Submitted to Ocean Engineering

September 2026

Modelling marginal and joint extremes – Introducing the EnvEx approach showcased for metocean related challenges

Patrik Bohlinger[a], Philip Jonathan[b], Ben Youngman[c], Theo Economou[c], Jan-Victor Björkqvist[a], Øyvind Breivik[a,d]

[a] *Division for Oceanography and Maritime Meteorology, Norwegian Meteorological Institute, Allégaten 70, Bergen, 5007, Vestland, Norway*
[b] *School of Mathematical Sciences, Lancaster University, Lancaster, United Kingdom*
[c] *Department of Mathematics and Statistics, University of Exeter, Exeter, United Kingdom*
[d] *Geophysical Institute, University of Bergen, Allégaten 70, Bergen, 5007, Vestland, Norway*

Abstract

The use of non-stationary and joint extremes have become increasingly popular in climate studies and poses a critical challenge in metocean related applications. We present the EnvEx (**Env**ironmental **Ex**tremes) approach along with tailored R code to model non-stationary and joint extremes for common problems in geophysics, demonstrate its performance for typical metocean related tasks, and compare against conventional approaches. EnvEx builds on recent advances in extreme value statistics and consists of a hierarchical model that produces a predictive distribution of annual or $\mathcal{T}$-year maxima. Multiple return value estimators are compared. At its core, the hierarchical model is built from generalised additive models to introduce non-stationarity and the Heffernan and Tawn conditional extremes model for modelling joint extremes. We show that not taking into account serial correlation or covariate effects can lead to severe biases for very long and short tails. Predictive model validation of EnvEx is based on synthetic data generation. Finally, we demonstrate the full workflow applied to a wave hindcast time series illustrating information rich results that can be harvested for various downstream applications.

Keywords: Non-stationary extremes, joint extremes, Generalised Additive Models, metocean, marine climate, Heffernan and Tawn

1. Introduction and motivation

In the wake of climate change and climate prediction, modelling extreme events and compound events of extremes have become increasingly important during the last decades [5]. It is of particular interest to correctly estimate the return values of severe geophysical conditions as those may shape mitigation measures to prevent the loss of life and property. To meet these challenges, we present a versatile framework that is validated across various scenarios and compared against a range of conventional and alternative methods. We tailor our study to metocean related challenges but underline that its applicability is of general nature.

In metocean applications, extreme environmental conditions are of particular interest during the design phase, as they can impose exceptionally large loads on structures. While this is true for single response variables, typically significant wave height (*Hs*) or the maximum individual

wave height ($H_{\max}$), it also applies to more complex situations where multiple variables can be extreme simultaneously, e.g. joint extremes of *Hs* and wind speed at 10m (*U10*), mean period (e.g. T_z), or wave steepness (s). Relevant guidelines on how to address this issue, e.g. NORSOK, DNV-RP-C205, or ISO19900 are neither straightforward nor unambiguous. In fact, sometimes even the question that should be answered is not clearly defined. It is, however, crucial to define what type of extreme and exceedance probability one tries to model, because the answers may differ considerably [10]. In the following, we describe typical metocean data and discuss three sources of confusion: the definition of the sought probability, the effect of correlated data, and finally, different modelling approaches.

Suppose that the primary variable of interest is X (e.g $X = Hs$). Then an arbitrary time series could look like the illustration in Fig. 1. This serially correlated and non-stationary time series is characterized by local minima and maxima where the local maxima represent storm peaks (*SP*) and are depicted as circles. The probability of non-exceedance can be written in terms of annual maxima X^A as $\mathbb{P}\left(X^A \leq q_{\mathcal{T}}\right) = 1 - \frac{1}{\mathcal{T}}$. The threshold that is exceeded at least once within a given $\mathcal{T}$-year period of time is called the $\mathcal{T}$-year return level for a return period $\mathcal{T}$ and denoted $q_{\mathcal{T}}$. Assuming *iid* and no covariates, the non-exceedance probability can equally be expressed in terms of storm peaks or the full sea-state data set, i.e.

$$\mathbb{P}(X^{SP} \leq q_{\mathcal{T}}) \approx 1 - \frac{1}{\lambda \mathcal{T}}$$
$$\mathbb{P}(X \leq q_{\mathcal{T}}) \approx 1 - \frac{1}{N\mathcal{T}}, \qquad \text{(for large } N, \mathcal{T}\text{)},$$

with N as the number of sea-states per annum and λ the expected number of storms per annum. Metocean data, however, is usually not *iid* often with heavy serial correlation and multiple covariates at play. In these cases, $q_{\mathcal{T}}$ cannot simply be computed by means of $\mathbb{P}\ (X \leq q_{\mathcal{T}})$ using all data (see Supplementary Section 1 for mathematical reasoning). In fact, neglecting whether data are *iid* or not when estimating $q_{\mathcal{T}}$ has been shown to introduce considerable biases [1, 17, 18].

It is generally challenging to model the tail of a distribution. Extreme value theory (EVT) provides a sound mathematical basis [7]. Provided that data are *iid*, conventional stationary EVT is sufficient for estimating global return levels. A challenge arises when a more efficient data usage is needed or where it is important to focus on covariate subspaces. For joint extremes, the Heffernan and Tawn conditional extremes model [12] (henceforth denoted as HT2004) has become increasingly popular. Relevant standards, however, also allow the use of the initial distribution method [13] (typically a Weibull fit to all data) and for multivariate problems variations of the LogNorm-Weibull method [2]. Justifications are based on experience and the strong assumption that one can confidently choose a specific distribution, i.e. a Weibull distribution, for waves. For multivariate cases one assumes that conditional dependence estimated from all data continues to be valid in the tail. Both assumptions can usually not be verified.

