2 Dependence and non–stationarity

The asymptotic results introduced in Part 1 have assumed the underlying process to be independent and identically distributed (i.i.d.). They also assume this process is stationary. In practice, extreme value data – particularly environmental time series – exhibit some form of departure from this ideal. The most common forms are:

— Local temporal dependence, where successive values of the time series are dependent, but values farther apart are independent (to a good approximation);

— Long term trends, where the underlying distribution changes gradually over time;

— Seasonal variation, where the underlying distribution changes periodically through time.

These departures can be handled through a combination of extending both the theory and the modelling. However, although a wide range of theoretical models for non–stationarity have been studied, only in a few cases have these been used for statistical modelling; the results have generally been too specific to be of use in modelling data for which the form of non–stationarity is unknown. Over the last decade or so, it has been more usual for practitioners to employ statistical procedures which allow the existing results to be applied. In Part 2, we will consider some of these in detail.

2.1 Extremes of dependent sequences

For the types of data to which extreme value models are commonly applied, temporal independence is usually an unrealistic assumption. In particular, extreme conditions often persist over several consecutive observations, bringing into question the appropriateness of models such as the GEV. A detailed investigation of this requires mathematical treatment at a level of sophistication beyond which we have time to capitulate in this short course; however, the general ideas are not difficult and the main result offers a simple, practical, interpretation. For the remainder of this section on dependent sequences, we shall assume that our process is *stationary*, corresponding to a series whose variables may be mutually dependent, but whose stochastic properties are homogeneous throughout time.

Dependence in stationary sequences can take many different forms. With practical applications in mind, it is common to assume a condition that limits the extent of dependence to short–range temporal dependence so that, for example, events X_i and X_j, both of which are extreme, are independent provided time points i and j are far enough apart. Indeed, many stationary sequences satisfy this property. By excluding the possibility of long–range dependence in this way, we focus our attention on dependence at a much shorter range. Effects of such short–range dependence, it turns out, can be quantified within the standard extreme value limits discussed in Part 1.

2.1.1 Maxima of stationary sequences

The book by Leadbetter *et al.* (1983) considers, in great detail, properties of extremes of dependent processes. A key result often used is 'Leadbetter's $D(u_n)$ condition', which ensures that long–range dependence is sufficiently weak so as not to affect the asymptotics of an extreme value analysis. This condition is stated more formally in the Definition below.

Definition (Leadbetter's $D(u_n)$ condition)
A stationary series $X_1, X_2, \ldots$ is said to satisfy the $D(u_n)$ condition if, for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l$,

$$\begin{aligned}\Big| \Pr\left\{X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n, X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n\right\} \\ -\Pr\left\{X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n\right\} \Pr\left\{X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n\right\} \Big| \quad &\le \quad \alpha(n, l), \end{aligned} \tag{7}$$

where $\alpha(n, l) \to 0$ for some sequence l_n such that $l_n/n \to 0$ as $n \to \infty$.

For sequences of independent variables, the difference in probabilities in the above expression is exactly zero for *any* sequence u_n. More generally, we will require that the $D(u_n)$ condition holds only for a specific sequence of thresholds u_n that increases with n. For such a sequence, the $D(u_n)$ condition ensures that, for sets of variables that are far enough apart, the difference in probabilities expressed in (7), while not zero, is sufficiently close to zero to have no effect on the limit laws for extremes.

Theorem
Let $\tilde{X}_1, \tilde{X}_2, \ldots$ be a stationary series satisfying Leadbetter's $D(u_n)$ condition, and let $\tilde{M}_n = \max\{\tilde{X}_1, \ldots, \tilde{X}_n\}$. Now let $X_1, X_2, \ldots$ be an *independent* series with X having the same distribution as $\tilde{X}$, and let $M_n = \max\{X_1, \ldots, X_n\}$. Then if M_n has a non–degenerate limit law given by $\Pr\{(M_n - b_n)/a_n \le x\} \to G(x)$, it follows that

$$\Pr\left\{(\tilde{M}_n - b_n)/a_n \le x\right\} \to G^\theta(x) \tag{8}$$

for some $0 \le \theta \le 1$.

The parameter θ is known as the *extremal index*, and quantifies the extent of extremal dependence: $\theta = 1$ for a completely independent process, and $\theta \to 0$ with increasing levels of (extremal) dependence. Since G in the above theorem is necessarily an extreme value distribution, and due to the *max–stability* property (see Leadbetter *et al.*, 1983), then the distribution of maxima in processes displaying short–range temporal dependence (characterised by the extremal index θ) is also a GEV distribution; the powering of the limit distribution by θ only affects the location and scale parameters of this distribution.

