Applied Ocean Research

Multivariate extrapolation in the offshore environment

S. Zachary[a],*, G. Feld[b], G. Ward[c], J. Wolfram[d]

[a]*Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, UK*
[b]*Seadata Division, Fugro GEOS, Seaforth Centre, Lime Street, Aberdeen AB11 5FJ, UK*
[c]*Department of Statistics, University of California, Berkeley, CA 94720, USA*
[d]*Department of Civil and Offshore Engineering, Heriot-Watt University, Edinburgh EH14 4AS, UK*

Abstract

We consider the estimation of the extremes of the metocean climate, in particular those of the univariate and joint distributions of wave height, wave period and wind speed. This is of importance in the design of oil rigs and other marine structures which must be able to withstand extreme environmental loadings. Such loadings are often functions of two or more metocean variables and the problem is to estimate the extremes of their joint distribution, typically beyond the range of the observed data. The statistical methodology involves both univariate and multivariate extreme value theory. Multivariate theory which avoids (often very inappropriate) prior assumptions about the nature of the statistical association between the variables is a fairly recent development. We review and adapt this theory, presenting simpler descriptions and proofs of the key results. We study in detail an application to data collected over a nine-year period at the Alwyn North platform in the northern North Sea. We consider the many problems arising in the analysis of such data, including those of seasonality and short-term dependence, and we show that multivariate extreme value theory may indeed be used to estimate probabilities and return periods associated with extreme events. We consider also the confidence intervals associated with such estimates and the implications for future data collection and analysis. Finally we review further both the statistical and engineering issues raised by our analysis. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Extreme value theory; Extreme loadings; Metocean data; Extreme wave statistics; Design waves

1. Introduction

One of the most critical features of the design process for offshore structures is the estimation of the worst loading conditions to which a given structure is likely to be exposed in its lifetime. Typically a structure must be designed to withstand, with some margin of safety, that loading which is expected to be exceeded with a frequency of, say, once in every hundred years [24]. This paper is concerned with the statistical problems which arise in such estimation, and with the application of statistical methodology to metocean data.

We distinguish two major statistical issues. The first of these is that estimation of extreme loadings may well require to be based on metocean data collected over a relatively short period. Reliable data collected over several hundred years will certainly not be available. Rather we may have useful data collected over a period of, say, 10 years. The problem is thus one of extrapolation of the observed distribution of data into its extreme region, typically lying

well beyond even the most extreme of the available observations. Statistical methodology for the extrapolation of the distribution of any single variable is well established, at least in the case where we have independent observations. This theory was initiated by Gumbel [13], and was originally concerned with the distribution of maximum observed values of the variable (for example, annual maxima). Asymptotic theory suggests that such maxima are well modelled by a generalised extreme value distribution (Fréchet, Gumbel, or Weibull, also known as Fisher–Tippett Types I, II and III and Gumbel Types I, II and III). Subsequently the theory has been much refined, and more efficient inference is now based on consideration of the excesses over a given threshold of all observations. See, in particular, Ref. [7] and, for an excellent and comprehensive account of both the probability and the statistical theory, Ref. [10]. In Section 2 we review briefly this univariate extreme value theory, and consider how it may be adapted and applied to metocean data, where we usually do not have independent observations and where underlying distributions have substantial seasonal variation. Some other approaches to the univariate

* Corresponding author

modelling of metocean data have been presented elsewhere [16,19,21].

The second issue is that of the extrapolation of multivariate distributions. The loading on a structure is estimated as a function of several variables, for example wave height, wave period, wind speed, wind direction and current, and we require to estimate probabilities corresponding to those combinations of these variables which result in extreme loadings. There are two reasonable approaches to this problem. The *structure variable* method identifies, prior to analysis, that function of the observed variables which best represents the loading on the specific structure of interest. Multivariate observations are then converted to univariate loadings, and univariate extreme value theory used to estimate the probabilities, or equivalently the return periods, associated with extreme events. The alternative approach, which we pursue in this paper, is to use multivariate extreme value theory to estimate directly (the extremes of) the *joint* distribution of the variables of interest. Possible extreme regions in the multidimensional variable space may then be identified and their associated probabilities estimated.

Multivariate extreme value theory is somewhat more complex than its univariate counterpart, and appropriate statistical methodology has only been developed in recent years. The basic theory is due to de Haan and Resnick [14], de Haan [15], and is further developed by Resnick [22], Joe et al. [17] and Coles and Tawn [4,5]. The major problem is that of correctly capturing the statistical association, or correlation, between the extremes of the variables concerned, without prior assumptions about the nature of this association, and in particular without assuming that estimates of association appropriate to the body of the distribution are also appropriate to its extremes. Earlier (and indeed some later) approaches to this problem did make such prior assumptions, which were frequently implicit, often very strong and often inappropriate to modelling in the extreme region of the variable space. Perhaps the best exposition of the modern theory is given by Coles and Tawn [4,5], who also discuss earlier multivariate approaches, and by Coles [3].

Coles and Tawn [5] also give detailed consideration of the relative merits of the univariate structure variable and the multivariate approaches, including discussion of the reliability of the statistical procedures associated with each. In particular, they argue, somewhat informally, that, for a given volume of data, the asymptotic theory underpinning both approaches is likely to be more accurate in the case of a multivariate analysis of the joint distribution of the original variables. For our present purposes a major disadvantage of the structure variable approach is that the load function must be fully identified prior to statistical analysis. Thus a long and complex analysis, which is here a far from automated procedure, must be performed for every possible structure under consideration. This creates considerable difficulties for design and optimisation. The multivariate approach requires (in principle) a single statistical analysis, the results of which may be applied to a wide variety of possible structures. Further, these results clearly provide considerable insight into the extremes of the metocean climate itself, in particular the association between the variables concerned. This is both of considerable scientific interest and important to good engineering design.

In outline, the multivariate approach is as follows. The first step is the application of *univariate* extreme value theory to estimate and extrapolate the marginal distribution of each of the variables under study. The relevant theory is described in Section 2.1. This step is in itself sufficient to answer questions about, for example, the 100-year return levels associated with individual variables—their likely values and associated confidence intervals.

The second step is the transformation of the multivariate observations so that the (marginal) distribution of each of the individual transformed variables has a standard (or unit) Fréchet distribution. This is followed by a *further* transformation of the data to (pseudo-) radial and angular components. Under these two successive transformations, the tail of the distribution of the radial component r turns out not to depend on the statistical association between the individual variables, and to be of known form (which is given simply, and unsurprisingly, by a rescaling of the tail of the standard Fréchet distribution). Further, the conditional distribution of the angular component $\mathbf{w}$ *given* r converges, as r is allowed to increase, to some limiting distribution. This limiting distribution therefore entirely captures the statistical association in, and beyond, the extremes of the data. It may be estimated from the the distribution of $\mathbf{w}$ in the set of those (transformed) observations for which r exceeds some suitably chosen threshold r_0. Hence the extremes of the joint distribution of r and $\mathbf{w}$ may be estimated. This theory is described in detail in Section 2.2.

The third, and final, step is simply the inversion of the transformations described above so as to recover an estimate of the extremes of the joint distribution of the original variables. This will both match the extremes of the observed data and extrapolate their joint distribution as is necessarily implied by the above asymptotic theory. The analytical details of this final step depend on the exact representation of the angular component $\mathbf{w}$ of the transformed observations. In the metocean application of Section 3 we show how this is achieved in practice.

This multivariate approach, as originally developed, was based on a point process representation of the data, which involved a not very intuitive renormalisation of observations by their total number. We take the opportunity in Section 2.2 to present a mathematically equivalent but considerably simpler description, which is essentially that outlined above. Section 2.3 and Section 2.4 consider problems of seasonality and short-term dependence such as are typical of metocean data.

In Section 3 we study in detail an application to metocean data. These consist of hourly observations of significant wave height, wave period, and wind speed, collected at

the Alwyn North platform in the northern North Sea over a nine-year period. We consider the many problems arising in the analysis of such data, including those of seasonality and short-term dependence, and we show that multivariate extreme value theory may indeed be used to estimate probabilities and return periods associated with extreme events. We consider also the confidence intervals associated with such estimates and the implications for future data collection and analysis.

In Section 4 we consider further both the statistical and engineering issues raised by our analysis.

2. Methodology

As discussed in Section 1, methodology for the analysis of multivariate extremes must address two problems. The first is the estimation and extrapolation of the (marginal) distributions of individual variables and the determination of the probabilities of extreme events associated with them. In Section 2.1 we review the relevant univariate extreme value theory and discuss its application.

The second problem is that of correctly modelling the statistical association between the variables involved, in particular the association in the *extreme region*, or *extremes*, of the (multidimensional) variable space. We define this to be that region in which *any* of the variables under study is extreme. This is the region which corresponds to extreme loads on, for example, offshore structures. In Section 2.2 we describe and discuss the necessary multivariate extreme value theory. This theory allows an arbitrary association structure in the extreme region to be estimated directly from the data. We also give a more accessible description of the relevant mathematics.

Throughout Section 2.1 and Section 2.2 we assume that the data for analysis may reasonably be modelled as independent and identically distributed multivariate observations $\mathbf{X}_1,\ldots,\mathbf{X}_n$ of a random vector $\mathbf{X} = (X_1,\ldots,X_d)$. Thus, for each $i = 1,\ldots,n$, $\mathbf{X}_i = (X_{i1},\ldots,X_{id})$ is the corresponding vector of observations of the d variables under study. Of course, in applications such as that to metocean data considered here, it will frequently be the case that this assumption does not hold: there will be considerable seasonal variation in the distribution of the data, and additionally the data will naturally exhibit short-term dependence over sufficiently short periods of time. We consider these problems in Section 2.3 and Section 2.4.

2.1. Estimation of marginal distributions

2.1.1. Probability theory

For each j ($1 \leq j \leq d$), let F_j be the distribution function of the variable X_j. In the estimation of this distribution, the main concern is to model its tail as accurately as possible. This is essential in order to permit reliable extrapolation to values more extreme than those actually observed. We use

well-established univariate extreme value theory. We review this briefly here, but much more detail is given elsewhere [10].

The current, much refined, approach to this theory associates a suitably chosen *threshold u_j* (see below) with each variable X_j and considers the annual rate $g(x)$ at which observations of X_j exceed any given level $x \geq u_j$. We assume that the number of observations per year is reasonably large, and that the threshold u_j is such that the probability of an individual observation exceeding it—and so also the probability of the observation exceeding any $x \geq u_j$—is reasonably small. (Both of these conditions will be comfortably satisfied in the application considered in Section 3 of this paper.) It then follows from the assumed independence of the observations of X_j that *exceedances* of any $x \geq u_j$ occur as an (approximate) Poisson process and that the maximum observed value of X_j in any interval of time of length t years has a distribution function G given, for $x \geq u_j$, by

$$G(x) = \exp\{ - tg(x)\} \tag{1}$$

Again provided that the threshold u_j is chosen sufficiently large, asymptotic theory [10] now implies that $G(x)$ should correspond to a *generalised extreme value* distribution for all $x \geq u_j$. This distribution is described by three parameters—*its shape ξ, location μ*, and *scale $\sigma > 0$*—and is a Fréchet, Gumbel, or Weibull distribution according as the shape parameter $\xi > 0$, $\xi = 0$, or $\xi < 0$. In order that G should be a generalised extreme value distribution, we require that the (annual) *exceedance rate function* $g = g_{\xi,\mu,\sigma}$ be given by

for $\xi > 0$

$$g_{\xi,\mu,\sigma}(x) = \left[1 + \frac{\xi(x - \mu)}{\sigma} \right]^{-1/\xi}$$

provided $x > \mu - \sigma/\xi$ (since we certainly require $u_j > \mu - \sigma/\xi$ this condition causes no problems);

for $\xi = 0$

$$g_{\xi,\mu,\sigma}(x) = \exp\left(- \frac{x - \mu}{\sigma} \right);$$

for $\xi < 0$

$$g_{\xi,\mu,\sigma}(x) = \begin{cases} \left[1 + \dfrac{\xi(x - \mu)}{\sigma} \right]^{-1/\xi} & \text{if } x < \mu - \sigma/\xi \\ 0 & \text{if } x \geq \mu - \sigma/\xi. \end{cases}$$

Here the appropriate values of the parameters (ξ,μ,σ) will of course depend on the variable X_j under consideration, and are to be estimated from the corresponding observations of that variable.

Observe that, despite its apparent complexity, the above definition of the exceedance rate function $g_{\xi,\mu,\sigma}$ has a natural mathematical coherence. In particular, for all x, $g_{0,\mu,\sigma}(x) = \lim_{\xi \to 0} g_{\xi,\mu,\sigma}(x)$, whether the limit is taken from below or

above. Observe also that (for all values of the parameters (ξ,μ,σ)) $g_{\xi,\mu,\sigma}(x)$ is decreasing in x and tends to zero as x tends to infinity.