The described issues can accumulate when the metocean engineer is tasked with estimating a return level on non *iid* data. NORSOK or DNV-RP-C205 allow for both the initial distribution method and EVT approaches, which effectively promotes the interchangeable use $P(X^A <= q_{\mathcal{T}}) = 1 - \frac{1}{\mathcal{T}}$ and $P(X <= q^*_{\mathcal{T}}) = 1 - \frac{1}{N\mathcal{T}}$. The second equation yields a biased estimate $q^*_{\mathcal{T}}$ as

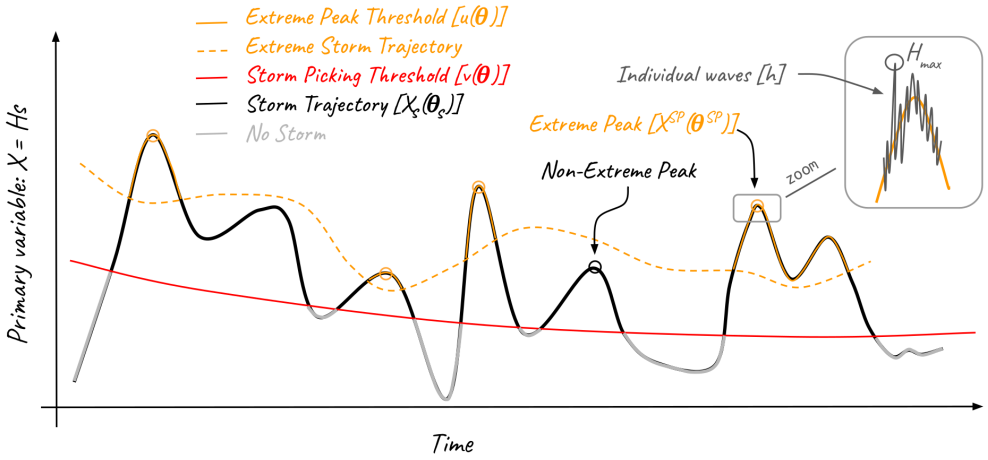


Figure 1: Sketch illustrating a typical metocean time series and the concept behind the `EnvEx` workflow. The primary variable X is chosen to be Hs with a set of covariates $\boldsymbol{\theta}$ and storm peak covariates $\boldsymbol{\theta}^{SP}$. The here introduced notation is consistent with other equations in the manuscript that describe the statistical models or workflow so that this sketch can serve as reference. The displayed timeseries is usually multivariate with contemporaneous values of an arbitrary number of available variables.

it ignores the effect of serial correlation of sea states and $q_{\mathcal{T}} \neq q^*_{\mathcal{T}}$. Resulting biases can propagate further to multivariate estimates for which DNV-RP-C205 recently admitted the use of the HT2004 model, but the LogNorm-Weibull method remains the recommended method.

The struggle to find experts that can implement and interpret more sophisticated methods has limited the experience regarding their quality in real world cases. However, during the past decades, methods for non-stationary modelling of metocean extremes were developed and their use demonstrated for non-stationary marginal and joint extremes [14, 15, 20]. In particular, the use of penalised splines provided a basis for efficiently modelling complex non-stationary features while at the same time avoiding overfitting and ensuring reproducibility. The use of splines in the context of generalised additive models (GAMs) was developed further and wrapped into versatile libraries by Wood [26] (`mgcv`), Youngman and Economou [30] (`ppgam`), and Youngman [29] (`evgam`), which are free and open-access `R` pakages.

We aim to bridge the aforementioned gap between methodology and practice by presenting a workflow based on GAMs in `R` for modeling non-stationary marginal and joint extremes. We call this workflow `EnvEx`, short for **Env**ironmental **Ex**tremes. This workflow allows the application of custom parametric and non-parametric models for a wide range of metocean environments, and also broader geophysical challenges that include various kinds of non-stationarities. Where possible, we avoid apriori assumptions about the characteristics of the data and use EVT for guidance in all steps. We design synthetic benchmark datasets for which the true return value is reasonably well known. Multiple methods, including stationary EVT models, `EnvEx` and clearly misspecified models, are subsequently challenged to estimate the return value. We investigate the effect of serial correlation, tail properties, and covariates on the return level estimates and discuss the consequence of attempting to estimate q_T incorrectly using $P(X <= q_{\mathcal{T}}) = 1 - 1/(N\mathcal{T})$. In all

cases, synthetic data are generated from a Weibull distribution, reflecting a common assumption in engineering applications.

The manuscript is structured as follows. First, we introduce the `EnvEx` approach (Sec. 2). We then present the experimental design, the model used to generate the synthetic data, and the various realizations that form the basis of our benchmark (Sec. 3). We continue with listing the approaches to be evaluated and describe how to calculate the 100-year event for each dataset and method (Sec. 4). We outline how uncertainty is quantified (Sec. 5) and finally describe and discuss the benchmarking results for multiple, typical scenarios (Sec. 6). Finally, we present results from applying `EnvEx` to hindcast data for a location in the central North Sea (Sec. 7).

2. `EnvEx` - modelling environmental extremes

`EnvEx` is a general framework for modelling non-stationary and joint extreme values and relies on non-parametric approaches where no prior assumption on the data distribution is needed. This is achieved by utilizing GAMs as provided by the `evgam` package in `R` [29]. In the following, we will outline the statistical model and present an intuitive interpretation. Details on the individual parts of `EnvEx` are provided later in dedicated chapters.

The primary goal in a metocean context is often to estimate the maximum response or the individual wave height $H_{\max}$ within a given period of time, i.e. the return period $\mathcal{T}$. This complicated, multidimensional problem can be simplified by first understanding the distribution of $H_{\max}$ in any random storm. As elucidated in [10], it is the maximum value in a storm evolving over a sequence of sea states that is of interest for design. This maximum, however, does not necessarily occur during the peak sea state of a storm, which is why the entire storm event needs to be considered. The characteristics of these storm events may depend on various conditions, e.g. season, direction, decadal and multi-decadal oscillations, climate trends, and so on. Given these conditions, storm events can usually be assumed independent from each other. Under this assumption, storm peaks can also be treated as independent and $q_{\mathcal{T}}$ can be calculated using either $P(X^A <= q_{\mathcal{T}}) = 1 - 1/\mathcal{T}$ or $P(X^{SP} <= q_{\mathcal{T}}) = 1 - \frac{1}{\lambda\mathcal{T}}$. They are essentially a manifestation of chaos, i.e. the stochastic realization of an instantaneous climate state, commonly called weather. In statistics, modelling independent events is more manageable, which provides a suitable starting point for building a hierarchy of assumptions and models for estimating $H_{\max}$.