The above theorem implies that if maxima of a stationary series converge – which, from Part 1, we know they will do – then, provided an appropriate $D(u_n)$ condition is satisfied, the limit distribution is related to the limit distribution of an independent series. The effect of dependence, as seen in expression (8), is just a replacement of G as the limit distribution with G^θ. In fact, if G corresponds to the GEV distribution with parameters (μ, σ, ξ), then

$$\begin{aligned} G^\theta(z) &= \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}^\theta \\ &= \exp\left\{-\left[1 + \xi\left(\frac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\}, \end{aligned}$$

where $\mu^* = \mu - \frac{\sigma}{\xi}\left(1 - \theta^{-\xi}\right)$ and $\sigma^* = \sigma\theta^\xi$. Thus, if the (approximate) distribution of M_n is GEV with parameters (μ, σ, ξ), then the (approximate) distribution of $\tilde{M}_n$ is GEV with parameters (μ^*, σ^*, ξ).

2.1.2 Modelling block maxima

Provided long–range dependence is weak, we can proceed to model block maxima from series with short–range extremal dependence as outlined in Part 1, since the distribution of block maxima falls within the same family of distributions as would be appropriate if the series were truly independent. This is fantastic news! Short–range temporal dependence is a much more plausible assumption than complete independence, and our modelling approach is still valid! However, the main difference – excluding the change in parameters from (μ, σ, ξ) to (μ^*, σ^*, ξ) – is that our implied n (the number we are taking the maxima over) is now effectively reduced due to the dependence, so convergence of maxima to the limit distribution will be slower. And shouldn't we be using threshold methods anyway, which use information on *all* extremes and not just those that are the maximum within their block?

2.1.3 Modelling threshold exceedances

Though the modelling procedure for fitting the GEV to a set of annual maxima is unchanged for series which display short–term temporal dependence, some revision is needed of the threshold exceedance approach. If all threshold exceedances are used in our analysis, and the GPD fitted to the set of threshold excesses, the likelihoods we use will be incorrect since they assume independence of sample observations. In practice, several techniques have been developed to circumvent this problem, including:

1. filtering out an (approximately) independent set of threshold exceedances

2. fitting the GPD to *all* exceedances, ignoring dependence, but then appropriately adjusting the inference to take into account the reduction in information

3. Explicitly modelling the temporal dependence in the process

Though the first approach above is by far the most widely–used, our research has focussed on the relative merits of the other two approaches. The third approach makes use of multivariate extreme value theory, and so we shall re–visit this idea in more detail in Parts 4 and 5 this afternoon. For now, let us consider the first two approaches, which we will call *removing* dependence and *ignoring* dependence, respectively.

2.1.4 Example: Cluster peaks or all excesses?

Figure 9 shows a series of 3–hourly measurements of sea–surge heights at Newlyn, a coastal town in the southwest of England, collected over a three year period. The sea–surge is the meteorologically induced non–tidal component of the still–water level of the sea. The practical motivation for the study of such data is that structural failure — probably a sea–wall in this case — is likely under the condition of extreme surges. Also shown in Figure 9 is a plot of the time series against the lag 1 time series.

A natural way of modelling extremes such time series is to use the Generalised Pareto Distribution (GPD) as a model for excesses over a high threshold. As already discussed in Part 1, this approach might be preferable to the block maxima approach which is highly wasteful of data (and precious extremes!). Figure 9 also shows the presence of substantial temporal dependence in the sequence of three–hourly surges. We will now consider approaches **1** and **2**, outlined above, to circumvent this problem.