Let n_y be the number of observations per year. It follows from Eq. (1), with $g(x) = g_{\xi,\mu,\sigma}(x)$, that the probability of an individual observation exceeding any $x \geq u_j$ is given by $\exp\{-(1/n_y)g_{\xi,\mu,\sigma}(x)\}$. Since under the above assumptions this probability is small, it further follows that, to a very good approximation, the distribution function F_j of X_j satisfies

$$F_j(x) = 1 - \frac{1}{n_y}g_{\xi,\mu,\sigma}(x), \quad \text{for all } x \geq u_j \tag{2}$$

(Indeed, under the above model, this result is asymptotically exact as either n_y or u_j increases.) We now treat Eq. (2) as exact, as is the usual practice, and base our inference directly on it. In the context of the above model, and for sufficiently large n_y or u_j, there is negligible loss in doing so, and there are considerable analytical advantages. Further discussion on the choice of threshold u_j is given in Section 2.1.2. However, it must be such that there lie above u_j a sufficient number of observations to permit the inference described below, and of course the extreme and generally unobserved values of X_j in which we are primarily interested will also lie above it.

Note that the alternative, and earlier, approach of consideration of only the maxima of sequences of observations sufficiently long as to require the use of the generalised extreme value distribution itself (for example, yearly maxima) involves a loss of information which is perhaps considerable, and so is an inefficient basis for inference [10]. For some further discussion here see also Section 2.4.

Note also that, in the above model, the parameters (ξ,μ,σ) have an interpretation in terms of yearly exceedance rates. Thus, in principle, estimates of these parameters should be reasonably stable under variation of n_y, provided that the total number of observed exceedances per year of each $x \geq u_j$ remains at least approximately constant—again see Section 2.4.

2.1.2. Estimation

We now consider the estimation of a suitable threshold u_j and of the parameters (ξ,μ,σ) of the model defined in the previous section. We do this by deriving, in terms of this model, the probability of threshold exceedance by any individual observation and the distribution of the associated threshold *excess*. We show below that comparison with the observed distribution of threshold excesses, for a varying threshold, enables u_j to be estimated. Then, for the estimated u_j, the available data may be used to further estimate both the above exceedance probability and the parameters of the excess distribution. From these it is straightforward to recover estimates of the original parameters (ξ,μ,σ).

Eq. (2) is further equivalent to the requirement that, for any $x \geq u_j$ and $z \geq 0$

$$\Pr\{X_j > x + z | X_j > x\} = \frac{1 - F_j(x+z)}{1 - F_j(x)} = \frac{g_{\xi,\mu,\sigma}(x+z)}{g_{\xi,\mu,\sigma}(x)}$$

$$= g_{\xi,0,\sigma(x)}(z) \tag{3}$$

Here the last equality follows, after some straightforward manipulation, from the definition of the function $g_{\xi,\mu,\sigma}$ (whatever the sign of ξ) and

$$\sigma(x) = \sigma + \xi(x - \mu) \tag{4}$$

Hence, the distribution of the excesses of x is a *generalised Pareto distribution* (GPD) with distribution function $1 - g_{\xi,0,\sigma(x)}$. This distribution is described by its shape parameter ξ, which is independent of x and unchanged from that in Eq. (2), and by its scale parameter $\sigma(x)$ which is given by Eq. (4). Its mean is given by

$$\frac{\sigma(x)}{1-\xi} = \frac{\sigma - \xi\mu}{1-\xi} + \frac{\xi}{1-\xi}x \tag{5}$$

(See, for example, Ref. [10] for the relevant background on the GPD.) Additionally, again from Eq. (2), the probability of an exceedance of x occurring is

$$1 - F_j(x) = \frac{1}{n_y}g_{\xi,\mu,\sigma}(x) \tag{6}$$

We may therefore take the threshold u_j to be any value whose excesses (in the observations $(X_{1j},...,X_{nj})$ of X_j) are well modelled by a GPD, provided only that there are sufficient exceedances of u_j to permit the inference we now describe. (This of course requires that we have available a sufficiently large data set.) We may estimate the corresponding parameters ξ and $\sigma(u_j)$ of this GPD by maximum likelihood estimation, and may similarly estimate the threshold exceedance probability $1 - F_j(u_j)$ by the proportion of observed exceedances of u_j (this is again the maximum likelihood estimate of this probability). Once these estimates are available, the parameters σ and μ of the underlying model may be recovered by using Eqs. (4) and (6), in each case with $x = u_j$, and by recalling the definition of the function $g_{\xi,\mu,\sigma}$. Maximum likelihood estimation has good statistical properties and additionally permits assessments of uncertainty, for example confidence intervals, for the parameters (ξ,μ,σ), and also for those quantities, such as return levels, which are functions of them. For a further discussion see Refs. [7,12,19].

The distribution of X_j below the threshold u_j may be estimated by smoothing the empirical distribution of the corresponding observations $(X_{1j},...,X_{nj})$ in this region. In practice the threshold u_j is such that this is not a problem (see Section 2.1.1). In the metocean application of Section 3 we use kernel density estimation [23], but the estimate of the distribution in this region will be relatively insensitive to the choice of (sensible) smoothing procedure. Further, our primary interest is in the estimate of extremal behaviour

which is unaffected by the exact choice of smoothing procedure below the threshold.

One commonly used aid to the identification of a suitable threshold u_j is the *mean excess plot*, in which the mean of the excesses of each x is plotted against x. Since, for the given model (with fixed (ξ,μ,σ)), the excesses of u_j, and so also of each $x \geq u_j$, follow a GPD, Eq. (5) suggests that this plot should be approximately linear beyond any suitable threshold u_j. In reality considerable experience is required in the interpretation of such plots, for the typically long tail of the GPD ensures that they are visually dominated by a large range of values of x to which there correspond only a small, or very small, number of exceedances. Because these observations represent only a small sample from the underlying distribution, they considerably distort the plot over much of its range. For further discussion, and for some very instructive simulations, see Ref. [10].

In the application of Section 3 we supplement mean excess plots by more direct checks on the suitability of the chosen thresholds, notably comparison of observed and fitted distributions above the thresholds, and by appropriate sensitivity analysis.

2.1.3. Calculation of return levels

We may use the above theory to calculate the return levels of the variable X_j to be associated with specified return periods. Under the model of Section 2.1.1, the return level x_p associated with a return period of $1/p$ years is given by the solution of

$$g_{\xi,\mu,\sigma}(x_p) = p \tag{7}$$

that is, by

$$x_p = \begin{cases} \mu + \dfrac{\sigma}{\xi}(p^{-\xi} - 1) & \text{if } \xi \neq 0 \\[2mm] \mu - \sigma \log p & \text{if } \xi = 0 \end{cases} \tag{8}$$

provided only that, as will be the case in applications, x_p lies above the threshold u_j.

Confidence intervals associated with a given estimate of the return level x_p are best obtained by determining the associated profile likelihood function [10]: for each possible value of x_p, we calculate the maximum value $l(x_p)$ of the log-likelihood of the observed data over the set of those parameters (ξ,μ,σ) such that Eq. (7) holds. (In practice this involves reparametrisation of the log-likelihood in terms of x_p and two remaining parameters.) The value $\hat{x}_p$ of x_p which maximises $l(x_p)$ is of course the maximum likelihood estimate of the return level and, based on the maximum likelihood ratio test, a confidence region for x_p of size α is given by

$$\left\{ x_p : l(\hat{x}_p) - l(x_p) \leq \tfrac{1}{2}c_\alpha \right\}$$

where c_α is the upper α point of the chi-squared distribution with one degree of freedom [6]. For a 95% confidence interval we take $c_{0.95} = 1.92$. The typically non-normal shape of

the profile likelihood function l confirms the necessity of an approach such as this—in contrast to basing the interval on the assumption of a normal distribution for the maximum likelihood estimate.

2.2. Multivariate extreme value theory

In estimating the joint distribution of the variables $X_1,\ldots,X_d$, it is of crucial importance to capture correctly the dependence structure in its extremes. We again appeal to asymptotic theory and seek a representation of the data in which the measure of dependence in the extreme region is necessarily stable (in a sense to be made clear below), and may be estimated from the data without prior modelling assumptions. The relevant theory is based on the properties of the Fréchet distribution with shape parameter 1 (to which the marginal distributions of the data may be transformed) and is described by Coles and Tawn [4,5], who also consider estimation issues. We give a somewhat different presentation of this theory. We then consider how it may be applied to the analysis of data.

2.2.1. Asymptotic theory

Suppose that the random vector $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d)$ is such that the (marginal) distribution of each of its components $\tilde{X}_j$ has a standard Fréchet distribution, i.e. a distribution function F on $\mathbb{R}_+ = [0,\infty)$ given by

$$F(x) = \exp(-1/x) \tag{9}$$

Define radial and angular components of the vector $\tilde{\mathbf{X}}$ by

$$r = \tilde{X}_1 + \ldots + \tilde{X}_d \tag{10}$$

$$w_j = \tilde{X}_j/r, \; j = 1, \ldots, d \tag{11}$$

Note that the random vector $\mathbf{w} = (w_1,\ldots,w_d)$ takes values in the space $S_d = \left\{ \mathbf{w} \in \mathbb{R}_+^d : \sum_{j=1}^d w_j = 1 \right\}$. In particular it can be specified by giving any $d - 1$ of its components.

We now have the following results (for the derivation of which see below). The density function $\hat{f}$ of the radial component r satisfies

$$\hat{f}(r) = \frac{d}{r^2} + o\left(\frac{1}{r^2}\right) \text{ as } r \to \infty \tag{12}$$

(where $o(1/r^2)$ denotes further terms of a higher order which become negligible in relation to $1/r^2$ as r increases). Further, the conditional distribution of the angular component $\mathbf{w}$ given r converges, as $r \to \infty$, to a probability measure μ on S_d—so that in particular it is asymptotically independent of r. This probability measure satisfies

$$\int_{S_d} w_j \, d\mu(\mathbf{w}) = \frac{1}{d}, \; j = 1, \ldots, d \tag{13}$$

that is, the expected value of each of the d margins of the measure μ is $1/d$. (Where μ has a density, as is the generally the case in applications, the left-hand side of Eq. (13) is

simply the integral of w_j with respect to this density.) In other respects the value of μ is arbitrary.

Now define an *extreme* of the random vector $\tilde{\mathbf{X}}$ to be any value for which the radial component r is sufficiently large. The significance of the above results is that (from Eq. (12)) the distribution of r is asymptotically independent of the dependence structure in the joint distribution of $\tilde{\mathbf{X}}$. Further, the dependence structure in the extremes of this joint distribution (the area of interest here) is captured entirely by the limit measure μ. Hence, estimation of this extreme dependence structure reduces to estimation of μ. Finally, from Eq. (12) again, knowledge of μ implies knowledge of the entire joint distribution of the extremes of $\tilde{\mathbf{X}}$.

Note that the measure μ need not have a density on S_d. For example, if $\tilde{X}_1, ..., \tilde{X}_d$ are independent, then μ assigns weight $1/d$ to each of the d extreme points of the space S_d (each of which has one component which takes the value 1 and the remaining components equal to 0).

These results are usually presented in terms of point process convergence. Let $\tilde{\mathbf{X}}_1, ..., \tilde{\mathbf{X}}_n$ be n independent observations of the random vector $\tilde{\mathbf{X}}$. Then, as $n \to \infty$, the point process on $\mathbb{R}_+^d$ defined by $\tilde{\mathbf{X}}_1/n, ..., \tilde{\mathbf{X}}_n/n$ converges in distribution on $\mathbb{R}_+^d \setminus \{0\}$ (i.e. away from the origin) to a heterogeneous Poisson process. The radial and angular components of the limit distribution are independent and that of the angular component is arbitrary, except for a normalisation condition which corresponds to Eq. (13) above. For a description of this theory, which is due to de Haan [15], see Refs. [3–5]. (Note that the intensity measure H on S_d used by Coles and Tawn differs by a factor of d from the measure μ given here. The reason for this is that the independence in the above point process limit distribution corresponds to a factorising of its intensity into radial and angular components, and an arbitrary multiplicative constant may be transferred from one component to the other. Here, we have chosen to state the key results in terms of the probability distribution of the random vector $\tilde{\mathbf{X}}$, so that the factor d naturally belongs to the density function $\hat{f}$ of r, rather than to the probability measure μ.)

The results as stated here (Eq. (12) describing the asymptotic distribution of r, the convergence to μ of the conditional distribution of $\mathbf{w}$ given r, and the *normalisation* condition, Eq. (13)) may be regarded as a simple restatement of the more usual point process description of the theory: they are implied easily and directly by the point process convergence referred to above, together with the limiting form of the intensity measure. However, a direct proof of these results, as presented here, is straightforward. For a simple, and slightly informal proof, see Appendix A.