We consider a collection of p storm variables $\mathbf{X}$. Let $\widetilde{X}$ denote its primary variable, which we refer to again in Section 2.1. Our first assumption is that storm events are independent given their storm peak values and appropriate covariates. This can be expressed in terms of the probability density of $\mathbf{X}^{SP}$, which denotes the subset of $\mathbf{X}$ corresponding to storm peaks, and a variety of corresponding covariates $\boldsymbol{\theta}^{SP}$. Secondly, we assume that storm events, i.e. the multivariate trajectory as illustrated in Fig. 1, can be sufficiently characterized by their respective storm peak variables and the storm peak covariates. The storm evolution itself can be described by the probability density of a multivariate time series of sea state characteristics $\mathbf{X}_s$ throughout the storm event, sea state covariates $\boldsymbol{\theta}_s$, and the unknown, variable storm length T. Thirdly, we assume that the dependence between individual waves given their sea state is negligible, which is essentially the fundamental assumption of short-term wave statistics condensed in distribution functions for individual waves. The assumption of independence of the sea state depending $H_{\max}$ allows for straightforward modelling of $H_{\max}$ in a storm event. Considering uncertainty introduced e.g.

by user specific choices on hyperparameters and following the notation developed by [21], we formulate the distribution of the maximum individual wave of any random storm in Eq. (1).

$$
\begin{aligned}
F_{H_{\max},s|\mathbf{Z}}\left(h|\zeta\right) = & \int_{\boldsymbol{\theta}^{SP}} \int_{\mathbf{x}^{SP}} \int_{(\{(\mathbf{x}_s,\boldsymbol{\theta}_s)\},\tau)_{s\in\mathcal{S}_\tau}} \\
\text{A)} \quad & \prod_{s\in\mathcal{S}_\tau} F_{H_{\max,L}|(\mathbf{x}_s,\boldsymbol{\theta}_s,\mathbf{Z})}\left(h|\mathbf{x}_s,\boldsymbol{\theta}_s,\zeta\right) \\
\text{B)} \quad & \times f_{(\{(\mathbf{X}_s,\boldsymbol{\theta}_s)\},T)_{s\in\mathcal{S}_\mathcal{T}}|\mathbf{x}^{SP},\boldsymbol{\theta}^{SP},\mathbf{Z}}\left(\{(\mathbf{x}_s,\boldsymbol{\theta}_s)\}_{s\in\mathcal{S}_\tau},\tau\middle|\,\mathbf{x}^{SP},\boldsymbol{\theta}^{SP},\zeta\right) \\
\text{C)} \quad & \times f_{\mathbf{X}^{SP}|\boldsymbol{\theta}^{SP},Z}\left(\mathbf{x}^{SP}|\boldsymbol{\theta}^{SP},\zeta\right) \\
\text{D)} \quad & \times f_{\boldsymbol{\theta}^{SP}|\mathbf{Z}}\left(\boldsymbol{\theta}^{SP}|\zeta\right) \\
& \mathrm{d}\left(\{(\mathbf{x}_s,\boldsymbol{\theta}_s)\}_{s\in\mathcal{S}_\mathcal{T}},\tau\right)\mathrm{d}\mathbf{x}^{SP}\mathrm{d}\boldsymbol{\theta}^{SP}
\end{aligned}
\tag{1}
$$

This is a hierarchical model consisting of four essential modules for: A) $H_{\max}$ within a storm event with random length $\tau \in T$ consisting of sea states s of length L, B) the evolution of all multivariate storm events characterized by all multivariate storm peaks given their covariates, C) the independent storm peak variables given their covariates, and D) the occurrence of storm peak covariates. Uncertainties formulated as parameters $\mathbf{Z}$ can be integrated out to obtain a full predictive distribution containing all considered uncertainties. As in [21], we need to move from the distribution of $H_{\max}$ of any random storm to the distribution of $H_{\max}$ per annum and finally the return period $\mathcal{T}$. Assuming the number of storms within any given year is Poisson-distributed with expectation $\lambda \in (0,\infty)$ per annum, we can formulate the distribution of $H_{\max}$ in a year as:

$$
F_{H_{\max},A}(h) = \sum_{m=0}^{\infty}[F_{H_{max},s}(h)]^m \frac{\lambda^m e^{-\lambda}}{m!} = \exp\left\{-\lambda\left[1 - F_{H_{max},s}(h)\right]\right\} \tag{2}
$$

with $m \in \mathbb{N}$ representing the number of events. $H_{\max}$ per annum can be translated to the quantile of the desired return period as:

$$
q_\mathcal{T} = F^{-1}_{H_{\max},A}(1 - 1/\mathcal{T}). \tag{3}
$$

Once the `EnvEx` models are established, one can generate the predictive distribution $F_{H_{\max},A}$ of annual-maxima or $F_{H_{\max},A_\mathcal{T}}$ of $\mathcal{T}$-maxima. Subsequently, one can retrieve different return value estimators $q = Q_A\left(1 - 1/\mathcal{T}\right)$ and $q' = Q_{A_\mathcal{T}}\left(\exp[-1]\right)$ from the respective quantile functions Q_A or $Q_{A_\mathcal{T}}$. The implemented return level estimators $q2$, $q'2$, $q3$, $q4$, and $q5$ are defined as in Jonathan et al. [16] and Towe et al. [23]. Further explanation, formulations and known biases can be found therein and in Supplementary Section 2.

For now, `EnvEx` models the joint tail behavior of n joint contemporaneous variables, where the primary variable, i.e. the variable defining the storm trajectory in Fig. 1, has to be chosen by the user depending on the needs at hand. If other needs arise in the future this can in principle be expanded. For more information on this "full integrated forward modelling approach" see, Towe et al. [22] and Speers et al. [21]. In this study, we focus on describing, illustrating, and validating the model from Eq.(1). For our purpose, we follow the basic steps of Ewans and Jonathan [8], i.e. (1) isolating storms and their respective peaks using a non-stationary threshold model; (2) modelling the occurrences of extreme storm peaks separated by another non-stationary threshold model; (3) modelling the severity of storm peaks using an exceedance model; (4) modelling joint extremes using the HT2004 model; and (5) modelling H_{max} using history matching and a short-term individual wave-height distribution.