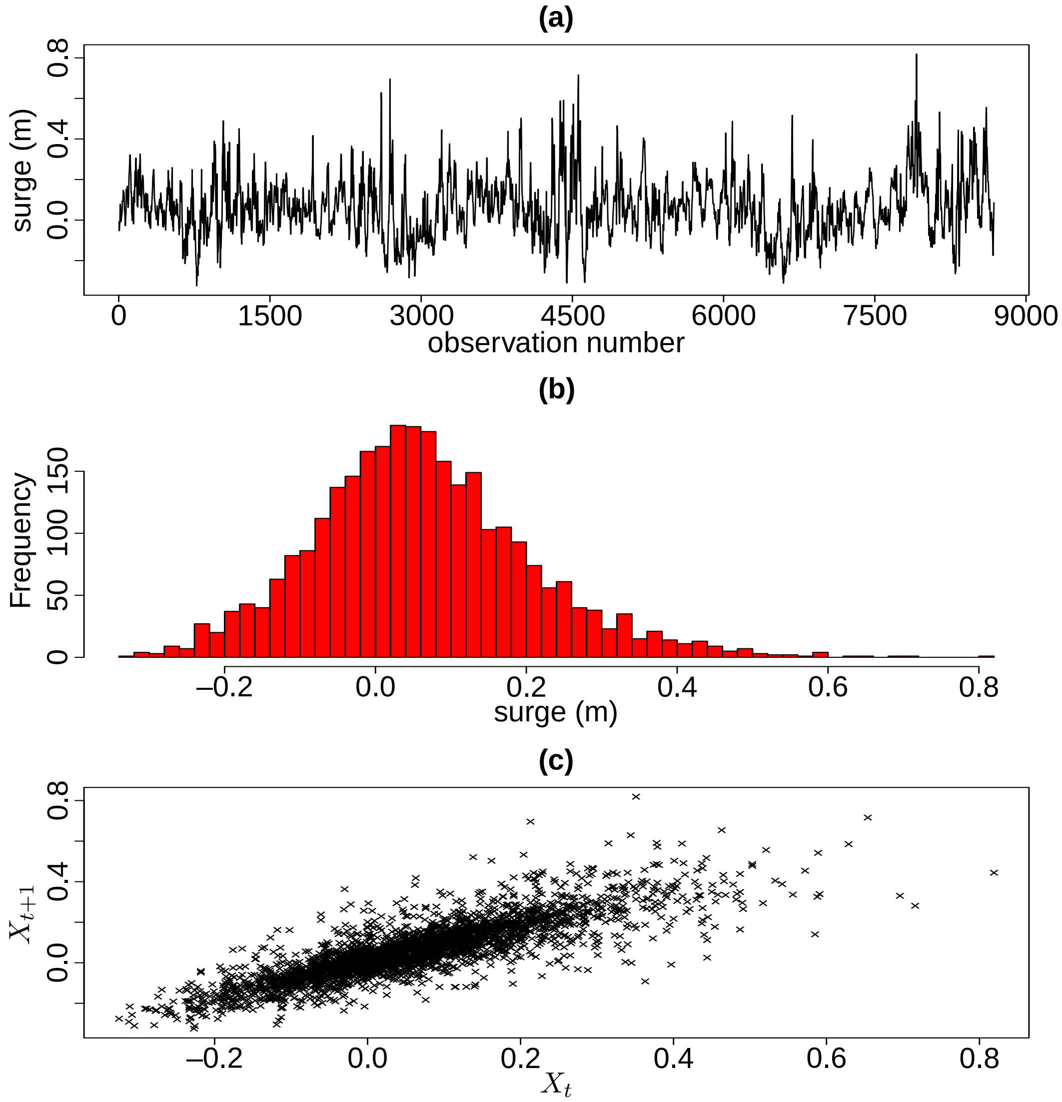

Figure 9: Newlyn sea–surge data: (a) Time series plot; (b) histogram; (c) plot of the time series against the series at lag 1.

'Removing' dependence

The most commonly adopted approach to circumvent the problems caused by such temporal dependence is to employ a declustering scheme to filter out a set of approximately independent threshold excesses. One method, which is often considered to be the most 'natural' way of identifying 'clusters' of extremes, is 'runs–declustering'. This is how it works:

1. Choose an auxiliary 'declustering parameter' (which we call κ)

2. A cluster of threshold excesses is then deemed to have terminated as soon as at least κ consecutive observations fall below the threshold

3. Go through the entire series identifying clusters in this way

4. The maximum (or 'peak') observation from each cluster is then extracted, and the GPD fitted to the set of cluster peak excesses.

This approach is often referred to as the *peaks over threshold* approach (POT, Davison and Smith, 1990) and is widely accepted as the main pragmatic approach for dealing with clustered extremes. Although this approach is quite easy to implement, there are issues surrounding the choice of κ; if

- κ is too small, the cluster peaks will not be far enough apart to safely assume independence
- κ is too large, there will be too few cluster exceedances on which to form our inference

It has also been shown that parameter estimates can be sensitive to the choice of κ. In this example, we use a separation interval of 60 hours (and so $\kappa = 20$) following the example of Coles and Tawn (1991), which should be large enough to safely assume independence between successively identified clusters allowing for wave propagation time. We used a mean residual life plot (see Part 1) to identify a suitably high threshold (0.3m).