2.2.2. Application

To apply the theory of Section 2.2.1, we first determine a transformation $\psi : \mathbb{R}^d \to \mathbb{R}_+^d$ such that the transformed vector observations $\tilde{\mathbf{X}}_i = \psi(\mathbf{X}_i), i = 1, ...n,$ have unit Fréchet marginal distributions. The appropriate transformation here is given by $\tilde{X}_{ij} = \psi_j(X_{ij})$ where, for each $j = 1, ...d,$ the function ψ_j is given by $\psi_j(x) = -\{\log F_j(x)\}^{-1}$ and where F_j is the marginal distribution function of the observations X_{ij} estimated as described in Section 2.1.2.

The transformed vectors $\tilde{\mathbf{X}}_1, ..., \tilde{\mathbf{X}}_n$ are now regarded as independent observations of a random vector $\tilde{\mathbf{X}}$ as described in Section 2.2.1. The problem is to estimate the limit measure μ defined there, and so the joint distribution of the extremes of $\tilde{\mathbf{X}}$. Inversion of the transformation ψ then gives the corresponding estimate of the joint distribution of the extremes of the original vector observations $\mathbf{X}_1, ..., \mathbf{X}_n$. (Details of how this is achieved in practice are given in Section 3.2.4.)

Thus, for each observation i, we *further* transform the vector $\tilde{\mathbf{X}}_i$ to $(r_i, \mathbf{w}_i)$ where

$$r_i = \tilde{X}_{i1} + ... + \tilde{X}_{id}$$

$$w_{ij} = \tilde{X}_{ij}/r_i, \ j = 1, ..., d$$

and $\mathbf{w}_i = (w_{i1}, ..., w_{id})$. That is, for each i, $(r_i, \mathbf{w}_i)$ is the corresponding observation of the random pair $(r, \mathbf{w})$ defined by Eqs. (10) and (11). For each $r_0 \geq 0$, let N_{r_0} denote the set of those observations i such that $r_i > r_0$. It now follows from the convergence to μ of the conditional distribution of $\mathbf{w}$ given r (see Section 2.2.1) that, for some sufficiently large threshold r_0, the limit measure μ may reasonably be estimated from the observed distribution of those $\mathbf{w}_i$ with $i \in N_{r_0}$.

The problem of choosing this threshold is analogous to that of choosing the thresholds u_j in the estimation of marginal distributions as described in Section 2.1. That is, r_0 must be sufficiently high for the limit distribution μ to be well approximated by the conditional distribution of $\mathbf{w}$, given $r > r_0$; however, it must also be low enough for there to be sufficient observations in the set N_{r_0} to permit reliable estimation. Some further discussion is given elsewhere [5]: essentially, for varying r, we examine the distribution of $\mathbf{w}_i$ for $i \in N_r$ and then take r_0 to be the lowest value of r above which this distribution is stable. We observe in Section 3.2.4 that in practice our results appear to be less sensitive to the exact choice of r_0 than is the case with the choice of the marginal thresholds u_j.

Once r_0 is determined the limit measure μ is estimated, as indicated above, using those observations in the set N_{r_0}. Coles and Tawn [5] take a parametric approach to this estimation: from within a large parametrised family of possible distributions they use maximum likelihood estimation to determine that which best fits the data. In the application of Section 3 we have preferred to use instead non-parametric estimation, as the applicability of any known parametric family to the current metocean data is unclear. (The parametric approach does, however, have the advantage of making more straightforward the determination of assessments of uncertainty about μ.) In the present paper we have used kernel density estimation [23], with a variable kernel width, to estimate the distribution of $\mathbf{w}$ in N_{r_0} and hence the measure μ. The details are described in Section 3.2.4.

2.3. Seasonality

The distribution of metocean data does of course exhibit very substantial seasonal variation. In general this is not too serious a problem. In the consideration of extreme events—those with a return period measured in years—it is the distribution of the data over each year which is important. This is simply the mixture of the seasonal distributions and should itself be well modelled in the tail by a GPD. The alternative procedure is to explicitly model seasonal dependence by allowing parameters, and perhaps thresholds, to vary appropriately. The simplest possibility here is to partition the data according to season, analyse separately the data for each such season, and appropriately combine the results. However, this approach is messy, the partitioning involved is somewhat arbitrary, and more data are probably required in order to obtain reliable results [1].

However, in the application to metocean data considered in Section 3, we do partially take the latter approach in order to cope with problems of missing data. There are substantial numbers of missing observations for the summer months of May to August, when the extreme events of interest have an exceedingly small probability of ever occurring (see Section 3.1 and Appendix C for further details). We treat the probability of such occurrences during this period as being zero. Thus we analyse only those data corresponding to the remaining eight months of the year, and, in the calculation of return levels and other time-dependent quantities, adjust the number n_y of observations periods per year to correspond to this eight-month period. (There are also some further missing observations *within* this eight-month period. Because they are distributed approximately uniformly throughout this period, they do not cause a problem beyond the loss of data. Nor, of course, does n_y require any further adjustment on account of them. Again see Appendix C for further details.)

2.4. Dependence

The theory described in Section 2.1 and Section 2.2 assumes the independence of successive observations. In practice, however, there is considerable short-term dependence in such metocean data. In the application of Section 3 observations are available at hourly intervals, while extreme events are typically to be found in storms, which may last for many hours or even several days. It is therefore very important to consider the effect of such short-term dependence in the data.

To the extent that the tail of the distribution of the data is still well modelled by a GPD, the above, independence-based, methodology should still correctly estimate the frequencies of extreme events, provided that events which would occur in disjoint observation periods are to be regarded as distinct. (The reason for this is that these frequencies are essentially an extrapolation of the frequencies of the extreme events observed in the data.) However the tendency of extreme events to cluster affects our conclusions in two further ways:

1. For any given extreme event, the average time between successive clusters of this event is greater than the average time between individual events. In applications, the former is the most relevant definition of a return period, whereas the latter is the return period estimated on the assumption of independent observations. Therefore the use of this assumption tends to result in underestimation of return periods, as relevant for applications.
2. Assessments of uncertainty (standard errors and confidence intervals for parameters) are again based on the assumption of independent observations. However, if there is dependence in the data, then there is less information available for these assessments than is being assumed, so that they are unduly optimistic.

Various approaches to this problem have been discussed in the literature. Those which address both the above difficulties typically either (a) 'de-cluster' the data, by explicitly identifying clusters of observations, which may reasonably be regarded as independent of each other, and then choosing the most extreme observation from within each such cluster, or (b) attempt to estimate the degree of clustering in the data—by, for example, estimation of the *extremal index* [10,18]—and to adjust for it, or (c) attempt to explicitly model the dependence structure in the data. There are some problems with the last two approaches. Estimation of the extremal index is difficult and it is not even clear that it is consistently defined throughout the range of extreme values of the variables. Hence its application to the adjustment of return periods, or return levels, appears hazardous. Similarly, explicit modelling of the dependence structure is still very much a topic of current theoretical research and further inevitably involves greatly increased computational complexity.

In the analysis of Section 3, we therefore take an approach which is essentially equivalent to (a) above. We divide the interval over which the data were collected into *blocks* or *periods* $i = 1,...,n$ of equal length (say 24, 48 or 72 h). With each such period i we associate a multivariate observation $\mathbf{X}_i$, which is defined to be the most extreme (in an appropriate sense—see below) of the hourly multivariate observations made during that period. The periods are chosen to be of sufficient length that it is reasonable to treat the observations $\mathbf{X}_1,...,\mathbf{X}_n$ as independent, and to analyse them as described in the preceding sections. There is of course a slight problem where the extremes of the original hourly observations cluster at the boundaries between successive periods, but the effect of this is relatively minor, and, as described above, to the extent that there is any further failure of independence its effect with respect to return periods and return levels is conservative.

Note that even if a period length longer than necessary is chosen, it is relatively rare that an extreme observation is lost because it is masked by another from which it might

have been considered genuinely independent. Hence the results of our analysis—the univariate parameters which describe exceedance rates as defined in Section 2.1, their associated confidence intervals, and the multivariate analysis of Section 2.2—should not be too sensitive to the choice of period length. (See also Section 3.2.6 for some direct experimentation here.) This increases our confidence that there is little loss of reliably useful information in the above approach.

We study the choice of an appropriate period length below. However, it is sensible to consider first the appropriate determination of the multivariate observation $\mathbf{X}_i$ to be associated with each period i. As remarked above, this is in some sense the most extreme observation associated with that period.

The simplest approach is to define each component observation X_{ij} to be the most extreme of the corresponding individual observations of variable X_j made during the period i. However, in the multivariate context this may well be unduly conservative as different variables may attain their most extreme values at different times within this period.

We allow the above (conservative) approach as one possibility. However, we also consider the approach whereby one particular variable X_j is considered to be of primary importance for the application concerned. For each period i we then define X_{ij} to be the maximum observed value of X_j within this period. For each remaining variable $X_k (k \neq j)$ we define X_{ik} to be the *concomitant* observation of this variable, i.e. the observation of X_k which is made at the same time as the observation X_{ij}. (In the application considered in Section 3 this is the observation of X_k made in the same hour as the observation X_{ij}.) This is a slight departure from a truly multivariate approach to the problem, and the results of the analysis will vary slightly according to which variable is in fact considered to be of primary importance.

There is also a further possibility, which is, however, structure specific. Each of the observations within each period i is converted to an overall *loading* for the structure of interest, and the multivariate observation $\mathbf{X}_i$ is then defined to be the most extreme of the original observations as defined by this loading. This has similarities with the univariate structure variable approach discussed in Section 1, but it has the advantage of allowing information regarding the constituent elements of the total loading to be retained. For a *known* structure, this would be perhaps the most satisfactory solution to this particular problem.

We now consider the problem of choosing an appropriate period length. There is considerable discussion in the literature about this problem in the context of metocean data [11,20,26]. The simplest way to examine the suitability of any given choice is to look at the residual dependence structure of the corresponding associated observations $\mathbf{X}_1,\ldots,\mathbf{X}_n$. These should ideally be independent. It seems sufficient to do this separately for each of the component variables, since any serial dependence in the sequence of multivariate observations will almost certainly show in the corresponding sequence for at least one of the component variables. In the analysis of Section 3 we use two techniques: (a) examination, for each j, of the serial autocorrelation structure of $X_{1j},\ldots,X_{nj}$; and (b) examination, for each j, of the serial dependence of the occurrence of threshold exceedances by the observations $X_{1j},\ldots,X_{nj}$. For the latter we simply note whether or not each X_{ij} exceeds some suitable threshold (in practice it seems sensible to use that threshold above which the observations appear reasonably modelled by a GPD) and look at the first-order dependence of these events. This is very simple, but focuses attention on the dependence structure in the extremes of the data, and this is what is required.

It is important to note that both of the above techniques require that the data be separated more carefully—for this purpose—according to season. Otherwise the inevitable seasonal dependence introduces an essentially spurious appearance of short-term dependence in the data.

3. Application to metocean data

In this section we consider an application to data consisting of observations of three metocean variables collected in the northern North Sea. We restrict attention to univariate and bivariate analyses. This is sufficient both to illustrate the above theory and, as we shall show, to obtain practically useful results.

3.1. The data

The data analysed here consist of observations of the following three metocean variables measured from the Alwyn North platform, which is situated in the northern North Sea (60°48.5′N and 1°44.17′E, water depth approximately 130 m):

Hs:	Significant wave height (m) derived as $4 \times$ (variance of sea surface)
Tz:	Zero up-crossing wave period (s)
Ws:	Mean wind speed (m s^{-1}) measured at 103 m above MWL

Although wind direction data were also collected they were not incorporated into the present analysis.