2.1. GAMs, overfitting, and covariate interactions

Recall the primary variable of interest $\widetilde{X}$. Let $f(\tilde{x}; \boldsymbol{\zeta}(\boldsymbol{\theta})$ denote its marginal distribution, which has parameters $\boldsymbol{\zeta}(\boldsymbol{\theta})$ that depend on covariates $\boldsymbol{\theta}$; henceforth we write $\widetilde{X}(\boldsymbol{\theta})$. We model all non-stationary marginal distribution parameters using GAMs. For a J-parameter distribution, $\boldsymbol{\zeta}(\boldsymbol{\theta}) = (\zeta_1(\boldsymbol{\theta}), \ldots, \zeta_J(\boldsymbol{\theta}))$. A GAM formulation means that its jth parameter, $\zeta_j(\boldsymbol{\theta})$, $j = 1, \ldots, J$, is characterised by M_j smooth functions:

$$\zeta_j(\boldsymbol{\theta}) = \zeta_{j0} + \sum_{m=1}^{M_j} g_{jm}(\boldsymbol{\theta})$$

where $g_{jm}(\boldsymbol{\theta})$ denotes a spline function of $\boldsymbol{\theta}$ with basis coefficients $\boldsymbol{\beta}_{jm}$. (In practice, link functions may be used to preserve legitimate parameter ranges). Typically, $g_{jm}(\boldsymbol{\theta})$ will denote univariate splines, e.g. $g_{jm}(\theta_l)$, for $l = 1, \ldots, L$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$, or the tensor product of univariate splines, e.g. $g_{jm}(\theta_l, \theta_{l'})$. Tensor products allow for covariate effects that cannot be assumed independent, such as the interaction between wave direction and season. If we observe a finite set of data, $\tilde{x}(\boldsymbol{\theta}_1), \ldots, \tilde{x}(\boldsymbol{\theta}_n)$, say, under probability distribution f we have likelihood $\mathcal{L}$. We fit extreme value distributions with parameters of GAM form by maximising a penalised log-likelihood of the form

$$\log\left(\mathcal{L}_p\right) = \log\left(\mathcal{L}\right) - \frac{1}{2}\sum_{j=1}^{J}\sum_{m=1}^{M_j} \rho_{jm}\boldsymbol{\beta}_{jm}^{\top}\mathbf{S}_{jm}\boldsymbol{\beta}_{jm}, \tag{4}$$

where smooth g_{jm} has corresponding penalty matrix $\mathbf{S}_{jm}$ and smoothing parameter ρ_{jm}. We estimate ρ_{jm} using Laplace's method [25, 27], as implemented in the `R` package `evgam` [29].

2.2. Threshold model for selecting storms and peaks

The primary variable, $\widetilde{X}$, is used to identify storms (Fig. 1). Storms are defined as the sequence of the $\widetilde{X}$ time series that is above a threshold defined by a non-stationary quantile. The threshold model used to distinguish storms can be non-stationary because it is allowed to vary with covariates.

Quantile regression is used to estimate the non-stationary threshold, denoted $v(\boldsymbol{\theta})$. We choose quantile τ so that $\Pr(\widetilde{X}(\boldsymbol{\theta}) > v(\boldsymbol{\theta})) = \tau$. Following Youngman [28], we use the asymmetric Laplace distribution (ALD), to estimate the threshold via:

$$\widetilde{X}(\boldsymbol{\theta}) \sim \mathrm{ALD}(v(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta})),$$

with $\boldsymbol{\zeta}(\boldsymbol{\theta}) = (\zeta_1(\boldsymbol{\theta}), \zeta_2(\boldsymbol{\theta}))$, where $\zeta_1(\boldsymbol{\theta}) = v(\boldsymbol{\theta})$ is the threshold and $\zeta_2(\boldsymbol{\theta}) = \log\sigma(\boldsymbol{\theta})$ is a scaling parameter.

Parameter $v(\boldsymbol{\theta})$ is the "storm-picking threshold" in Fig. 1. All storms are subsequently labeled. Let $\mathcal{I}_s = \{i \in \{1, \ldots, n\} : \widetilde{x}(\boldsymbol{\theta}_i) > v(\boldsymbol{\theta}_i)\}$ denote those values of the original time series corresponding to storms. If storms are too close to each other and therefore would violate the *iid* assumption, they are combined based on a declustering timescale of the order of days (typically 1-3 days). The result is a time series of storm sea states with the respective storm indices and all associated variables that are chosen to be in the dataset is illustrated in Figure 2.

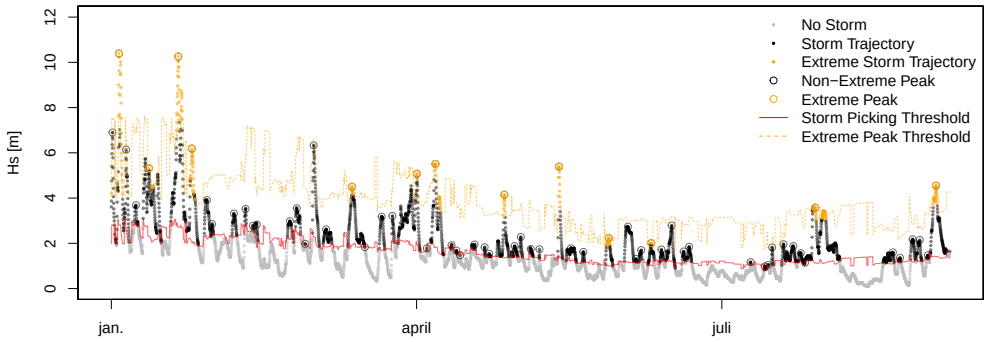


Figure 2: Illustration of a non-stationary threshold applied to storm picking and extreme storm peak picking on NORA3 wave hindcast [4] data at a location in the North Sea. The time series consists of approximately half a year worth of hourly data. The storm picking threshold is set to the median. The effect of the non-stationarity on the threshold, i.e. the dependence on covariates, is clearly visible as a trend and fluctuations along the threshold line.