The table below shows maximum likelihood estimates of the GPD scale and shape parameters σ and ξ, along with the associated 95% confidence intervals, fitted to the set of cluster peak excesses using $\kappa = 20$. Shown for comparison are the corresponding estimates using *all* threshold exceedances, ignoring temporal dependence. Note the discrepancy in the estimation of the two parameters under the two approaches; however, when allowing for sampling variability, these differences are not significant.

	$\hat{\sigma}$	$\hat{\xi}$
Cluster peaks	0.187	–0.259
95% confidence interval	(0.109, 0.265)	(–0.545, 0.027)
All excesses	0.104	–0.090
95% confidence interval	(0.084, 0.125)	(–0.215, 0.035)

Table 1: Maximum likelihood estimates, and associated 95% confidence intervals, for the GPD scale and shape parameters

'Ignoring' dependence

Table 1 above shows that, although there is a slight discrepancy in parameter estimation when using (i) cluster peak exceedances and (ii) *all* exceedances, these discrepancies are non–significant. Therefore, why bother declustering? Surely we're better off using *all* excesses?

The confidence intervals for the estimates using all excesses are too narrow – fitting to all exceedances when there is clearly evidence of short–term temporal dependence will result in underestimated standard errors. Smith (1991) suggests a procedure in which the usual asymptotic likelihood calculations are supplemented by empirical information on dependence, in order to produce a modified covariance matrix for the parameters, which is approximately correct after the dependence has been taken into account.

Under the model fitting procedure which assumes independence, denote the observed information matrix by H. If independence were a valid assumption, then the covariance matrix of the maximum likelihood estimates (m.l.e.s) would be approximately H^{-1}. Smith (1991) shows that to account for dependence this approximation should be replaced by $H^{-1}VH^{-1}$, where V is the covariance matrix of the likelihood gradient vector. Furthermore, V can be estimated by decomposing the log–likelihood sum into its contributions by year (which should be independent up to a good approximation) and obtaining the appropriate covariance matrix empirically.

Similar arguments can be applied to modify the procedure for testing hypotheses. Specifically, denoting model parameters by $\psi = (\rho, \zeta)$ where ρ and ζ are of dimensions p and q respectively, suppose that a test of $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$ is required, ζ being a nuisance parameter. Assuming independence, test procedures are usually based on the asymptotic distribution of

$$2\{\ell(\hat{\psi}_1) - \ell(\hat{\psi}_0)\}, \tag{9}$$

which is χ^2_p. Here, $\ell(\hat{\psi}_0)$ and $\ell(\hat{\psi}_1)$ denote the log–likelihood evaluated at the maximum likelihood estimate under H_0 and H_1 (respectively). Now suppose we wish to account for dependence. Partitioning

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

where H_{11}, H_{12}, H_{21} and H_{22} are the appropriate sub–matrices of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$ respectively, then we partition the inverse of H as

$$H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix},$$

where each sub–matrix $H^{\cdot\cdot}$ has the same dimensions as $H_{\cdot\cdot}$. Now let

$$C = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} - H_{22}^{-1} \end{pmatrix}.$$

Then Smith (1991) shows that the approximate distribution of expression (9) is given by

$$\sum_{i=1}^{p} \lambda_i z_i^2 \tag{10}$$

where the z_i, $i = 1, \ldots, p$, are standard normal variates and the λ_i are the non–zero eigenvalues of $V^{1/2}CV^{1/2}$. This replaces the usual χ^2_p–distribution, which is valid in the case of independence, and which would be recovered if all the λ_i were set equal to 1. It is then easy to simulate from the modified distribution (10) to estimate any required quantile of the test statistic. Profile likelihood confidence intervals then arise as the set of values of $\hat{\psi}_1$ such that the test statistic (9) is smaller than the quantile which represents the desired level of significance.

Table 2 reports maximum likelihood estimates for the GPD scale and shape parameters, along with their 95% confidence intervals, for analyses using *all excesses* and just *cluster peak excesses* (as before); in the analysis using information on all extremes, though, standard errors have now been inflated to account for temporal dependence via Smith's method (1991).