The data were collected between September 1987 and April 1996 and consist of hourly observations over this period made by Paras Ltd and by Heriot-Watt University. The details of the data collection are described in Appendix B. For the reasons described in Appendix C, we have omitted from the analysis data collected during the months of May to August each year. However, the extreme events in which we are interested have negligible probability of occurring during these months. (See Section 2.3 for a
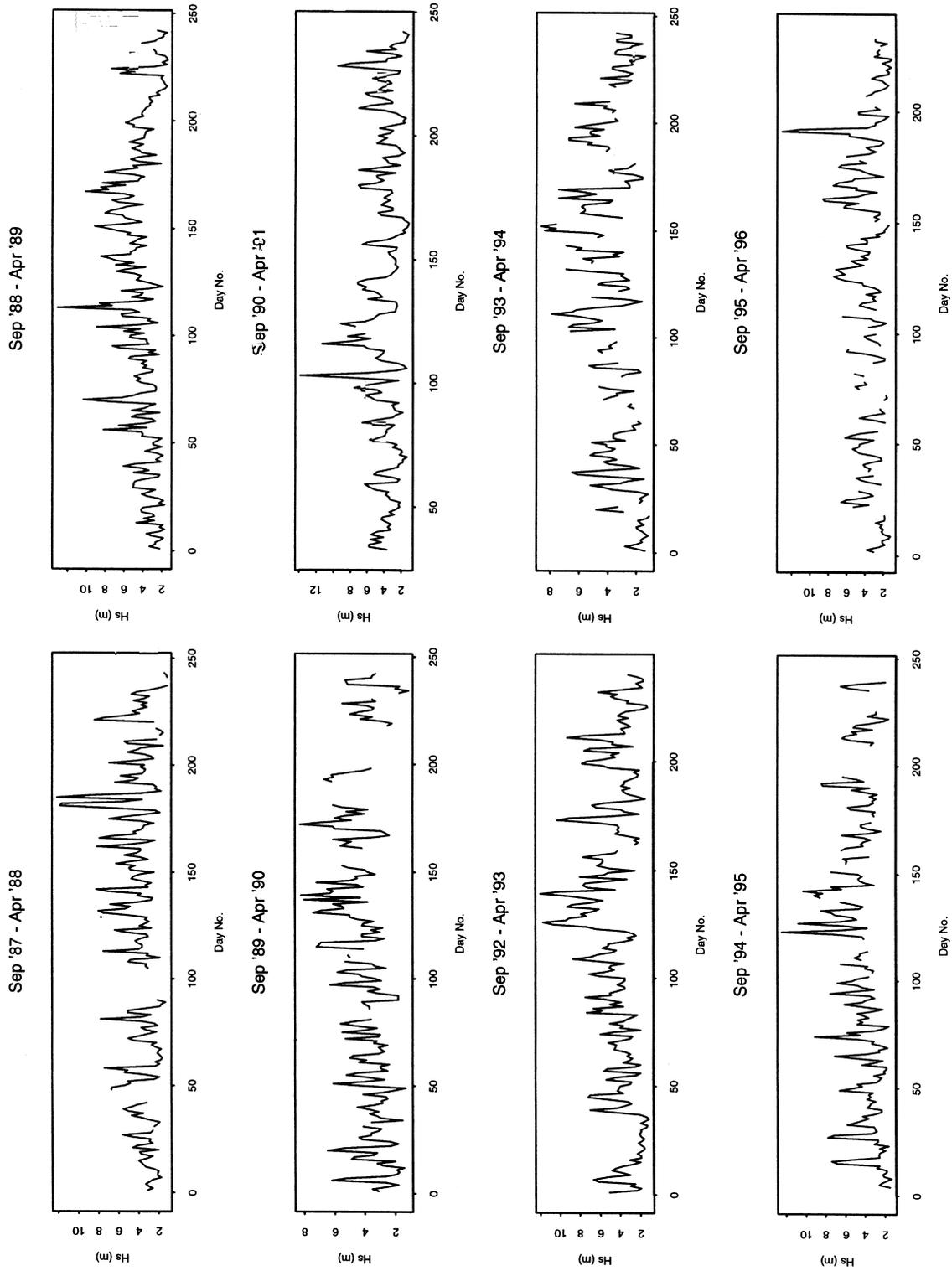
Fig. 1. Daily maxima of observations of wave height Hs.

discussion of the methodological issues here.) We have also omitted data collected during the interval September 1991 to April 1992, owing to the poor quality of the record for this period. Finally, there are occasional other smaller gaps in the observation record for one or more of the variables, because of individual sporadic sensor problems or other difficulties (not associated with the severity of the metocean conditions). Thus we have chosen for analysis a fairly complete record corresponding to eight 'winter' periods of eight months each.

Fig. 1 shows a plot of the daily maxima of the corresponding observations of wave height *Hs*. The observations of the remaining variables *Tz* and *Ws* exhibit similar patterns of variability, in particular seasonal dependence within each of the eight-month periods, and have similar patterns of missing observations. These occasional missing observations are sufficiently uniformly distributed over the seasons under study so as not to introduce any bias into the present analysis.

The two variables *Hs* and *Ws* are those of most interest and significance for structural loading and we concentrate on a analysis of their joint distribution, following the methodology of Section 2. We subsequently consider the variable *Tz* and its relation to each of the former variables.

3.2. Bivariate analysis of wave height and wind speed

3.2.1. Choice of period length and associated observations

The methodology presented in this paper assumes that the data for analysis consist of at least approximately independent observations. We have already argued in Section 2.3 that such seasonality as is present in the data does not present a serious problem. However, there is also very considerable short-term dependence in the hourly observations which constitute our primary data record. In particular extreme events tend to cluster in storms. We therefore seek to follow the approach of Section 2.4 and choose a length of period such that, if n is the number of such periods and $\mathbf{X}_i = (Hs_i, Ws_i)$ is the bivariate observation associated with each period i as described there, then $\mathbf{X}_1,...,\mathbf{X}_n$ can reasonably be treated as independent.

Previous studies of offshore data [11,20,26] have used period lengths varying from 24 to 60 h, and we investigate the suitability of period lengths of this order. The details of our analysis are given in Appendix D, and we suggest that,

for the present data, a suitable length of period is given by 48 h. We use this period length in the analysis below of both the marginal and joint distributions of *Hs* and *Ws*. However, the criticality of this choice is further investigated in the sensitivity analysis of Section 3.2.6.

For the present we take Hs_i and Ws_i to be the maximum observed values of *Hs* and *Ws*, respectively, during each period i. We discuss subsequently possible variations of these definitions.

3.2.2. Fitting marginal distributions

The most difficult statistical issue in fitting the marginal distribution of each variable X_j is the estimation of the threshold u_j above which the data are assumed to be well modelled by a GPD. Fig. 2 shows the mean excess plots described in Section 2.1 for each of the variables *Hs* and *Ws*. These are based on the observations $\mathbf{X}_i = (Hs_i, Ws_i)$, $i = 1,...,n$, associated with a 48-hour period length as described above. Fig. 3 shows the corresponding histograms of the distributions of the two variables.

We have already remarked that mean excess plots are quite difficult to interpret, as they are visually dominated by the few most extreme observations. However, both they and, more so, the histograms in Fig. 3 (see below) suggest that reasonable choices of threshold are 6.5 m for *Hs* and 16.5 m s^{-1} for *Ws*. We use these thresholds below, but in Section 3.2.6 we further investigate the criticality of these choices.

For each of the variables *Hs* and *Ws* the (marginal) distribution of the observations Hs_i or Ws_i is now estimated by using kernel density estimation for those observations below the threshold and maximum likelihood fitting of a GPD for those observations above it (see Section 2.1). The parameters (ξ, μ, σ) (as defined in Section 2.1.1 and representing shape, location and scale, respectively) associated with each of these fitted distributions are shown in Table 1. The threshold associated with each variable and the density of the associated fitted distribution are also shown in Fig. 3, the close fit to the observed data confirming the reasonableness of the choice of threshold. Note that, as is typical, there is a slight discontinuity at the threshold in the smoothed density. This is at a level which is well below the area of interest and does not present any difficulties for the extrapolation process.
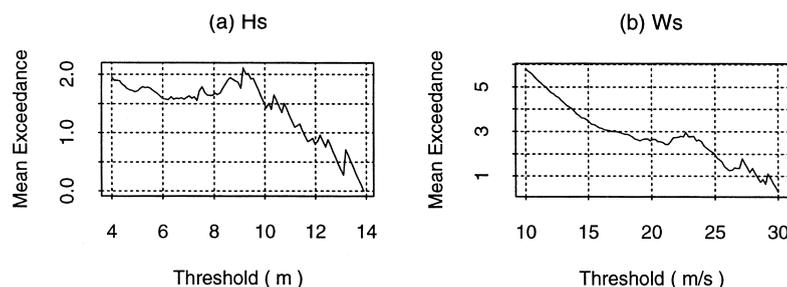
Fig. 2. Mean excess plots for *Hs* and *Ws* (48-hour period length).
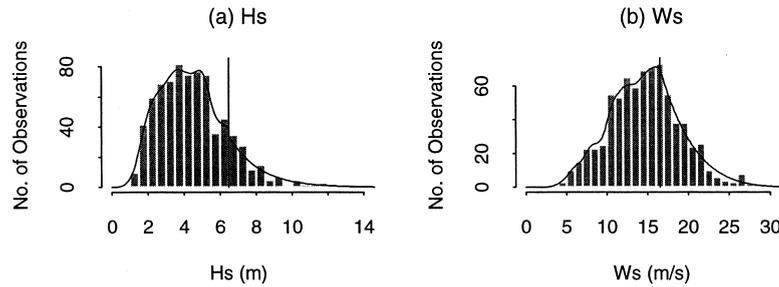
Fig. 3. Observed and fitted distributions of *Hs* and *Ws* (48-hour period length).

3.2.3. Calculation of return levels

For each of the variables *Hs* and *Ws*, the return levels associated with given return periods are determined as described in Section 2.1.3. Fig. 4 shows, for each of these variables, a plot of return level against return period. Each plot gives point estimates of the return level as the return period is increased, along with 95% confidence intervals calculated using profile likelihoods.

The confidence intervals indicate a sometimes considerable degree of uncertainty in these return levels. This is scarcely surprising since the quantity of data available as the basis for extrapolation corresponds to an interval of time whose length is less than one-tenth of that of the longest return periods considered. However, these confidence intervals are wider than those sometimes reported in other studies using comparable quantities of data [2,8,21]. There are two reasons for this: (a) no prior assumptions are made here regarding the tail shapes of the distributions of the variables beyond those which are implied by the asymptotic theory used (however, further assumptions which are compatible with this approach, if warranted, could be incorporated into the present methodology); and (b) some other studies apply results which are essentially asymptotic while using thresholds which are considerably below those used here, thereby increasing the quantity of data apparently available for assessments of uncertainty; this decreases the variance of parameter estimates at the expense of increasing their (unknown) bias. (Note, however, that consideration of water depth places a natural upper bound on *Hs* return levels. If this were to be incorporated into the current methodology the effect would be to truncate the set of attainable *Hs* values, rather than to alter the shape of the return level curve at lower return periods.)

3.2.4. Estimation of the joint distribution

We now consider estimation of the joint distribution of

Hs and *Ws*. We use the methodology described in Section 2.2 and based on (a) transformation of the individual variables so that each has a standard Fréchet distribution (Eq. (9)), followed by further transformation to pseudo-polar coordinates $(r, \mathbf{w})$ as defined by Eqs. (10) and (11), and (b) estimation of the asymptotic distribution μ of $\mathbf{w}$. The transformations described in (a) are straightforward (given the estimated marginal distributions of the variables), while the estimation in (b) is based on the observed distribution of $\mathbf{w}$ for those observations in the set N_{r_0} (see Section 2.2.2) for a suitably chosen threshold r_0.

Note that, since here $\mathbf{w} = (w_1, w_2)$ and, as already observed, $w_1 + w_2 = 1$ always, we may replace the vector $\mathbf{w}$ by the scalar quantity $w = w_1$. We take this to be the component of $\mathbf{w}$ associated with the variable *Hs*.

The threshold r_0 is the value of r above which the conditional distribution of w, given r, is reasonably stable. The choice of this threshold is perhaps best made [5] by the consideration of a range of possible values of r_0 and the construction, for each such value, of a histogram of the distribution of w for those observations in the set N_{r_0}. Fig. 5 shows these histograms for a range of increasing values of r_0. (As is customary, we consider values of r_0 of the form exp k for simple values of k.)

Here, for values of r_0 greater than or equal to exp (1.5), the distribution of w in the set N_{r_0} seems to be fairly stable until, for very high values of r_0, the number of observations in the set N_{r_0} is too small to be able to represent the underlying distribution at all reliably. We therefore take $r_0 =$ exp (1.5). However, the sensitivity of the results to this choice is further investigated in Section 3.2.6.

The asymptotic distribution of w on the interval [0,1] is now estimated from those observations in the set N_{r_0}. For the reasons explained in Section 2.2.2, we here take a non-parametric approach and use kernel density estimation [23]. As might be expected for a distribution with a finite range, it

Table 1
Bivariate analysis of *Hs* and *Ws*: summary of marginal analyses with $\mathbf{X}_i = (Hs_i, Ws_i)$

Variable	No. of observations	Threshold	No. of observations above threshold	Distribution parameters		
				ξ	μ	σ
Hs	762	6.5 m	119	−0.01	11.2	1.58
Ws	762	16.5 m s^{-1}	256	−0.15	26.6	2.03
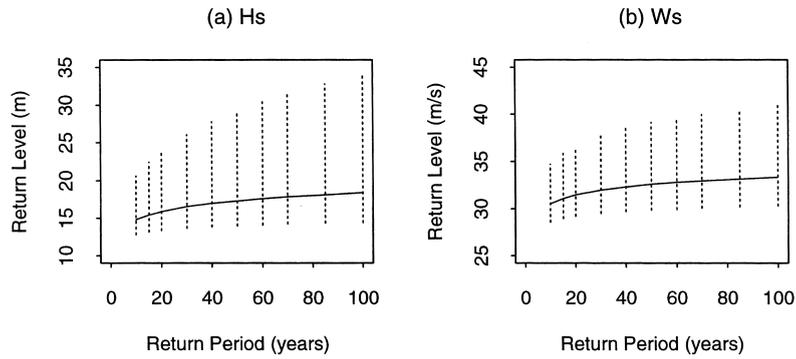
Fig. 4. Return level plots for *Hs* and *Ws* with 95% confidence intervals.

seems desirable to use a variable window width which becomes narrower close to the boundaries of the interval [0,1]. In particular this may be used to prevent the kernel density estimate from spilling beyond this interval. We achieve this by a logistic transformation $z = \log w - \log (1 - w)$ of w to the entire real line, followed by (fixed width) kernel density estimation of the distribution of z and then the inverse transformation back to the interval [0,1]. Of course the choice of kernel width is still necessarily somewhat subjective.