2.3. Threshold model for extreme storm peaks

We focus now on picking and modelling only the extreme storm peaks from the storm dataset from the previous step. This corresponds to picking the $\mathbf{X}^{SP}$ given $\boldsymbol{\theta}^{SP}$ in Eq. (1) that we model in section 2.4. Let $u(\boldsymbol{\theta}^{SP})$ denote the extreme value threshold, which must be sufficiently high to separate the storm peak data. We also estimate $u(\boldsymbol{\theta}^{SP})$ by quantile regression and choose $\omega \gg \tau$ so that $\Pr(\tilde{X}(\boldsymbol{\theta}^{SP}) > u(\boldsymbol{\theta}^{SP})) = \omega$. The result is a dataset of extreme storm peak exceedances that can further be ingested into an exceedance model and an occurrence model (Fig. 1, Fig. 2).

2.4. Exceedance model

Given the extreme value threshold $u(\boldsymbol{\theta}^{SP})$, excesses of the threshold are modelled using a non-stationary generalised Pareto (GP) distribution, again following Youngman [28]. Our GP model takes the form

$$\tilde{X}(\boldsymbol{\theta}^{SP}) - u(\boldsymbol{\theta}^{SP}) \mid [\tilde{X}(\boldsymbol{\theta}^{SP}) > u(\boldsymbol{\theta})^{SP}] \sim \text{GPD}\left(\psi(\boldsymbol{\theta})^{SP}, \xi(\boldsymbol{\theta}^{SP})\right) \tag{5}$$

with scale parameter $\psi(\boldsymbol{\theta}^{SP}) > 0$ and shape parameter $-1 < \xi(\boldsymbol{\theta}^{SP}) < 0.5$ constrained to ensure finite variance. See Supplementary Section 3 for the GP distribution's parameterisation. For the GP model we take $\boldsymbol{\zeta}(\boldsymbol{\theta}^{SP}) = (\zeta_1(\boldsymbol{\theta}^{SP}), \zeta_2(\boldsymbol{\theta}^{SP}))$, where $\psi(\boldsymbol{\theta}^{SP}) = \exp\{\zeta_1(\boldsymbol{\theta}^{SP})\}$ and $\xi(\boldsymbol{\theta}^{SP}) = 1.5/[1 + \exp\{-\zeta_2(\boldsymbol{\theta}^{SP})\}] - 1$. The same covariates and GAM forms used for $u(\boldsymbol{\theta}^{SP})$ are used for $\zeta_1(\boldsymbol{\theta}^{SP})$ and $\zeta_2(\boldsymbol{\theta}^{SP})$. Likelihood $\mathcal{L}$ in Eq. (4) is calculated from the n_u storm peaks that exceed $u(\boldsymbol{\theta}^{SP})$, i.e. $\{\widetilde{x}(\boldsymbol{\theta}_i^{SP})\}_{i \in \mathcal{I}_u}$, where $\mathcal{I}_u = \{i : \widetilde{x}(\boldsymbol{\theta}_i^{SP}) > u(\boldsymbol{\theta}_i^{SP})\}$, $n_u = \sum_{i=1}^{n} I[\widetilde{x}(\boldsymbol{\theta}_i^{SP}) > u(\boldsymbol{\theta}_i^{SP})]$, and $I[\,]$ denotes the indicator function.

2.5. Occurrence model

In order to compute the correct return values of exceedances or peaks, the correct number of occurrences has to be modelled and factored in. We use the non-stationary Poisson point

process GAM of Youngman and Economou [30] via the `R` package `ppgam` to model the rate of occurrence, i.e. $\{i : i \in \mathcal{I}_u\}$. Occurrences are assumed to have a non-homogeneous rate, $\lambda(\boldsymbol{\theta}^{SP})$, where $\log \lambda(\boldsymbol{\theta}^{SP})$ is assumed to have GAM form, which corresponds to the above specification with $\boldsymbol{\zeta}(\boldsymbol{\theta}^{SP}) = \zeta_1(\boldsymbol{\theta}^{SP})$, where $\zeta_1(\boldsymbol{\theta}^{SP}) = \log \lambda(\boldsymbol{\theta}^{SP})$.

We acknowledge that circular distributions can be used, e.g. von Mises distribution as applied in [3] or a non-stationary Poisson distribution [11]. The advantage with the suggested solution, however, is that we do not assume a particular shape or distribution type of the circular density and any binning to establish counts is unnecessary.

2.6. Model fitting marginal models for storm peak extremes

Marginal extremes are a fundamental building block of `EnvEx` because they are needed even if considering joint extremes. In the case of joint extremes that is discussed in section 2.8, the marginal models lay the foundation in terms of the necessary probability integral transform to Laplace space. When considering univariate extremes the marginal models are used directly. Either way, the models for peak picking and extreme peaks, as well as occurrence and exceedence, need to be established and combined, i.e. part C and D in Eq. (1). The algorithm is designed as described in SM3.

2.7. Predicting marginal extremes

Marginal extremes can be predicted for one or several user-specified regions in the covariate space. The goal is to simulate maxima given a return period of interest. The model prediction starts by simulating storms, i.e. covariates, worth a time period defined by the return period $\mathcal{T}$. Once the covariates are known, they are propagated to the threshold and exceedence models that predict conditional values. The total maximum is picked as the simulated extreme storm peak. Executing this procedure multiple times results in a predictive distribution of $\mathcal{T}$-year maxima simulated under the corresponding covariates space. The algorithm is outlined in Supplementary Section 5 & 6. So far we use simple numerical simulations but we acknowledge the existance of other methods like closed form estimates or the use of importance sampling [24].

2.8. Model for joint extremes

The HT2004 model [12] can reproduce the dependence structure of a broad variety of copulas and model both asymptotic independence and asymptotic dependence. Asymptotic dependence is a corner case. The HT2004 model involves modelling the conditional distribution of one variable when the value of the conditioning variable is large. An advantage is that no prior knowledge of the forms of the distributions is required. Instead, asymptotic distributional forms are used.

