

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

We thank the reviewers and the Editor for taking the time to read our paper. We feel that the minor revisions they have suggested have greatly enhanced the paper and make for a much more complete manuscript. Although the recommendation was for minor revisions to the original submission, we have taken much time to carefully review each and every query raised by the reviewers, and we thank the journal for their patience here (and for agreeing to extend the original re-submission date).

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

## Reviewer 1's report

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

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

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

Secondly, taking into account the consequences of error. We think the reviewer makes an excellent point here, and our response to this query overlaps slightly with our response to query (c) below. As the reviewer suggests, we think the best way to do this, formally, would be through the incorporation of a loss function. We now include a specific discussion in the paper on our thoughts here (please see the first paragraph of Section 4.2), and we hope the reviewer agrees this makes the manuscript much more complete. We do not directly implement any of the material discussed here as we feel this would be too much work for the current paper, although (as we discuss in Section 4.2) our intention is certainly to follow this up in future work. As we discuss, the choice a specific loss function might need careful consideration – in response to query (3) below we now examine “linex errors” for our four estimators, to enable over-estimation of  $r^{-1}$  to Pm punished more severely than under-estimation (as opposed to squared errors which punish over- and -under-estimation symmetrically), and we give a practical rationale for doing so. Thus, linex errors might be incorporated more formally into a loss function for finding an estimator which minimises the Bayesian *a posteriori expected loss* (see Eq. (13) of the revised manuscript), or perhaps the *predictive a posteriori expected loss* (see Eq. (14)) as an equivalent in the predictive setting. These ideas will form part of the second author’s work in her new Ph.D position at our institution.

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

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

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

could be considered ill-informed) and although we noticed increased biases in estimates of  $r^{-1}$ , and a loss of precision, across the board, estimates based on the posterior predictive return level were always superior. Again, full details can now be found in Section 3.3.4 of the paper.

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

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

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

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

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

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

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

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

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

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

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

*ll22-23 and later: the crucial censoring device makes it possible to apply a model built for consecutive observations when data are from exceedances that occur only intermittently. It is described here in the context of the Markov Chain model for asymptotic dependence but presumably is used also for the AR model described later. If true, it seems important to make that clear (or if not, to say how else the problem is solved).*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The reviewer makes very good suggestions here. In our simulations we did not account for non-stationarity (e.g. trend or seasonal variation). Given the length of the manuscript already, and the computation time required to run further simulations, we have also refrained from considering non-stationarity in the revised version of the paper. Of course, we could have generated data from the seasonal piecewise model we apply to our wind speed extremes in

§2 of the paper; however, given the reviewer’s comments (and our response) in (8) above, the method chosen for handling such variability in practice is highly dependent on the variable being studied (and the location at which it was measured), perhaps making it difficult to generalise any findings from the simulation study. Trends could have been built into our generated sequences, perhaps by allowing the GP scale parameter  $\sigma^*$  to be time-dependent. However, as already discussed the computation time and manuscript space necessary to cover this were, we felt, prohibitive.

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

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

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

## Reviewer 2's report

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

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

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

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

We completely agree with everything the reviewer says here, and we have made the necessary changes. Our plots in Figs. 3 and 4 now show our estimates on a logarithmic scale, to make the plots legible for all values of return period  $r$ , and we have presented the ratio of  $p^*$  to  $1/r$  as the reviewer suggests. We have also made sure the size of the plots is determined at the generation stage in R, as opposed to re-sizing within L<sup>A</sup>T<sub>E</sub>X, to ensure maximum clarity. Given the information being displayed in the plots, some readers may need to give the plots careful attention if they intend to understand fully what they reveal, but with the reviewer's recommendations (and with careful treatment being given in the production process of the plots) we believe the plots in the revised manuscript are a big improvement over those displayed in the original submission. In Table 2, results are now shown for *RMSE* (as opposed to *MSE*) for the reason the reviewer suggests. We also give results for the *mean linex error*, in response to a comment made by reviewer 1 (to penalise over-estimation of  $r-1$  more heavily than under-estimation; see §3.3.1 for our rationale here).

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

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

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

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