Fig. 6 shows a histogram of the distribution of w for those observations in the set N_{r_0} ($r_0 = \exp(1.5)$) together with its probability density estimated as above. Note the general shape of this density, which is very different from the U-shaped densities of Coles and Tawn [5] and corresponds to a very strong association between the extremes of the two variables. This is typical of wind and wave data.

Finally, we consider the estimation of the joint density function of the (untransformed) variables *Hs* and *Ws*. This may now be determined from the theory of Section 2.2. Consideration of the transformations (a) and (b) above shows that the joint density function f_X of $\mathbf{X} = (Hs, Ws)$ is

given by

$$f_{\mathbf{X}}(x_1, x_2) = \frac{2h(w)}{r^3}\psi_1'(x_1)\psi_2'(x_2) \qquad (14)$$

where h is the density function of w as estimated above, ψ_1' and ψ_2' are the derivatives of the functions ψ_1 and ψ_2 defined in Section 2.2.2 (and implementing the transformation (a) above), and the pair (r,w) is obtained from (x_1,x_2) via the transformations (a) and (b), i.e.

$$r = \psi_1(x_1) + \psi_2(x_2) \text{ and } w = \psi_1(x_1)/r$$

A proof of this result is given in Appendix A. Note that the above representation of the joint density function depends both on the fact that here $\mathbf{X}$ is bivariate ($d = 2$) and on the chosen parametrisation of the space S_d (by $w = w_1 = 1 - w_2$).

Fig. 7(a) shows geometrically spaced contours of the joint density superimposed on a plot of the observations (Hs_i, Ws_i). The horizontal and vertical lines correspond to the thresholds used in fitting the marginal distributions of the variables. The remaining, curved, line corresponds to the threshold r_0 used in the bivariate analysis. The area of
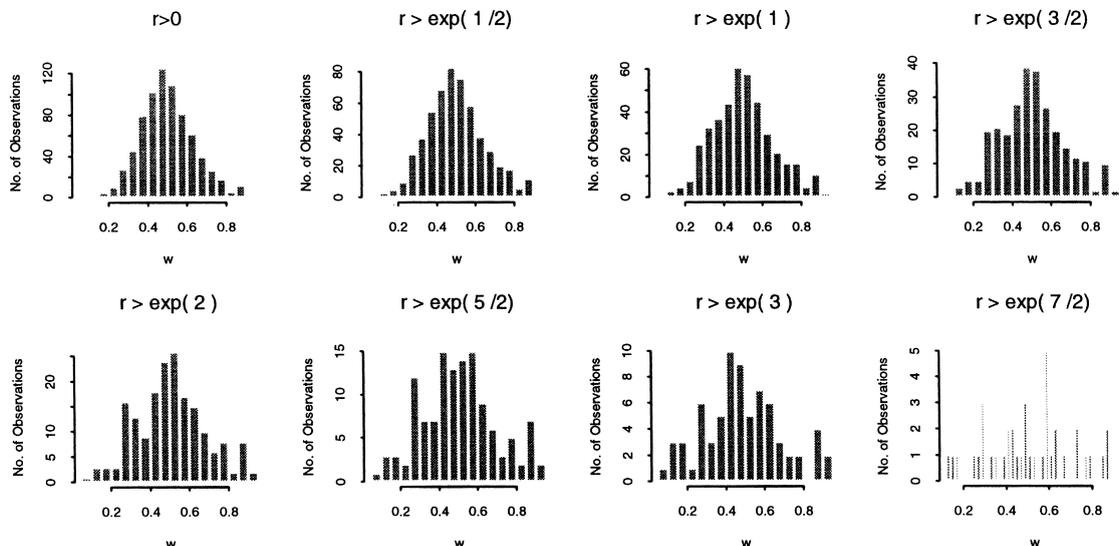
Fig. 5. Bivariate analysis of *Hs* and *Ws*: histograms of w for increasing r_0.
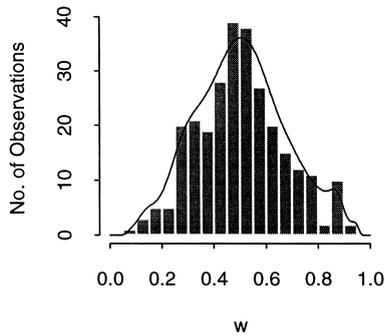
Fig. 6. Bivariate analysis of *Hs* and *Ws*: distribution of *w* in the set N_{r_0} and estimated probability density.

interest, corresponding to the region $r > r_0$ in (r,w) space, lies to the right of and above this curve and contours of the density are shown only in this region. (Recall that it is in this region that the association between the two variables, as measured by *w*, is assumed to have stabilised, and that it is the observations in this region which are used to estimate this association.) Observe that the estimated density function is compatible with the observed data in the region where the latter exists, as would be expected, and provides a realistic extrapolation of the density into the region where observations do not exist.

3.2.5. Alternative observation definitions

A traditional analysis of the most extreme wave height and wind speed combination likely to occur with, for example, a 100-year return period would simply combine the individual 100-year return levels for these two variables. In general this is excessively conservative as it is most unlikely that these two return levels will be achieved at precisely the same time. By focusing on the joint distribution of the two variables, the present bivariate analysis attempts to address this problem. However, our current definition of the bivariate observation $\mathbf{X}_i$ to be associated with

each 48-hour period *i* still couples the maximum wave height and maximum wind speed within this period and thus retains some of the conservatism inherent in the more traditional approach.

We now consider two other possible definitions of $\mathbf{X}_i$, as discussed in Section 2.4:

$\mathbf{X}_i = (Hs_i, Ws_i')$ where, as previously, Hs_i is the maximum observation of the variable *Hs* within the period *i* and where Ws_i' is the *concomitant* observation of the variable *Ws*;

$\mathbf{X}_i = (Hs_i', Ws_i)$ where Ws_i is the maximum observation of the variable *Ws* within the period *i* and where Hs_i' is the *concomitant* observation of the variable *Hs*.

These two definitions are appropriate for applications to structures for which the loadings are wave- and wind-dominated, respectively. (The remaining possibility discussed in Section 2.4 is structure-specific and therefore not pursued here.) Tables 2 and 3 summarise the marginal analyses corresponding to these two further definitions, and are analogous to Table 1 for our earlier definition. Note that, for each of the three definitions of the bivariate observations $\mathbf{X}_i$, there is a slightly different number of these observations available for analysis. This is accounted for by the locations of the occasional gaps in the original series of hourly observations. Note also that the observations Hs_i' exceeds their threshold of 6.5 m on many fewer occasions than do the observations Hs_i; this is as would be expected from their concomitant definition. A similar remark applies to the threshold exceedances of Ws_i' in relation to those of Ws_i.

Fig. 7(b) and (c) show the corresponding joint density functions of the variables *Hs* and *Ws*. These figures are based on similar bivariate analyses to that leading to Fig. 7(a), again with $r_0 = \exp(1.5)$ in each case, and are interpreted similarly. We discuss the significance of these alternative analyses in Section 3.4.

Table 2
Bivariate analysis of *Hs* and *Ws*: summary of marginal analyses with $\mathbf{X}_i = (Hs_i, Ws_i')$

Variable	No. of observations	Threshold	No. of observations above threshold	Distribution parameters		
				ξ	μ	σ
Hs	792	6.5 m	121	0.00	11.2	1.59
Ws	792	16.5 m s^{-1}	155	−0.15	24.3	1.92

Table 3
Bivariate analysis of *Hs* and *Ws*: summary of marginal analyses with $\mathbf{X}_i = (Hs_i', Ws_i)$

Variable	No. of observations	Threshold	No. of observations above threshold	Distribution parameters		
				ξ	μ	σ
Hs	777	6.5 m	63	0.01	10.0	1.55
Ws	777	16.5 m s^{-1}	257	−0.16	26.6	1.99

(a) Hs and Ws

(b) Hs and Ws'

(c) Hs' and Ws
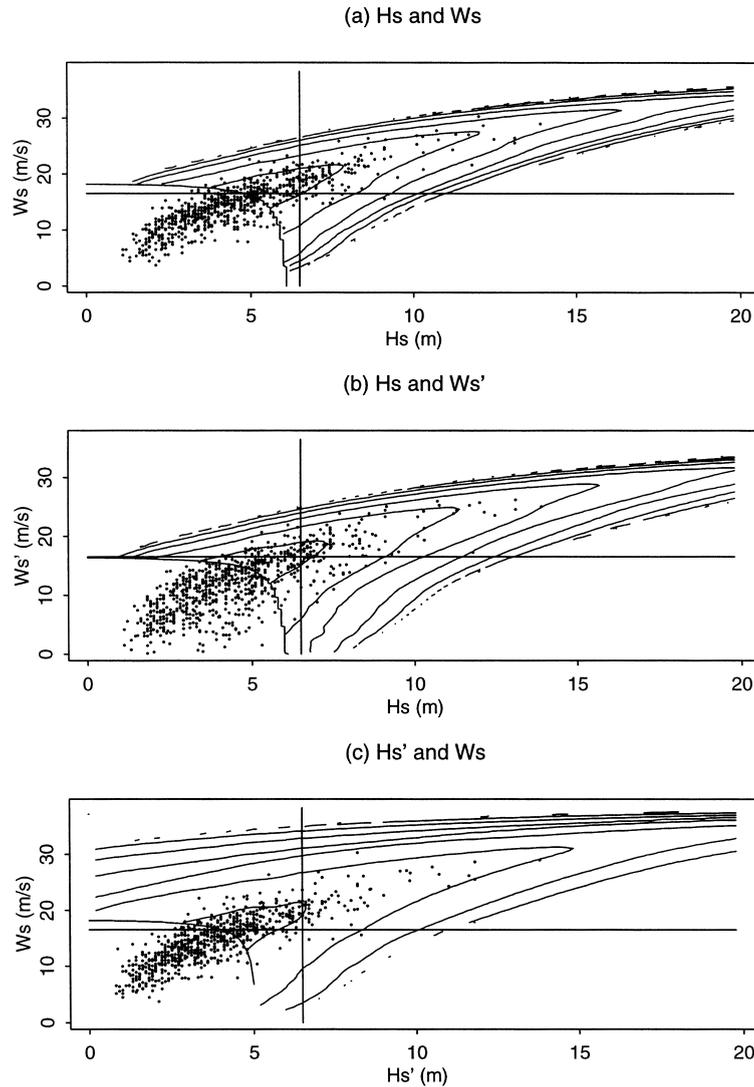
Fig. 7. Joint density of wave height and wind speed based on (a) $\mathbf{X}_i = (Hs_i, Ws_i)$, (b) $\mathbf{X}_i = (Hs_i, Ws_i')$, (c) $\mathbf{X}_i = (Hs_i', Ws_i)$.

3.2.6. Sensitivity analysis

We examine further the sensitivity of the preceding analysis to the choice of period length (Section 3.2.1), marginal thresholds (Section 3.2.2), and bivariate threshold r_0 (Section 3.2.4). Rather than attempt a general description of the sensitivity to these quantities of the entire joint distribution of Hs and Ws (through, for example, the estimated

Table 4
Sensitivity analysis: 100-year return levels for Hs (m) and associated values of Ws (m s^{-1})

Thresholds	Period length					
	36 h		48 h		60 h	
	Hs	Ws	Hs	Ws	Hs	Ws
6 m, 15 m s^{-1}	18.17	31.40	18.28	31.77	17.71	31.06
6.5 m, 16.5 m s^{-1}	17.78	32.59	18.36	32.98	17.85	31.86
7 m, 18 m s^{-1}	19.70	32.81	17.69	32.56	19.01	33.34

parameters of the marginal distributions and estimated distribution of w—all of which are difficult to interpret), we focus instead on 100-year return levels as described below.