Let $\mathbf{Y}$ denote the collection of storm variables, $\mathbf{X}$, with margins transformed to Laplace scale. Following the notation of Speers et al. [21], which has a regression-like form for data above an appropriately high threshold v,

$$\left(\mathbf{Y}_{-j}|Y_j = y\right) = \boldsymbol{\alpha}_{|j} y + y^{\boldsymbol{\beta}_{|j}} \mathbf{Z}_{|j}, \qquad y > v_j \tag{6}$$

where $\mathbf{Y}_{-j}$ represents the $(p-1)$ dimensions of $\mathbf{Y}$ with element j removed. The parameter vectors $\boldsymbol{\alpha}_{|j}$ and $\boldsymbol{\beta}_{|j}$ are estimated by applying the quasi-Newton method as modified in Byrd

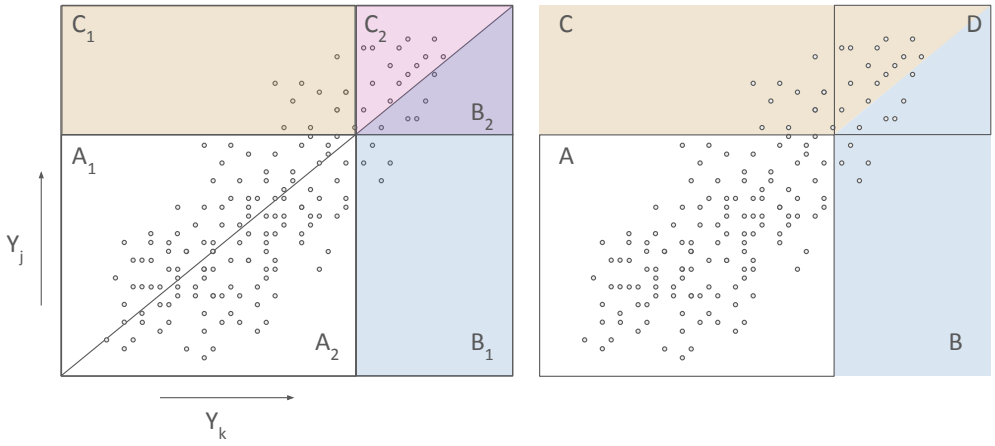


Figure 3: Sketch to illustrate regions of different categories with respect to the prediction of joint extremes.

et al. [6] with the constraints $0 \leq \alpha \leq 1.2$ and $-100 \leq \beta \leq 1.2$. The upper limit is greater than 1 to allow for probability to achieve the corner cases of $\alpha, \beta = 1$ given uncertain and sub-asymptotic data. Diagnostic plots can help to evaluate the parameter stability and are exemplified in Supplementary Figure 3.

2.9. Simulation of joint extremes

Once all p-models are established, values from the p-dimensional joint extreme value model can be simulated. In principle many different questions can be answered regarding the joint distribution. In this paper, we specifically simulate the joint 100-year maxima under the HT2004 model following the steps:

1. Simulate $\mathcal{T}$-year return values from the respective marginal model to be used as the conditioning variable Y_k.
2. Simulate values for the conditioned variable Y_j using each HT2004 model for the given values from the conditioning variable Y_k.
3. Transform valid results from standard Laplace back to original margins and pool simulations. Ensure the correct occurrence rate for each sector where the respective models are valid and values were rejected.

The HT2004 simulations can be divided into 10 regions (A, A1, A2, B, B1, B2, C, C1, C2, D) of which 7 regions (B, B1, B2, C, C1, C2, D) can be modelled with the described procedure where non-stationary marginal EVT models where utilized (see Fig. 3 for orientation). Region B1+D and region C1+D represent regions where either Y_k or Y_j is extreme, respectively. In region D, both Y_k and Y_j are extreme. The practitioner can obtain results for any region by filtering and consolidating simulations accordingly. The established models should be checked using diagnostic plots of the simulations (e.g. Supplementary Figure 4).

2.10. Storm model

We use a technique broadly called "history matching" following Feld et al. [9]. This generally means that historic events are matched up with simulated storm peaks chosen based on

minimizing a dissimilarity metric. The storm is subsequently linearly scaled to match the magnitude of the simulated storm peak. Periods are rescaled to maintain the steepness of the historic waves keeping the relationship to the wave heights physically consistent with the aim of producing realistic and comparable sea states. For the storm integration later on and the directional sector dependent prediction of individual wave heights, we rotate the simulated storm trajectory according to the circular difference between the direction at the simulated storm peak and the direction of the historic storm peak.

The dissimilarity metric d is the simple euclidean distance between the multivariate simulated storm peak $\mathbf{X}^{sim}$ and the corresponding variables from the historic storm peak $\mathbf{X}^{hist}$. A user specific number of closest matches is chosen from which a random index is picked. Prior to computing the dissimilarity, we normalize and standardize all variables, i.e. $z = (x - \overline{x})/\sigma_x$, so that the distance for each variable is in the same units, namely standard deviations, and thus their importance is equally weighted. The assigned distance d_i for each storm i is thus defined as $d_i = \sqrt{\left(\mathbf{X}_i^{sim} - \mathbf{X}_i^{hist}\right)^2}$. For circular variables circular distance functions are used. Due to the random selection of storms, different storm lengths and storm shapes are also sampled, so that, provided there are enough historical storms, the simulated storms resemble the characteristics of the historic storms and approximate the characteristics of the true historic storm population. The chosen weighting is subjective for now but could be optimized to match best the considered location and its data at hand.

2.10.1. Storm integration algorithm

Depending on the sea state Hs^{SS} along the storm trajectory, there may be a considerable probability of H_{max} to occur before or after the storm peak Hs^{SP} [10]. This section outlines therefore an algorithm to compute H_{max} along a storm trajectory illustrated using the Rayleigh distribution for a narrow-band Gaussian sea state. The algorithm is provided in detail in Supplementary Section 8 together with additional validation.

