2 and Figs. 3, 4, 5 and 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 that 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 has 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 that 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. 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 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 Sect. 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/mis-specification, 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 into a single point estimate, perhaps to be used to aid structural design. Although the results of our simulation study in Sect. 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 $\delta(\boldsymbol{x})$ which minimises the *a posteriori expected loss* for some model parameter $\psi$:

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

In Eq. (13), $L$ represent a loss function: $L(\psi, \delta(\boldsymbol{x})) = (\psi - \delta(\boldsymbol{x}))^2$ gives squared errors, although as we discuss in Sect. 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 $(\boldsymbol{x})$ and averaging over the unknowns (e.g. parameter(s) $\psi$ and future observations $\boldsymbol{y}$), gives

$$\int_{Y} L(\boldsymbol{y}, \delta(\boldsymbol{x}))f_Y(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}, \tag{14}$$

where $f_Y(\boldsymbol{y}|\boldsymbol{x})$ is the posterior predictive density for $\boldsymbol{y}$. The optimal decision $\delta(\boldsymbol{x})$ can then be seen as the action that minimises this *predictive a posteriori expected loss*. From an inference point-of-view, $\delta(\boldsymbol{x})$ is a function whose output $\hat{\psi}$ is an estimate of $\psi$.

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 Eqs. (13) and (14).

The contribution of parameter uncertainty to the predictive return level can be estimated by comparing $\hat{z}_{r,\mathrm{pred}}$ to what we call naïve return level estimates. Figure 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 the solid and dashed lines in the plots in Fig. 7 can be seen as the average contribution to $z_{r,\mathrm{pred}}$ 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

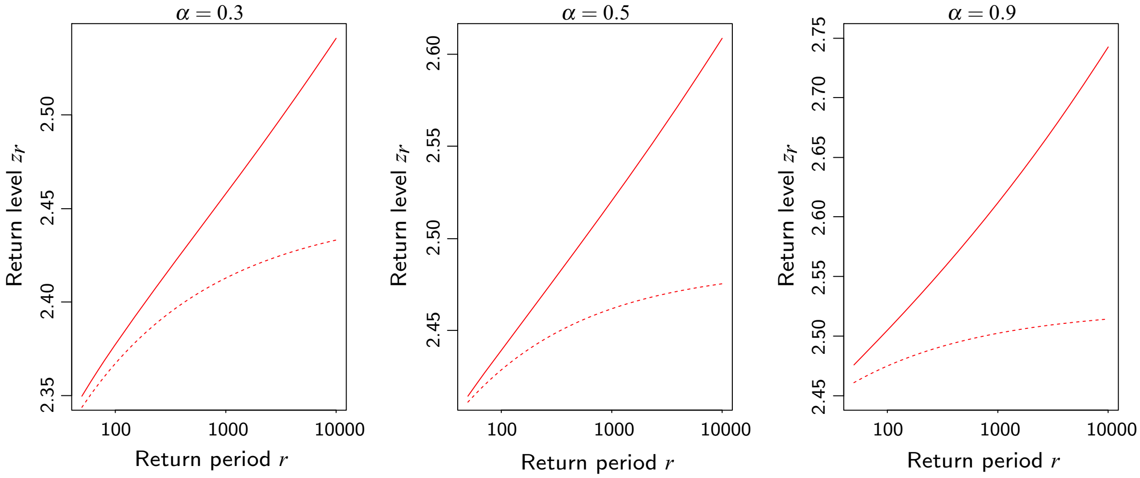

**Fig. 7** Sampling distribution means for predictive return levels $z_{r,\text{pred}}$ (solid lines) and "naïve" return levels (dashed lines). Simulated data are constructed with asymptotic dependence according to the bivariate logistic model with dependence $\alpha$. Marginally, $\xi = -0.4$ and $u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

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 the UK Meteorological Office for supplying the wind speed data that are used in this paper. Data supporting this publication are openly available under an 'Open Data Commons Open Database License'. Additional metadata are available at: http://dx.doi.org/10.17634/154300-87. Please contact Newcastle Research Data Service at rdm@ncl.ac.uk for access instructions. We would also like to thank two referees and an Associate Editor for making excellent suggestions, resulting in a significantly improved manuscript.