Table 4 shows estimates of the 100-year return level for wave height Hs together with, for each such estimate, the value of the wind speed Ws most likely to be associated with it (see Section 3.4.2). This latter quantity is determined from the estimated joint distribution of the two variables. For this analysis we again consider the case $\mathbf{X}_i = (Hs_i, Ws_i)$, so that the observation of each variable associated with the period i is its maximum during that period. The estimates are based on three choices of period length combined with each of three choices of marginal threshold pair. Recall that our main analysis corresponds to the use of a 48 h period length and marginal thresholds of 6.5 m and 16.5 m s^{-1} for Hs and Ws, respectively. Similarly, Table 5 shows analogous estimates of the 100-year return level for wind speed Ws together with the most likely associated values of the

Table 5
Sensitivity analysis: 100-year return levels for Ws (m s^{-1}) and associated values of Hs (m)

Thresholds	Period length					
	36 h		48 h		60 h	
	Ws	Hs	Ws	Hs	Ws	Hs
6 m, 15 m s^{-1}	31.56	18.20	32.04	18.20	31.10	17.40
6.5 m, 16.5 m s^{-1}	32.88	17.60	33.32	18.80	32.11	17.60
7 m, 18 m s^{-1}	32.86	19.60	32.77	18.00	33.48	18.80

wave height Hs. The bivariate analysis is again based on defining $\mathbf{X}_i = (Hs_i, Ws_i)$.

The results in Tables 4 and 5 show that these estimates are not especially sensitive to modest variations in either the choice of period length or that of marginal threshold. In particular the variation in the 100-year return levels of the two variables is substantially less than the width of the confidence intervals for our main estimates of these quantities (see Section 3.2.3).

Note also from Tables 4 and 5 that the value of Ws most likely to be associated with the 100-year return level of Hs is, for these analyses, quite close to the 100-year return level of Ws itself. A similar remark holds when the roles of the two variables are interchanged. (See Section 3.2.5 for some discussion of this.) This is relevant when we consider the effect of varying the threshold r_0 associated with the bivariate analysis. Clearly this does not affect the estimated 100-year return level associated with either variable, which is based only on the marginal analysis of that variable, but in each case it may affect the most likely associated value of the other variable. In Section 3.2.4 we take $r_0 = \exp(1.5)$. With the period length and marginal thresholds used in that section, variation of r_0 to $\exp(1.25)$ and to $\exp(1.75)$ has no effect on these associated values. Given the preceding remarks, this is perhaps not surprising for the present data.

Finally we remark briefly on the sensitivity of standard

errors to variation in any of the quantities considered here. This is as might be expected: to a very good approximation standard errors are inversely proportional to the square root of the number of observations available for their estimation. In particular, for a return level, this is the number of observations—of Hs or Ws—above the marginal threshold for the variable concerned.

3.3. Incorporation of wave period

In this section we consider, rather more briefly, the incorporation of the wave period Tz into the preceding analysis. In many applications one or other of the variables wave height Hs and wind speed Ws will be the main determinant of structural loading, and may be thought of as the primary variable of interest. We will then be further interested in values of the remaining two variables likely to be associated with, for example, given return levels of the primary variable. For this reason we consider here bivariate analyses of (a) wave height and wave period, (b) wind speed and wave period. This is in preference to a full trivariate analysis of the three variables which would seem unlikely to produce significant additional insight. We continue to use a 48-hour period length and consider the bivariate observations associated with the individual periods as independent. For each such period i, and for the analysis (a) above, we define the associated bivariate observation (Hs_i, Tz_i) of (Hs, Tz) by again taking Hs_i to be the maximum value of Hs observed during the period, and by taking Tz_i to be the concomitant value of Tz. Similarly, for each period i, and for the analysis (b), we define the bivariate observation (Ws_i, Tz_i') of (Ws, Tz) by taking Ws_i to be the maximum value of Ws observed during the period and Tz_i' to be the value of Tz observed concomitantly with Ws_i. These definitions are those appropriate when wave height and wind speed are the respective primary variables.

The fitted marginal distribution of each of the primary variables Hs and Ws remains that obtained in

Table 6
Bivariate analysis of Hs and Tz: summary of marginal analyses

Variable	No. of observations	Threshold	No. of observations above threshold	Distribution parameters		
				ξ	μ	σ
Hs	762	6.5 m	119	−0.01	11.2	1.58
Tz	762	7.5 s	173	−0.13	10.1	0.63

Table 7
Bivariate analysis of Ws and Tz: summary of marginal analyses

Variable	No. of observations	Threshold	No. of observations above threshold	Distribution parameters		
				ξ	μ	σ
Ws	777	16.5 m s^{-1}	257	−0.16	26.6	1.99
Tz	777	7.5 s	82	−0.07	9.7	0.80
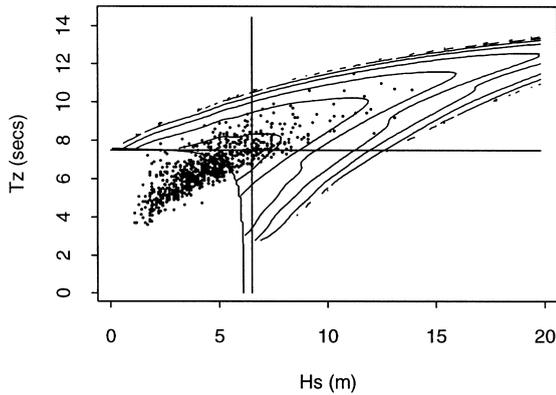
Fig. 8. Joint density of wave height *Hs* and wave period *Tz*.

Section 3.2.2. For each of the analyses (a) and (b) the marginal distribution of *Tz* is fitted as usual. The main issue here is again the choice of threshold above which its distribution is modelled by a GPD. Examination of mean excess plots and histograms suggested that in both cases a reasonable choice of threshold appears to be 7.5 s and we use this threshold here.

Tables 6 and 7 give summary information and parameters for the fitted marginal distributions associated with these two bivariate analyses. Note that in the analysis (b) of wind speed and wave period there are relatively few threshold exceedances by the observations of the latter variable. The reason is that Tz_i' is the observation of Tz concomitant with the maximum observation of *Ws* during each period *i*, and is likely to be far from the maximum period *i* observation of the variable Tz itself.

For each of the two analyses, the joint distribution of the bivariate pair is estimated analogously to that of the variables *Hs* and *Ws* in Section 3.2.4. In each case a suitable value of the threshold r_0 is again given by $r_0 = \exp(1.5)$. Figs 8 and 9 show the resulting estimated joint density functions. These are interpreted similarly to those in Fig. 7.

3.4. Interpretation and application of results

The preceding analyses have many applications, ranging
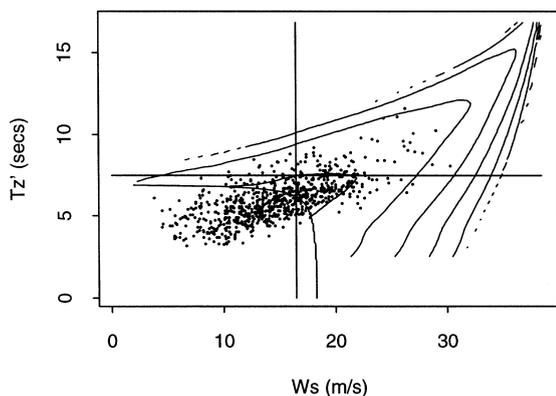
Fig. 9. Joint density of wind speed *Ws* and wave period *Tz*.

from the calculation of return levels for individual variables to the determination of likely combinations of extreme events and the estimation of their probabilities. These latter determinations depend on the joint distributions which are illustrated in Figs 7–9. They are useful for design engineers and offshore operators who require answers to specific questions about loadings on structures. They are also of considerable interest for oceanographers and others who require a greater understanding of the metocean climate at any particular site. We give below some qualitative and quantitative conclusions.

3.4.1. Associations between variables

The joint density function plots in Fig. 7 show the estimated joint distribution of wave height and wind speed. This distribution shows very considerable association between the two variables. In particular, this association continues to be present in the region where either of these variables is extreme and appears to extend well beyond the range of the observed data.

Note that, while the overall form of the joint distribution does not depend greatly on the precise definition of the bivariate observations $\mathbf{X}_i$ ($\mathbf{X}_i = (Hs_i, Ws_i)$, $\mathbf{X}_i = (Hs_i, Ws_i')$ or $\mathbf{X}_i = (Hs_i', Ws_i)$ as defined earlier), the association between the two variables is nevertheless greatest in the case $\mathbf{X}_i = (Hs_i, Ws_i)$, where the period *i* observation of each variable is its maximum during that period. This is illustrated in Fig. 7(a) and is in contrast to the lower associations in the cases of the alternative 'concomitant' definitions leading to Fig. 7(b) and (c). The likely reason for this is that the wave height at any given time is a function not just of the current wind speed, but of the past history of this process over a variable time period (depending on such factors as wind direction and rate of increase and decrease of wind speed). The statistical association between the observed values of the two variables is thus greatest when the definition of these observations allows for a (generally variable) time lag between wind speed and resulting wave height.

Similarly, the joint distribution of wave height and wave period illustrated in Fig. 8 again shows considerable association between these two variables, and in particular between their extremes. Very similar remarks again apply to the joint distribution of wind speed and wave period illustrated in Fig. 9.

3.4.2. Return levels and their concomitants

An analysis of the return levels for each of the primary variables *Hs* and *Ws* is given in Section 3.2.3. As further comment, note that the data provide no indication of an upper bound on the wave height. The best estimate of the shape parameter ξ_{Hs} is close to zero (corresponding to an exponential distribution of wave height) but its standard error is such that *Hs* might either have an upper bound ($\xi_{Hs} < 0$) or be unbounded ($\xi_{Hs} \geq 0$). Of course physical considerations imply that there does in reality exist some
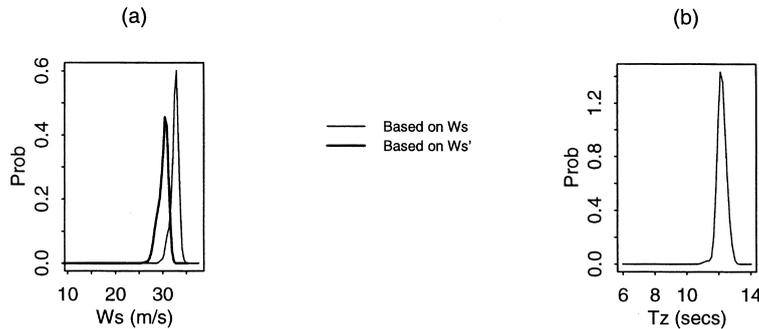
Fig. 10. Conditional densities of (a) wind speed; (b) wave period, for 100-year wave height.

such upper bound, but this lies beyond the point to which the distribution of the variable may reliably be extrapolated. The data do provide a fairly clear indication of an upper bound on the wind speed—here the standard error of the shape parameter ξ_{Ws} is such that its 95% confidence interval lies entirely within the region for which $\xi_{Ws} < 0$.

For any given return period, it is of practical importance to be able to determine the corresponding return level of either of these two variables, together with likely associated values of the other variable and of the wave period Tz. For example, the most recent design code requirements for calculating loads on offshore structures in the North Sea [24] give a choice of three possible design conditions which the structure must be able to withstand:

1. a 100-year return period wave height with 'associated' wave period, wind and current;
2. any 'reasonable' combination of wind speed, wave height, and current speed that results in the 100-year platform load;
3. the 100-year wave height combined with the 100-year wind speed and the 100-year current speed.

We have already observed in Section 1 that the second of these possibilities has the drawbacks associated with the structure variable approach, while the third is unduly conservative. The present methodology can readily be applied to analyses such as that required by the first of these options. Here, for any given value of the wave height Hs, for example its 100-year return level, we may easily calculate the corresponding *conditional* distribution of the wind speed

Ws by using the joint distribution of these two variables. The mode, or other suitable location measure, of this conditional distribution then defines a most likely value of Ws associated with the given value of Hs. Similarly, we may determine the corresponding associated value of the wave period Tz by consideration of its joint distribution with Hs.

Fig. 10(a) shows, as a density, the conditional distribution of wind speed associated with the estimated 100-year wave height of 18.4 m. This is determined from the estimated joint distribution of the two variables. As usual, for the calculation of the return level for wave height, we take the observations Hs_i of this variable to be the maximum observed values of Hs during successive 48-hour periods. The two curves shown in Fig. 10(a) derive from the two possible definitions Ws_i and Ws_i' of the observations of Ws to be associated with each period i. For the former definition (in which Ws_i is the maximum wind speed in each period i) the most likely wind speed associated with the 100-year wave height is 33.0 m s^{-1}. For the latter definition (in which Ws_i' is the observation of wind speed observed concomitantly with Hs_i in each period i) the wind speed associated with the 100-year wave height is 30.5 m s^{-1}. It is clear that the latter definition seems more appropriate for applications (see the discussion of Section 3.2.5). Note also, from Section 3.2.3, that the independently determined 100-year wind speed is 33.3 m s^{-1}. These results are very compatible with the general understanding that the 100-year wave height and 100-year wind speed may well occur within the same storm, but would not be expected to occur at the same time.
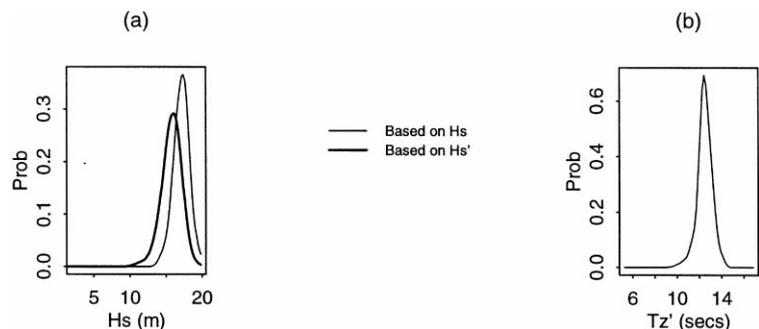
Fig. 11. Conditional densities of (a) wave height; (b) wave period, for 100-year wind speed.