Assuming individual wave events are independent, for a single sea state with length D the exceedance probability of $H_{\max}$ can be expressed as:

$$\mathbb{P}\left(H_{max} > x \mid Hs, T_z, D\right) = 1 - \left[1 - \exp\left(-\frac{2x^2}{Hs^2}\right)\right]^N, N = D/T_z \tag{7}$$

For a storm with multiple time steps and different sea states, this generalises to:

$$P(x) = \mathbb{P}(H_{max,storm} > x) = 1 - \prod_{i=1}^{n}\left[1 - \exp\left(-\frac{2x^2}{Hs_i^2}\right)\right]^{N_i}, N_i = D_i/T_{z,i} \tag{8}$$

assuming that wave events between different sea-states are independent. Evaluating $P(x)$ over all possible thresholds results in an exceedance curve that we can use to compute the expected storm maximum and the most likely storm maximum. For the most likely storm maximum, let $F(x)$ be the CDF of the storm maximum, i.e. $P(x) = 1 - F(x)$ and the derivative $f(x) = \frac{dF}{dx}$. $x^* = \arg\max_{x \geq 0} f(x)$ is the most likely value we are after. The expected storm maximum can be computed by integrating $P(x)$, i.e.

$$\mathbb{E}[H_{max,storm}] = \int_0^\infty x\, f(x)\, dx = \int_0^\infty (1 - F(x))\, dx = \int_0^\infty P(x)\, dx$$
$$= \int_0^\infty \left(1 - \prod_{i=1}^{n} \left[1 - \exp\left(-\frac{2x^2}{Hs_i^2}\right)\right]^{N_i}\right) dx \tag{9}$$

Further on in this manuscript, we will use the mode corresponding to the most likely H_{max} given a storm trajectory.

3. Design of validation experiment using synthetic data

The validation experiment focuses on storm peak *Hs* because the estimation of H_{max} is contingent on the predicted peaks. We chose a WB3P as the baseline model for producing synthetic data with the density defined as:

$$f(x) = \left(\frac{C}{B}\right)\left(\frac{x-A}{B}\right)^{C-1} exp\left[-\left(\frac{x-A}{B}\right)^{C}\right] \tag{10}$$

where $A \in \mathbb{R}$, $B, C > 0$, and A represents the location parameter, B is the scale parameter, and C defines the shape. When $A = 0$, Eq. (10) is reduced to a 2-parameter Weibull distribution, when $C = 1$ this reduces to an exponential distribution.

We produce $N^{sim} = 10000$ years worth of data in a complexity cascade featuring four different categories of auto-correlation in the data:

1. This is the reference and the most basic level where *iid* simulations from a WB3P distribution produce one value a day given parameters $A_{const} = 4$, $B_{const} = 8$, $C_{const} = 1$. This is reproduced for other shape parameters with increasingly shorter tail, i.e. $C_{const} \in (1, 2, 3, 4, 5, 6)$. The parameter choice is somewhat arbitrary but was chosen here to create a dataset that resembles characteristics similar to typical time series of *Hs*.
2. The effect of a covariate, i.e. seasonal dependency, is introduced into the data. This is done by letting the WB3P parameters vary with a single seasonal cycle defined by a cosine function. Location and scale parameters are set to be largest in winter and smallest in summer, whereas the change in the shape parameter is shifted by π to produce longest tails in winter and shortest in summer. The parameters are varied as follows:

$$A = 3 \cdot cos(2\pi\frac{d}{\tau}) + A_{const} \tag{11}$$

$$B = 2 \cdot cos(2\pi\frac{d}{\tau}) + B_{const} \tag{12}$$

$$C = .3 \cdot cos(2\pi\frac{d}{\tau} + \pi) + C_{const} \tag{13}$$

The period τ is fixed to 365 and the date d is just an endless vector of days from 1 to 365 repeating themselves, creating a long cosine function from which N^{sim} years of daily realizations are chosen.

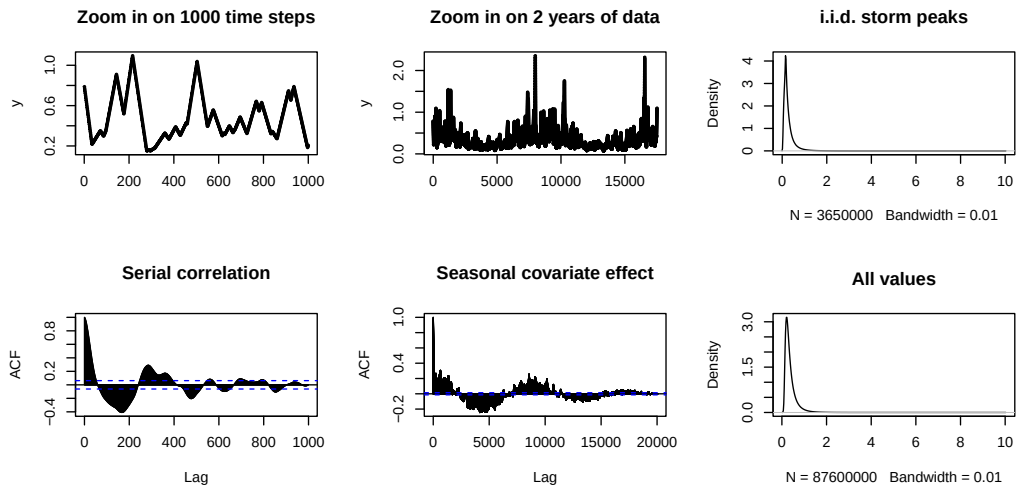


Figure 4: Time series figures for different chunks of the data, with densities of all storms peaks and all data, as well as the empirical auto-correlation function (AFC) on different time scales. The synthetic data is based on category 4 featuring a WP3P $C_{const} = 1$ and a power storm model with $\lambda = 1$. N depicts the number of values which differs for the peaks (daily) and the serially correlated data (hourly)

3. Serial correlation is introduced into the data without covariate effects using the power storm model [18] with various λ-parameters (0.5, 1.0, 2.0) where $\lambda = 1$ is equivalent to a triangular storm shape. We apply the storm model parameterized as:

$$Hs(t) = A\left[1 - \left(\frac{2|t - t_0|}{B}\right)^{\lambda}\right], \lambda > 0, -\frac{B}{2} \leq |t - t_0| \leq \frac{B}{2} \tag{14}$$

   For each WB3P value a storm is being produced and all data is consolidated. Wherever storms are overlapping, always the largest value is chosen. This leads to a series of triangular storms ($\lambda = 1$) where the largest storms dominate small storms and therefore last longer.
4. Finally both covariate effects and serial correlation are introduced combining steps 2 and 3.

Figure 4 illustrates the characteristics of a synthetic dataset using $\lambda = 1$ and non-stationary WB3P parameters based on shape $C_{const} = 1$. In general, the WB3P shape as well as the serial correlation influences considerably the shape and width of the empirical densities of the synthetic data (compare e.g. with $C_{const} = 6$ in Supplementary Figure 2). The autocorrelation functions (ACF) show the seasonal covariate effect as well as the serial correlation clearly. For stationary timeseries with no serial correlation, the ACF is flat.