## Appendix

Here, we present some results from a simulation study to support the discussion in part (1) of Sect. 2.2.2 and the main simulation study in Sect. 3. The aim is to establish a simple polynomial approximation to the extremal index $\theta$ 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 $\alpha$ (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 Eqs. (8) or (9).

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

$$\theta_n = -\frac{\log \Pr(\max\{X_1, \ldots, X_n\} \le x_n)}{\log 2}, \tag{15}$$

and so $\theta_n \to \theta$ as $n \to \infty$. We can use Eq. (15) to investigate the relationship between $\theta$ and $\alpha$ in the logistic model. Specifically, we simulate $N$ first-order Markov chains, each of length $n$, with logistic dependence $\alpha$ governing the strength 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 $\alpha = \{0.05, 0.10, \ldots, 1.00\}$, using $N = n = 10{,}000$; the other two plots show corresponding results when using the bilogistic model (see Sect. 2.2.2) and another model occasionally used for bivariate extremes [the negative logistic model with dependence parameter $\rho$; 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 Sect. 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 Sect. 3, with lag 1 autocorrelation $A$. To establish a polynomial

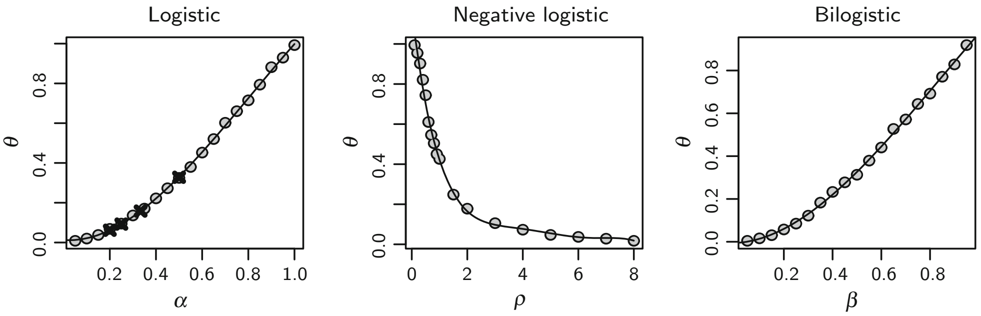

**Fig. 8** Simulated values of the extremal index $\theta$ for $\alpha$ (logistic), $\rho$ (negative logistic) and $\beta$ (bilogistic with $\alpha = 0.6$). The solid lines correspond to fitted polynomials: $\theta = 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3$; $\theta = 1.153 - 1.107\rho +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

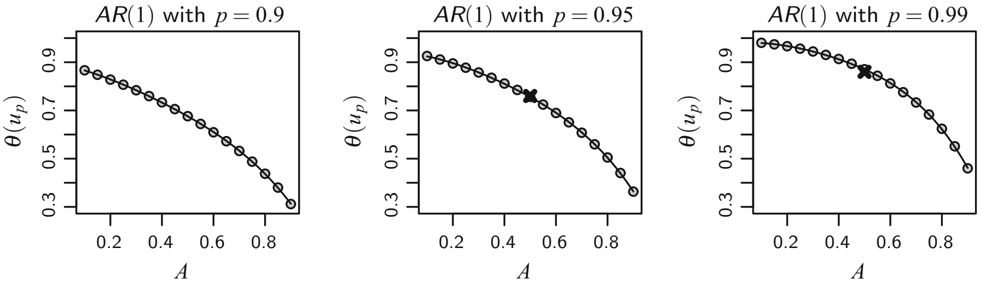

**Fig. 9** Simulated values of the threshold-based extremal index $\theta(u_p)$ for an $AR(1)$ process with a threshold set at the 90% marginal quantile $u_{0.9}$ (left); 95% marginal quantile $u_{0.95}$ (middle); 99% marginal quantile $u_{0.99}$ (right). The solid lines correspond to fitted polynomials: $\theta(u_{0.9}) = 0.891 - 0.168A - 0.920A^2 + 1.234A^3 - 0.886A^4$ (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)