Table 3 shows maximum likelihood estimates for return levels for four return periods — $s = 10$, 50, 200 and 1000 years. The corresponding 95% confidence intervals have been obtained using the method of profile likelihood, where the appropriate cut–off for the test statistic (9) has been obtained using the modified distribution (10). In this way the profile likelihood confidence intervals have been inflated to account for the dependence in a way which is consistent with the modifications proposed by Smith (1991). Figure 10 shows a plot of the profile likelihood for one of these return levels — z_{50} — illustrating the severe asymmetry which is commonly observed for return levels. This plot is for the analysis using all threshold exceedances. The 95% profile likelihood confidence interval for z_{50}, after adjusting for dependence, is identified on the plot. Also shown is the much narrower interval which would have been obtained if dependence had been ignored.

Table 2 shows that, when the analysis is restricted to a set of cluster peak exceedances, the GPD scale parameter σ is overestimated, and the shape parameter ξ underestimated, relative to the approach which uses all exceedances. However, when we account for sampling variability, we see that these differences are not significant.

Of greater practical interest are the estimated return levels. Table 3 shows that estimates barely differ for the ten year return period, but are consistently smaller in the cluster peaks analysis for the other three periods studied — in fact, quite substantially so for the 200 and 1000 year return periods. Since estimates of such long–range return levels are often used as a design requirement in oceanographic situations (e.g. for the height of sea walls), designing to a level specified by an analysis based on cluster peak excesses could result in substantial under–protection.

	$\hat{\sigma}$	$\hat{\xi}$
Cluster peaks	0.187	–0.259
95% Confidence Interval	(0.109, 0.265)	(–0.545, 0.027)
All excesses	0.104	–0.090
95% Confidence Interval	(0.082, 0.126)	(–0.217, 0.037)

Table 2: Maximum likelihood estimates, and associated Wald 95% confidence intervals, for the GPD scale and shape parameters and the threshold exceedance rate when using all excesses, and just cluster peak excesses.

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{200}$	$\hat{z}_{1000}$
Cluster peaks	0.868	0.920	0.951	0.975
95% Confidence Interval	(0.770, 1.031)	(0.813, 1.099)	(0.838, 1.008)	(0.858, 1.063)
All excesses	0.867	0.947	1.007	1.068
95% Confidence Interval	(0.736, 1.067)	(0.790, 1.193)	(0.844, 1.257)	(0.891, 1.335)

Table 3: Maximum likelihood estimates, and associated 95% profile likelihood confidence intervals, for four return levels (units are in metres).

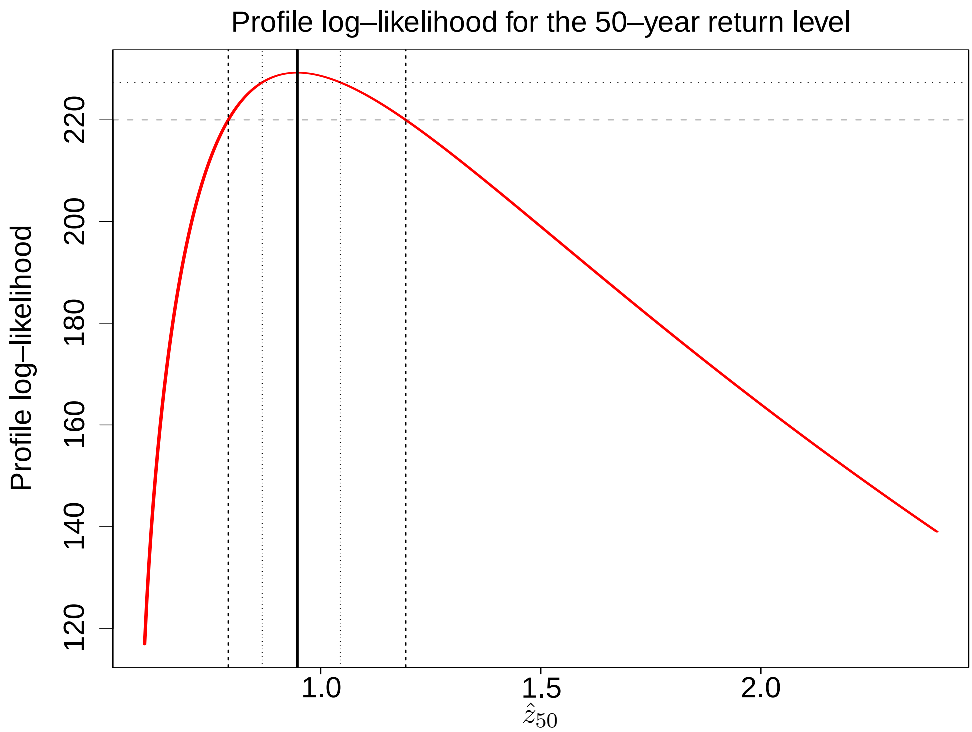

Figure 10: Profile log–likelihood surface, with corresponding 95% confidence intervals, for the 50 year return level $\hat{z}_{50}$. The dashed lines show the construction of the interval which has been inflated to account for temporal dependence in the sea–surge data (since in this example all threshold excesses were used). The dotted lines show how the interval would be constructed if dependence had been ignored.