Similarly, Fig. 10(b) shows, again as a density, the conditional distribution of the wave period Tz associated with the estimated 100-year wave height. Here, for each period i, the associated observation Tz_i is is taken to be that observed concomitantly with Hs_i. The most likely value of the wave period associated with the 100-year wave height is 12.1 s, resulting in an estimated significant steepness of the sea state of $2\pi Hs/gTz^2 = 1/12.5$. This is marginally more severe than the steepest sea states of about 1/13 usually observed in the North Sea [25].

For applications to structures for which the loading is wind dominated, a similar analysis may be performed to determine the distributions of wave height and wave period conditional on the estimated 100-year wind speed of 33.3 m s^{-1}, as noted above. Fig. 11(a) shows the resulting probability density of the wave height Hs, the two curves here corresponding to the two possible definitions, Hs_i and $Hs_i{}'$ of the observation of Hs to be associated with each period i. The most likely value of the wave height associated with the 100-year wind speed is 18.8 m for the analysis based on the use of Hs_i (the maximum observed value of Hs during period i) and 17.0 m for the analysis based on the use of $Hs_i{}'$ (the observation of Hs made concomitantly with Ws_i in each period i). Again, the latter analysis seems the more appropriate.

Fig. 11(b) shows the corresponding conditional probability density for the wave period Tz—here based on the use of the associated observations $Tz_i{}'$. The most likely value of the wave period associated with the 100-year wind speed is 12.4 s. This figure is very similar to the wave period associated with the 100-year wave height.

4. Conclusions

4.1. General conclusions

The methodology considered here for the analysis of multivariate extremes has two objectives: (a) the general estimation of the joint distribution of the extremes of the variables concerned; and (b) the estimation of the probabilities of particular extreme events, notably those corresponding to extreme loadings on given structures.

The first of these objectives is here successfully achieved. Indeed the methodology used makes no prior assumptions about the joint distribution of the variables and only requires that there is sufficient data for the applicability of asymptotic results. In particular, no modelling assumptions are made about the statistical association between the extremes of the variables. Rather this is estimated and extrapolated directly from the association observed in the extremes of the data. For the bivariate analyses of the metocean data considered in Section 3, the estimated joint distributions are shown in the density plots of Figs 7–9.

The second of the above objectives is usually achieved in applications through the determination of events corresponding to a given return period—for example 100 years. For multivariate data, there will be a wide choice of such events. Thus, for a given structure, one possibility will be to choose that event which is most extreme in terms of a suitable measure of the structural loading. In the application of Section 3, which is not structure-specific, we determine instead the corresponding return level for a specified variable (wave height Hs or wind speed Ws) together with the most likely associated values of the remaining variables. This procedure is in accordance with the most recent design code requirements, and is distinctly less conservative than the remaining possibility of simply combining the individually determined return levels of each of the variables. We return to this point in Section 4.3 below.

4.2. Methodological issues

4.2.1. Seasonality

The metocean data studied in this paper are of course highly seasonal, with extreme observations occurring almost entirely during winter months. However, this does not appear to result in serious problems for their analysis. Rather, what is modelled here is the distribution of the data averaged over the year, and this is entirely appropriate for the determination of, for example, 100-year return periods. While in principle there might be some small efficiency gain to be had by splitting the data according to, for example, month, this approach would suffer from the disadvantages discussed in Section 2.3.

Note that in the application of Section 3, we omit entirely data collected during the months of May to August (see Section 3.1). As indicated in Appendix C, the occurrence of extreme events in this period is quite negligible, and analysis may safely be based on the data collected in the remaining months, with the effective assumption of an eight-month year for the calculation of return levels.

4.2.2. Short-term dependence

The existence of short-term dependence in time-series data such as that considered here is a matter which requires very considerable care. As discussed in Section 2.4, the present methodology is based on the assumption that the data to be analysed consist of independent observations of the given variables. The analysis of dependent observations as if they were independent leads to (a) over-optimistic assessments of uncertainty for estimated quantities and (b) generally conservative estimates of return levels. We describe in Section 2.4, and further in Appendix D, how to identify blocks or periods of time such that the (multivariate) observations associated with successive periods may reasonably be regarded as independent. An alternative for univariate data is the explicit time-series modelling of the short-term dependence structure in the original data. It is

not clear how this might extend to the present multivariate methodology, but in any case the investigations referred to above suggest that the current approach is probably close to optimal for the data considered here.

4.2.3. Choice of marginal thresholds

The choice, for each variable, of a threshold above which its marginal distribution is well modelled by a GPD again requires care, and there is necessarily some subjectivity in the final determination of these thresholds. This reflects the general difficulty inherent in extrapolation. The issue is examined in detail in Section 2.1.2 and, for the metocean data of this paper, in Section 3.2.2. The further sensitivity analysis of Section 3.2.6 suggests that the marginal thresholds we use for these data are reasonable.

4.2.4. Estimation of joint distributions

The two issues involved in the estimation of joint distributions are (a) the choice of threshold r_0 such that the limiting distribution of the transformed variable **w** may reasonably be estimated from its distribution over those observations in the set N_{r_0}, and (b) the details of this estimation procedure. These issues are discussed in Section 2.2.2. For the present metocean data, they are further considered in Section 3.2.4 and Section 3.2.6.

Here the results appear relatively insensitive to the choice of r_0, except only of course that, as this threshold is increased, the quantity of data available for reliable estimation decreases.

For our metocean data and for the reasons explained in Section 2.2.2, we take a non-parametric approach to the estimation of the (limiting) distribution of **w**. Hence, it is difficult to make a *formal* assessment of the uncertainty involved in this estimation. However, for such a non-parametric approach, one issue which again requires care is that of the estimation of **w** at its boundaries. This is discussed in Section 3.2.4. Beyond this, there do not appear to be any major difficulties or uncertainties in the estimation of the distribution of **w**. See, for example, Fig. 6, which shows its observed and fitted distributions (for the chosen set N_{r_0}).

4.3. Implications for metocean design parameters

Estimation of extreme loadings on the basis of extrapolation from observed data requires (a) reliable extrapolation of the estimated distributions of the individual variables into their extreme regions, and (b) reliable estimation of the statistical association between these variables, again in their extreme regions.

The present paper addresses both these issues, but its more novel aspects are concerned with the second of them, where we develop and apply the methodology of Coles and Tawn [4,5]. This is the only approach to the estimation of association between the extremes of variables

which does not make prior assumptions about the nature of such association, but simply requires that there is available sufficient data for the reasonable applicability of asymptotic results (as described in Section 2.2.1).

More traditional methods of calculating extreme loading—for example, the third of the possibilities discussed in Section 3.4.2—typically do not attempt to estimate the association between the variables concerned. Rather they assume a worst-case scenario, for example by assuming that the 100-year wave height will always occur at just the same time as the 100-year wind speed. The greater precision of the present approach to the description of the metocean climate therefore makes it possible to remove some of the conservatism inherent in the design process and operating criteria for offshore structures. This has considerable potential economic benefits.

As we remark above, we also consider estimation and extrapolation of the marginal distributions of the individual variables. This is a necessary first step to the determination of the association between their extremes, but it is also the source of much of the inevitable uncertainty in the calculation of extreme loadings. We have again avoided making prior assumptions about these distributions and appealed only to asymptotic extreme value theory, leading to the use of a GPD for fitting their tails. At least for the data considered here, this approach reveals the large degree of uncertainty inherent in the estimation of 100-year return levels (and associated values of other variables) on the basis of extrapolation from less than 10 years of data. Where there is no sound theoretical argument to indicate that some prior and more specific tail shape is likely, this uncertainty should be properly reflected in the design process itself.

Appendix A. Proofs of results

We here give proofs of various results in the present paper. These are (a) the asymptotic results concerning the joint distribution of $(r,\mathbf{w})$ which are stated in Section 2.2.1, and (b) the determination, in Section 3.2.4 and *for a bivariate analysis*, of the joint density (Eq. (14)) of the extremes of the original variables $\mathbf{X} = (X_1, X_2)$, following the estimation of the limiting density function h of $w = w_1$.

The distribution function F of the standard Fréchet distribution (given by Eq. (9)) satisfies

$$F(x) = 1 - \frac{1}{x} + o\left(\frac{1}{x}\right) \text{ as } x \to \infty \tag{15}$$

In order that this result should hold the conditional distribution $\hat{F}_{\mathbf{w}}$ of r given **w** must similarly converge to 1 at a rate which is comparable to $1/r$ as $r \to \infty$ (of course the exact conditional distribution will depend on **w**). That is, under

very mild regularity conditions, we may write

$$\hat{F}_{\mathbf{w}}(r) = 1 - a(\mathbf{w})\frac{1}{r} + o\left(\frac{1}{r}\right) \text{ as } r \to \infty \tag{16}$$

for some non-negative function a on S_d. Now let the probability measure λ denote the marginal distribution of $\mathbf{w}$ on S_d (so that λ and $\hat{F}_{\mathbf{w}}$, $\mathbf{w} \in S_d$, together determine the joint distribution of $(r,\mathbf{w})$). Then, for each $j = 1,...,d$, the marginal distribution F of each of the random variables $\tilde{X}_j$ is given by

$$F(x) = \int_{S_d} \hat{F}_{\mathbf{w}}\left(\frac{x}{w_j}\right) d\lambda(\mathbf{w}) = 1 - \frac{1}{x}\int_{S_d} w_j a(\mathbf{w}) d\lambda(\mathbf{w})$$
$$+ o\left(\frac{1}{x}\right) \text{ as } x \to \infty \tag{17}$$

where the second equality in the above expression follows from Eq. (16) and on recalling that λ is a probability measure. By comparing Eqs. (15) and (17) it then follows that

$$\int_{S_d} w_j a(\mathbf{w}) d\lambda(\mathbf{w}) = 1, \ j = 1, ..., d \tag{18}$$

By summing Eq. (18) over $j = 1,...,d$ and recalling that $\sum_{j=1}^{d} w_j = 1$ for all $\mathbf{w} \in S_d$, we also obtain

$$\int_{S_d} a(\mathbf{w}) d\lambda(\mathbf{w}) = d \tag{19}$$

It further follows from Eq. (16) that, for large r, the conditional density of r given $\mathbf{w}$ is $a(\mathbf{w})/r^2$ (plus a term which is negligible in comparison to $1/r^2$ as $r \to \infty$). Since also the probability measure λ gives the marginal distribution of $\mathbf{w}$, the conditional distribution of $\mathbf{w}$ given r is defined by a probability measure μ, where

$$d\mu(\mathbf{w}) = ka(\mathbf{w})d\lambda(\mathbf{w}) \tag{20}$$

Here k is defined by the requirement that μ should be a probability measure ($\int_{S_d} d\mu(\mathbf{w}) = 1$) and so is independent of r; indeed consideration of Eq. (19) shows that $k = 1/d$. As $r \to \infty$ the conditional distribution of $\mathbf{w}$ given r converges to exactly the probability measure μ as required.

Similarly, from Eqs. (16) and (19) (with $k = 1/d$), the marginal distribution function $\hat{F}$ of r satisfies

$$\hat{F}(r) = 1 - \frac{1}{r}\int_{S_d} a(\mathbf{w})d\lambda(\mathbf{w}) + o\left(\frac{1}{r}\right) = 1 - \frac{d}{r} + o\left(\frac{1}{r}\right)$$

as $r \to \infty$, so that the result of Eq. (12) follows. Finally the result of Eq. (13) follows immediately from Eqs. (18) and (20).