approximation for the extremal index in this setting, for each value $A = \{0.05, 0.10, \ldots, 0.95\}$ 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 (2) of Sect. 2.2.2 of this paper), our approximation to the extremal index $\theta$ 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 $\theta$ itself), a threshold-based penultimate approximation to $\theta$. Figure 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 $A$.

## References

Ancona-Navarrete MA, Tawn JA (2000) A comparison of methods for estimating the Extremal Index. Extremes 3:5–38

Beirlant J, Goegebeur J, Teugels J, Segers J, De Waal D, Ferro C (2004) Statistics of extremes. Wiley, New York

Berger JO (2010) Statistical decision theory and Bayesian analysis. Springer, London

Chavez-Demoulin V, Davison AC (2005) Generalized additive models for sample extremes. J R Stat Soc C 54(1):207–222

Coles SG (2001) An introduction to statistical modeling of extreme values. Springer, London

Coles SG, Powell EA (1996) Bayesian methods in extreme value modelling: a review and new developments. Int Stat Rev 64(1):119–136

Coles SG, Tawn JA (1991) Modelling extreme multivariate events. Biometrika 53(2):377–392

Coles SG, Tawn JA (1996) A Bayesian analysis of extreme rainfall data. J R Stat Soc C 45:463–478

Davison AC, Smith RL (1990) Models for exceedances over high thresholds. J R Stat Soc B 52:393–442 **(with discussion)**

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

Eastoe EF, Tawn JA (2012) Modelling the distribution for the cluster maxima of exceedances of sub-asymptotic thresholds. Biometrika 99(1):43–55

Eugenia Castellanos M, Cabras S (2007) A default Bayesian procedure for the generalized Pareto distribution. J Stat Plan Inf 137(2):473–483

Fawcett L, Walshaw D (2006) A hierarchical model for extreme wind speeds. J R Stat Soc C 55(5):631–646

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

Fawcett L, Walshaw D (2008) Bayesian inference for clustered extremes. Extremes 11:217–233

Fawcett L, Walshaw D (2012) Estimating return levels from serially dependent extremes. Environmetrics 23(3):272–283

Fawcett L, Walshaw D (2016) Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers. Stoch Environ Res Risk Assess 30:463–480

Ferro CAT, Segers J (2003) Inference for clusters of extreme values. J R Stat Soc B 65:545–556

Gamerman D, Lopes HF (2006) Markov Chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall, Boca Raton

Jenkinson AF (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quart J Roy Met Soc 81:158–171

Leadbetter MR, Rootzén H (1988) Extremal theory for stochastic processes. Ann Probab 16:431–476

Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3(1):119–131

Sang H, Gelfand AE (2009) Hierarchical modeling for extreme values observed over space and time. Environ Ecol Stat 16:407–426

Sang H, Gelfand AE (2010) Continuous spatial process models for extreme values. J Agric Biol Environ Stat 15:49–65

Smith RL (1992) The Extremal Index for a Markov Chain. J Appl Probab 29:37–45

Smith RL (1999) Bayesian and frequentist approaches to parametric predictive inference (with discussion). Bayesian Stat 6:589–612

Smith EL, Walshaw D (2003) Modelling bivariate extremes in a region. Bayesian Stat 7:681–690

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

Walshaw D (1994) Getting the most from your extreme wind data: a step by step guide. J Res Natl Inst Stand Technol 99:399–411

Yee TW, Stephenson AG (2007) Vector generalized linear and additive extreme value models. Extremes 10:1–19

Zellner A (1986) Bayesian estimation and prediction using asymmetric loss functions. J Am Stat Assoc 81:446–451