Simulation study

So we know there are differences – some significant – in return level estimation when we use (i) cluster peak excesses and (ii) all threshold excesses. Which approach are we to trust?

— The usual approach is to use cluster peaks, then we have effectively removed temporal dependence

— However, return levels using this approach are underestimated relative to the procedure which uses all threshold excesses

— Using cluster peak excesses could result in substantial under–protection (i.e. not building a sea–wall high enough to protect against the 1 in 1000 year surge)

Figure 11 below shows some results of a simulation study undertaken by Fawcett and Walshaw (2007), in which the GPD was fitted to a simulated dataset for which the true values of σ, ξ and various return levels were *known*, and the strength of temporal dependence was similar to that of which is often observed in real–life environmental time series. The bold lines correspond to sampling distributions for the GPD parameters (and two return levels) using all threshold excesses, the thin lines correspond to the equivalent when using just cluster peak excesses. Clearly, for all parameters, the analysis using all threshold excesses outperforms that which uses just cluster peak excesses. Of most concern are the result shown for the two return levels; Fawcett and Walshaw (2007) found systematic underestimation of return levels when using cluster peak excesses (remember, this is the approach most commonly adopted to circumvent the problem of temporal dependence), whereas estimates of these return levels were much more accurate under the approach using all threshold excesses.

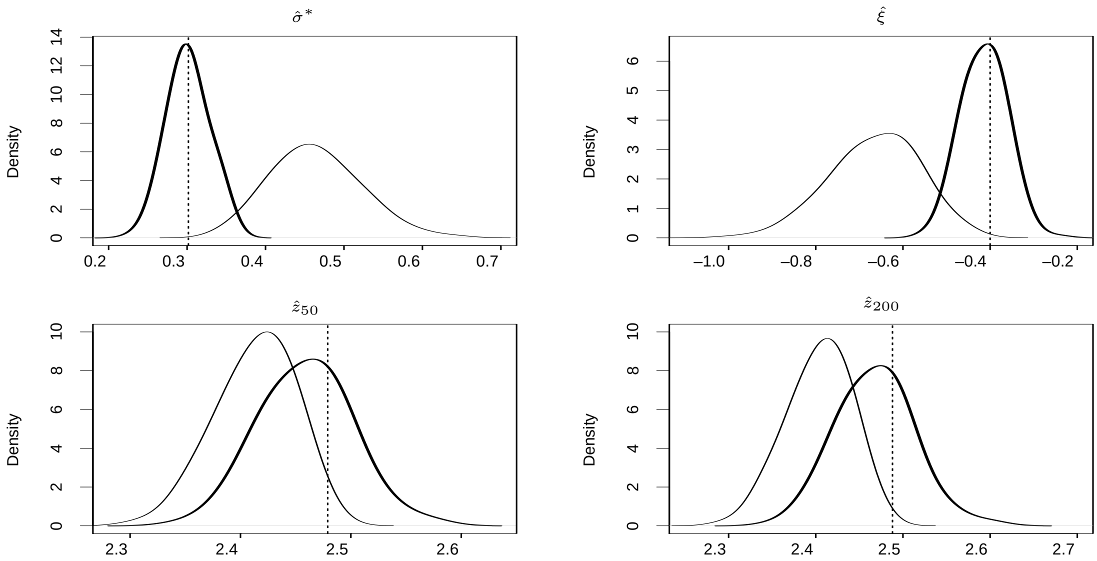

Figure 11: Sampling distributions of $\hat{\sigma}^*$, $\hat{\xi}$, $\hat{z}_{50}$ and $\hat{z}_{200}$ when $\alpha = 0.2$, and using all threshold excesses (heavy line) and cluster peak excesses (thin line). The *true* values for each parameter are shown by the vertical lines.