We now consider the derivation of the joint density, Eq. (14), in Section 3.2.4. It follows from Eq. (12) that, in the bivariate case $d = 2$ and with $w = w_1$, the estimated joint density of the extremes of (r,w) (corresponding to large r) is

given by $2h(w)/r^2$. The Jacobian appropriate to the transformation of joint density of (r,w) to that of the (vector) variable $\tilde{\mathbf{X}}$ (i.e. to the inversion of the transformation (b) of Section 3.2.4) is given by

$$\begin{vmatrix} \dfrac{\partial r}{\partial \tilde{X}_1} & \dfrac{\partial r}{\partial \tilde{X}_2} \\ \dfrac{\partial w}{\partial \tilde{X}_1} & \dfrac{\partial w}{\partial \tilde{X}_2} \end{vmatrix} = \frac{1}{r}$$

The Jacobian appropriate to the transformation of joint density of $\tilde{\mathbf{X}}$ to that of the original variable $\mathbf{X}$ is simply the product $\psi_1'(X_1)\psi_2'(X_2)$—where ψ_1' and ψ_2' are as in Section 3.2.4—since, in the inversion of the transformation (a) of that section, each margin transforms separately. The result of Eq. (14) is now immediate.

Appendix B. The Alwyn North data

The metocean data analysed in the present paper are based on observations made at Alwyn North during the period August 1987 to September 1996. They consist largely of measurements made by Paras Ltd (at a sampling rate of 2 Hz) and supplemented by data recorded by Heriot-Watt University (the HW data—recorded at 5 Hz). The Paras data consist of series of hourly observations from August 1987 to December 1994 inclusive, while the HW data are based on hourly observations from August 1994 to April 1996 inclusive. (For the HW data, observations are available for every 20-minute interval during the period concerned, but for consistency we only consider those from every third interval.) While both data sets come from essentially the same instrumentation, there were some differences in the sampling regime and statistical handling of the raw data. The overlap period of August to December 1994 permitted a direct comparison between the two sets and enabled an assessment of the validity of combining them.

For the Paras data the observations of mean wind speed Ws are calculated as 10 min averages every hour, while averages of 20 min are used for the HW data. A comparison of data from the two sources during the overlap period shows that no correction is necessary in order to combine the two sets of observations of this variable. However, *all* wind speed observations have been multiplied by 0.774 so that they correspond to the internationally accepted reference height for wind speed of 10 m above mean water level. The formula used here is $Ws_{10} = Ws_h(10/h)^e$ where h is the height of the anemometer and e the empirically determined exponent of 0.11 [9]. (Of course, for the analysis of this paper, this adjustment is merely a scale change for Ws.)

Observations of the significant wave height Hs and the mean wave period Tz are calculated over a 20-minute period in every hour, for both the Paras and the HW data. Both statistics are calculated from their time domain definitions.

Table 8
Distributions of *Hs* and *Ws* during 'summer' and 'winter' periods

Hs (m)	Summer	Winter	*Ws* (m s^{-1})	Summer	Winter
0–1	5162	1186	0–2	1724	1482
1–2	11 576	10 029	2–4	3575	4193
2–3	3710	12 746	4–6	4602	5846
3–4	634	9386	6–8	4508	6239
4–5	158	5103	8–10	3685	6633
5–6	21	2360	10–12	2189	6139
6–7	7	1142	12–14	728	4886
7–8	7	417	14–16	198	3508
8–9	3	180	16–18	69	1987
9–10	0	66	18–20	17	1004
10–11	0	44	20–22	27	388
11–12	0	22	22–24	5	123
12–13	0	9	24–26	0	72
13–14	0	8	26–28	0	28
14–15	0	0	28–30	0	6
15–16	0	0	30–32	0	1

The observations of *Hs* from the two data series are consistent in the overlap period. However, the original hourly observations of the wave period *Tz* obtained from the Paras data are consistently higher than those obtained from the HW data. The reason for this is the higher HW sampling rate (5 as opposed to 2 Hz), which results in a greater resolution of the smallest waves. In order to give a consistent series of observations of the wave period *Tz*, those obtained from the Paras data are multiplied by an empirically-derived constant of 0.86. Thus all the observations of this variable may be regarded as corresponding to a 5 Hz sampling frequency.

Appendix C. Comparison of winter and summer months

Throughout the collection period for the data studied here, there were various gaps in the recording process. These were caused by the need for periodic maintenance and calibration of the instrumentation. Most of these gaps occurred during the 'summer months' of May to August. To remove the bias caused by the uneven distribution of the data throughout the year, such data as do exist for the months of May to August are omitted entirely from the main study, and the resulting analysis—in particular the calculation of return levels and associated values of other variables—based on the assumption of an eight-month year. This corresponds to the assumption that we may neglect any probability that the extreme events of interest in the present study, which are typically associated with winter storms, can occur during the four months omitted from the analysis. Table 8 is a frequency table showing, for each of our primary variables *Hs* and *Ws*, the distribution of those hourly observations which *were* collected during the above summer period of May to August, together with the distribution of those observations collected during the

remaining 'winter period' of September to April. The table clearly shows the difference between the summer and winter distributions—in particular the very much shorter tail of the summer distribution—of each of the two variables. These differences thus confirm the reasonableness of treating as negligible any probability that the extreme events of interest can occur during the summer period, and of basing our analysis as described above on the September to April data.

Appendix D. Choice of period length for independent observations

We consider the problem, discussed in Section 3.2.1, of choosing a length of period such that, in the analysis of the joint distribution of wave height and wind speed, the bivariate observations associated with the successive periods can reasonably be treated as independent. As suggested in that section, we investigate the suitability of period lengths varying from 24 to 60 h.

As discussed in Section 2.4, it should be sufficient to consider separately each of the two variables. For each possible choice of period length, we take Hs_i and Ws_i to be the maximum observed values of *Hs* and *Ws*, respectively, during each period *i*, *i* = 1,…,*n*. We consider two approaches: (a) examination of the serial autocorrelation structure of the associated observations Hs_i and Ws_i; and (b) examination of the serial dependence of the occurrence of threshold exceedances by these observations. Recall that both of these techniques require the data to be fairly carefully separated according to season.

To examine the serial autocorrelation structure of each of the above variables, the data were split into series corresponding to each of the 64 months during which they were collected, so that each of these series might reasonably

Table 9
Threshold exceedances: analysis of first-order dependence

Month	Hs					Ws				
	n_{00}	n_{01}	n_{10}	n_{11}	p-value	n_{00}	n_{01}	n_{10}	n_{11}	p-value
Sep.	84	4	5	0	0.69	61	7	8	8	< 0.01
Oct.	87	6	7	1	0.26	51	19	19	10	0.32
Nov.	71	7	6	2	0.21	42	18	16	11	0.23
Dec.	48	12	13	5	0.35	21	22	20	18	0.55
Jan.	46	16	17	23	< 0.01	29	14	15	46	< 0.01
Feb.	47	19	15	7	0.39	27	20	21	19	0.40
Mar.	60	10	9	3	0.30	40	12	12	14	< 0.01
Apr.	76	2	3	3	< 0.01	62	10	10	5	0.08

be considered stationary. A similar approach was used by Walshaw [26] and gives a reasonable compromise between allowing for gradual seasonal change and obtaining sufficiently long series to allow useful statistical analysis. For each of the two variables and for period lengths of 48 and 60 h there is very little evidence of significant autocorrelation in the 64 monthly series, so that, on the basis of this analysis, it seems reasonable to take a period length of 48 h as sufficient to enable the associated observations to be treated as independent.

As a second method of identifying a suitable period length, we consider the serial dependence of threshold exceedances by the associated observations. As before, to minimise problems of seasonality we perform a separate analysis for each month. However, by restricting attention to first-order dependence (in which any short-term dependence will certainly show) we may effectively combine each month's data across the eight years for which it is available. This permits a more powerful analysis than that considered above.

Consider a fixed (trial) period length and variable X_j, where here $X_j = Hs$ or Ws. Let u_j be the threshold defined in Section 2.1, above which the distribution of the observations X_{ij} to be associated with this period length is modelled by a GPD. For each period i define

$$y_i = \begin{cases} 0 & \text{if } X_{ij} \leq u_j \\ 1 & \text{if } X_{ij} > u_j \end{cases}$$

Further, for $r = 0,1$, define p_r to be the (underlying) probability that $y_{i+1} = 1$, conditional on the event that $y_i = r$. Now, for each month, consider all pairs of successive periods within that month (over all years), and construct the two-by-two contingency table (n_{rs}, $r,s = 0,1$) where n_{rs} is the number of such pairs $(i, i + 1)$ with $y_i = r$ and $y_{i+1} = s$. Each of these monthly contingency tables may now be analysed for evidence of association in the counts n_{rs}. Any such association corresponds to $p_0 \neq p_1$, i.e. to first-order dependence in the process of threshold exceedances by the variable X_j, and so also to first-order dependence in the observations X_{ij}.

Table 9 shows the results for the period length of 48 h

identified earlier. We have here used the thresholds identified in Section 3.2.2 of 6.5 m for Hs and 16.5 m s^{-1} for Ws, but the results of this analysis are not particularly sensitive to the exact choice of threshold. For each variable and month, the table records the counts n_{rs} and the p-value (significance level) associated with the test of the null hypothesis that the underlying probabilities p_0 and p_1 are equal against the *one-sided* alternative that $p_0 < p_1$. These p-values are those which result from the standard test for the difference of two binomial proportions (with continuity corrections where appropriate). Some care must be taken in their interpretation, because of the sometimes small counts involved. However, it is clear that there is evidence of first-order dependence in the processes of threshold exceedances only for relatively few combinations of variable and month. The results therefore provide some very modest evidence of slight further serial dependence in these threshold exceedance processes, but this seems very unlikely to be so great as to seriously affect our subsequent analysis. No further improvement is obtained by the use of a period length of 72 h.

References

[1] Carter DJT, Challenor PG. Return wave heights at Seven Stones and Famita estimated from monthly maxima. IOS Report No. 66, 1978.

[2] Challenor PG. Confidence limits for extreme value statistics. IOS Report No. 82, Godalming, 1979.

[3] Coles SG. Extreme value theory and applications. http://www.maths.lancs.ac.uk/~coless/extremes/notes.ps, 1996.

[4] Coles SG, Tawn JA. Modelling extreme multivariate events. J R Statist Soc B 1991;53:377–392.

[5] Coles SG, Tawn JA. Statistical methods for multivariate extremes: an application to structural design. Appl Statist 1994;43:1–48.

[6] Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall, 1974.

[7] Davison AC, Smith RL. Models for exceedances over high thresholds. J R Statist Soc B 1990;52:393–442.

[8] Department of Energy, Metocean parameters—wave parameters. Offshore Technology Report OTH 89-300, HMSO, 1990.

[9] Department of Energy, Metocean parameters—parameters other than waves. Offshore Technology Report OTH 89-299, HMSO, 1990.

[10] Embrechts P, Klüppelberg C, Mikosch T. Modelling extremal events. Berlin: Springer, 1997.

[11] Guedes Soares C, Ferreira JA. Modelling long-term distributions of significant wave height. Proc Offshore Mechanics and Arctic Engineering, Am Soc Mech Enginrs 1995;2:51–61.

[12] Guedes Soares C, Henriques AC. On the statistical uncertainty in long term predictions of significant wave height. Offshore Mechanics and Arctic Engineering, ASME 1994;II:67–75.

[13] Gumbel EJ. Statistics of extremes. New York: Columbia University Press, 1958.

[14] de Haan L, Resnick SI. Limit theory for multivariate extremes. Z Wahrsch Theor 1977;40:317–337.

[15] de Haan L. Extremes in higher dimensions: the model and some statistics. In: Proc 45th Sess Int Stat Inst, paper 26.3, 1985.

[16] Isaacson M, Mackenzie NG. Long-term distributions of ocean waves: a review. J Waterways Port Coastal Ocean Div, Am Soc Enginrs 1981;107:93–109.

[17] Joe H, Smith RL, Weissman I. Bivariate threshold methods for extremes. J R Statist Soc B 1992;54:171–183.

[18] Leadbetter MR. Extremes and local dependence of stationary sequences. Z Wahrscheinlichkeitstheorie verw Gebiete 1983;65:291–306.

[19] Mathisen J, Bitner-Gregersen E. Joint distributions for significant wave height and wave zero-up-crossing period. Appl Ocean Res 1990;12:93–102.

[20] Morton ID, Bowers J. Extreme value analysis in a multivariate offshore environment. Appl Ocean Res, in press.

[21] Muir LR, El-Shaarawi AH. On the calculation of extreme wave heights: a review. Ocean Engng 1986;13:93–118.

[22] Resnick SI. Extreme values, regular variation and point processes. New York: Springer, 1987.

[23] Silverman BW. Density estimation for statistics and data analysis. London: Chapman and Hall, 1986.

[24] Smith D, Birkinshaw M. ISO (draft) offshore design code—wave loading requirements. Trans Marine Enginrs 1996;108:89–96.

[25] Tucker MJ. Waves in ocean engineering: measurement, analysis and interpretation. New York: Ellis Horwood, 1983.

[26] Walshaw D. Getting the most from your extreme wind data: a step by step guide. J Res Nat Inst Stand Technol 1994;99:399–411.