2.2 Non–stationarity: trend

In Section 2.1 we demonstrated that, subject to specified limitations, the usual extreme value limit models are still applicable in the presence of short–term temporal dependence. In fact, we can use the results for block maxima directly as they stand, though some thought is required when considering threshold models. The general theory cannot be extended for non–stationary series; instead, it is usual to adopt a pragmatic approach of using the standard extreme value models as basic templates that can be augmented by statistical modelling.

Figure 12 (over-leaf) below shows a time series plot of annual maximum sea levels observed at Fremantle, Western Australia, between 1900 and 1986; the right–hand–side plot shows these sea–levels plotted against the annual mean value of the *Southern Oscillation Index* (SOI), which is a proxy for meteorological volatility. There appears to be an increase in annual maximum sea levels through time, as well as an association between annual maximum sea levels and the mean SOI.

We can accommodate the time–trend shown in the plot on the left–hand–side of Figure 12 by fitting the GEV distribution (as we have annual maxima), but allowing for a linear trend in the underlying level of extreme behaviour. For example, if we define Z_t to be the annual maximum sea level at Fremantle in year t, then we might use

$$Z_t \sim \text{GEV}\left(\mu(t), \sigma, \xi\right)$$

where

$$\mu(t) = \beta_0 + \beta_1 t. \tag{11}$$

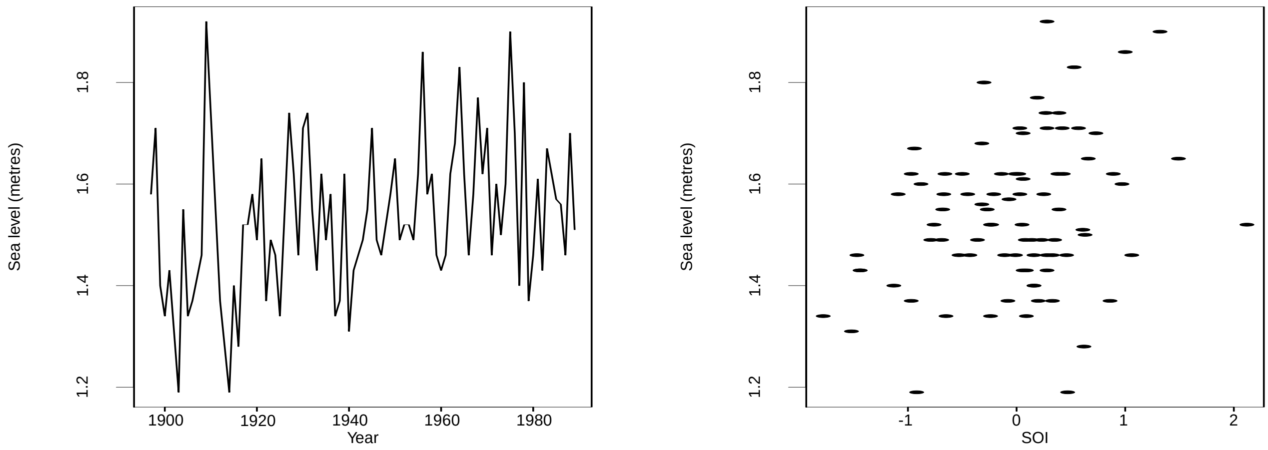

Figure 12: Time series plot of annual maximum sea levels observed at Fremantle (left), and a plot of the mean SOI against annual maximum sea level (right).

In this way, variations through time in the observed process are modelled as a linear trend in the location parameter of the appropriate extreme value model (the GEV in this case). We might choose to adopt the following model for $\mu(t)$:

$$\mu(t) \quad = \quad \beta_0 + \beta_1 \text{SOI}(t)$$

to allow for a linear association between the maximum sea level in year t and the SOI in year t. Or perhaps a textitmultiple linear regression model for $\mu(t)$, whereby

$$\mu(t) \quad = \quad \beta_0 + \beta_1 t + \beta_2 \text{SOI}(t); \tag{12}$$

we can then assess our preferences between the stationary model ($\mu(t) = \beta_0$), the models which allow for a dependence in time (alone), a dependence on SOI through time (alone), and the model which allows the underlying extremal behaviour to be determined by *both* a change in time *and* SOI, by referring to the usual likelihood ratio tests (since these models are nested). For example, fitting a stationary GEV distribution to these data, we get:

$$\hat{\mu} = 1.482(0.017) \qquad \hat{\sigma} = 0.141(0.011) \qquad \hat{\xi} = -0.217(0.064),$$

with a maximised log–likelihood of 43.6. Fitting the model which allows for a trend in time (the model shown in 11), we get:

$$\hat{\beta}_0 = 1.387(0.027) \qquad \hat{\beta}_1 = 0.002(0.0005) \qquad \hat{\sigma} = 0.124(0.010) \qquad \hat{\xi} = -0.128(0.068)$$

with a maximised log–likelihood of 49.79. Referring

$$\begin{aligned} D &= 2\{49.79 - 43.6\} \\ &= 12.38 \end{aligned}$$

to χ^2_1 tables, we have a significant result, suggesting that the model which includes a linear trend in time for μ explains substantially more of the variation in the data than the stationary model. Figure 13 shows the time series plot of the Fremantle sea level data with fitted estimates for μ superimposed. Also shown, for comparison, is the fitted estimate for μ under the stationary model.

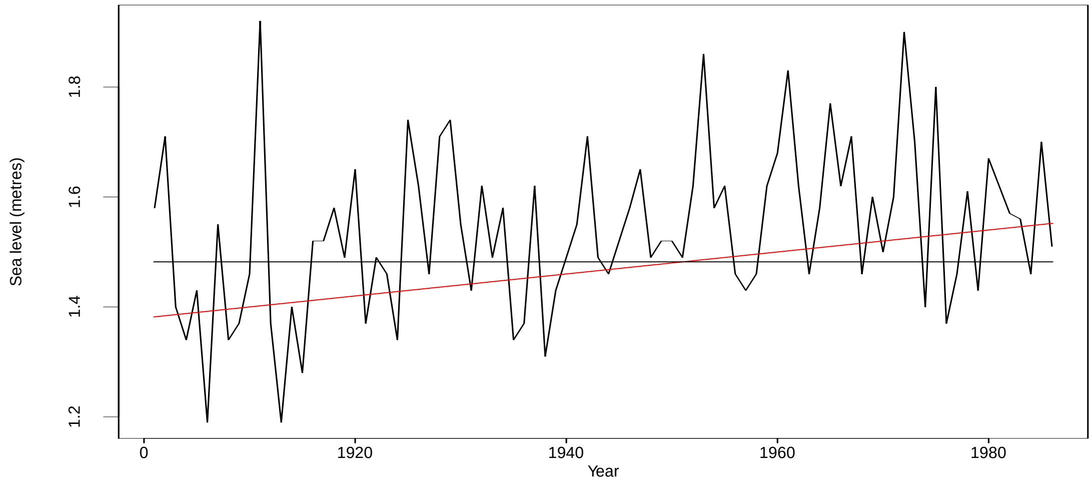

Figure 13: Time series plot of annual maximum sea levels observed at Fremantle, with fitted estimates for μ based on the stationary GEV model and the model which allows for a linear trend in time.

Similar methodology actually suggests that the model in equation 12 is the best model to use here, i.e. that which allows for a trend in μ depending on both time and SOI. In fact, we get:

$$\hat{\beta}_0 = 1.389(0.027) \qquad \hat{\beta}_1 = 0.002(0.0005) \qquad \hat{\beta}_2 = 0.055(0.020)$$

$$\hat{\sigma} = 0.121(0.010) \qquad \hat{\xi} = -0.154(0.064)$$

giving

$$\hat{\mu} \quad = \quad 1.389 + 0.002t + 0.055\text{SOI}(t).$$

Of course, more exotic model structures can be incorporated into this framework, including quadratic models, higher–order polynomial models, and models which allow for non–normal error structures. Trend can also be incorporated into the other GEV/GPD model parameters.

2.3 Non–stationarity: seasonality

The most widely adopted technique to deal with data which vary seasonally is to partition the data into seasons (within which we can assume the data to be homogeneous), and perform a separate extremal analysis on each season. Examples of such an approach can be found in Smith (1989) and Walshaw (1994). These seasons might be, for example, 'winter' and 'summer', or 'dry' and 'wet', where the seasonal variation is clearly understood. However, for data which exhibit less defined seasons, we can fit to separate months or years. Disadvantages of this approach are that a separate set of extremal parameters require estimating for each season, and that recombining these estimates is often non–trivial. To overcome these disadvantages, another approach is to allow the extremal parameters to vary continuously throughout the period of seasonality – for example, within the year. Fourier forms can be fitted to the parameters, and a model selected based on likelihood ratio tests. However, Walshaw (1991) suggests that inferences are barely altered in relation to a piecewise seasonality approach (for extreme wind gusts, anyway), and that the significant increase in computation time incurred by fitting continuously varying parameters is therefore not worthwhile.