4. Evaluating different methods

The central question to be answered is: how do off-the-shelf approaches and `EnvEx` perform when computing return levels for data featuring stationarity, non-stationarity, and serial

correlation, with different storm-peak WB3P shapes? The experiments isolate the effects of non-stationarity, serial correlation, and shape parameters on the performance of different statistical approaches. The type of statistical models we can choose between can be divided into common distributions and EVT distributions, all of which are stationary distributions. `EnvEx` is the only non-stationary approach. The following distributions are considered: the Normal, two-parameter Weibull (WB2P), three-parameter Weibull (WB3P), and three-parameter Gamma (G3) distributions are fitted to the full dataset. The Generalised Pareto (GP) distribution is fitted to peaks exceeding a constant threshold corresponding to the 98th percentile, while the Generalised Extreme Value (GEV) distribution is fitted to the annual block maxima and EnvEx to peaks exceeding a non-stationary threshold.

As stated in Section 1 we aim to estimate the return level $q_{\mathcal{T}}$. Each compared approach yields a model that can simulate *iid* values. Therefore, we let every model produce 100 years worth of data and choose the maximum. We repeat this many times to establish predictive distributions of $\mathcal{T}$-year maxima. The "truth" against which each approach is compared is defined as the $\mathcal{T}$-year quantile from the N^{sim} years of data produced for each scenario. This is essentially the q_3 estimator applied to the true underlying population.

Each dataset is scaled so that its maximum over the N^{sim} years equals 10. Further, deviations from the "truth" are displayed in relative terms by means of a percentage deviation. The percentage deviation Δ_v is calculated for each value v comparing it against the ground truth T (Eq. 15). T is different for each shown variable. For the box and whiskers plots the deviation is computed from comparing predicted $\mathcal{T}$-year maxima with $E[.]$ from the respective maxima in the data. Predicted return levels are compared with their respective counterpart in the data.

$$\Delta_v = \frac{v - T}{T} \cdot 100 \tag{15}$$

5. Quantifying uncertainty

Uncertainty on the model predictions is quantified by resampling the original N^{sim} years in intervals of $\mathcal{T}$-year periods. Additionally, we sample for each set of parameters from each fitted distribution function multiple times (e.g. 100 Monte Carlo simulations) approximating the aleatory uncertainty. For the choice of the threshold level we implemented a k-fold cross-validation scheme (here $k = 5$). For each iteration a cost-function is computed quantifying the deviation from the true values that had to be predicted. Since we are interested in the extremes the cost-function is applied to typically the highest 10% of the values. We apply this to a range of thresholds and sample around the minimum of the cost-function. Also a constant threshold can be chosen which is computationally more efficient. The cost-functions C_1 and C_2 are pre-defined (Eq. (16)–(17)) and represent a combination of the mean error (bias) and the root mean square error, and the mean error and the mean absolute error, respectively.

$$C_1\,(|u) = \left[\frac{\sum_n (p - t)}{n}\right]^2 + \frac{\sum_n (p - t)^2}{n} \tag{16}$$

$$C_2\,(|u) = \frac{\sum_n (p - t)}{n} + \frac{\sum_n (|p - t|)}{n} \tag{17}$$

Both cost functions yield similar minima, with the lowest bias often coinciding with a comparable low random error and the cost-function minimum. However, in some instances one is more noisy than the other and having a second cost-function helps. An example of the cross-validation output used for section 7 is shown in Supplementary Figure 6.

6. Application to synthetic data

6.1. Stationary distributions vs `EnvEx` vs truth

We estimated the 100 year return level for all described methods from 50 years worth of data for categories 1-4 including all WB3P shapes ($C_{const} = 1, ..., 6$) and $\lambda = 1$ where serial correlation applies.

In Fig. 5 a) the performance for i.i.d. data is displayed. As expected, the WB3P is performs well for all shapes as it is the data producing model. The EVT distributions predict the 100 year return value with a minor bias suggesting max-stability and that the extremes converge to an EVT distribution. `EnvEx`'s additional degrees of freedom that otherwise would allow for covariate effects if present do not seem to overfit and do not introduce a notable additional bias. In fact, the non-stationary threshold model is essentially penalised to a flat line and `EnvEx` reduces to a stationary peak over threshold model.

The other distributions struggle with deviations of close to 50% for long tails. When the tail in the data becomes shorter, the WB2P, gradually improves its estimate. The G3 distribution, featuring a heavy tail, can deal well with a long WB3P tail ($C_{const} = 1$) but predicts confidently erroneous and biased estimates for shorter tails. The Normal distribution has a sweet spot where the Weibull tail yields a distribution function similar to a Normal distribution ($C_{const} = 3$) but develops negative and positive biases for longer and shorter tails, respectively. The uncertainty is generally largest for a the long-tail scenario where comparably few values can inform the tail. The uncertainty thus decreases as the tail becomes shorter.

When a covariate effect is introduced by means of a seasonal cycle in panel b), an additional bias is introduced in the estimates of all stationary standard distributions. The main reason for this is that these models, and now also the WB3P, are misspecified and do not account for a seasonal cycle. The misspecification arises because annual maxima are seasonally constrained, whereas stationary distributions assume year-round homogeneity. Especially for the long and short tails, this bias is considerable with a confident underestimation of about 60% of the WB3P and a shape of 1. Compared to the stationary case, the number of values populating the tail are now even fewer, exacerbating the lack of information for the shape of the tail. As this is true in general, also for the EVT distributions and `EnvEx`, the uncertainties are largest for the long tails and reduce with an increasing shape. Misspecified models are thus increasingly certain about their biased results. `EnvEx`, which can take into account the seasonal cycle, demonstrates an efficient usage of the data and indicates no bias. The shift from a negative bias to a positive bias may stem from the fact that the body of the empirical density function grows much wider with increasing shape parameter. It becomes even bimodal for very short tails essentially distributing mass almost evenly from values 2-8 (Supplementary Figure 2). With a heavy tail (WB3P shape of 1) there is hardly any mass beyond a value of 1 (Fig. 4). EVT distributions are not affected by the seasonal cycle once the threshold or block size is sufficiently large and the stochastic process

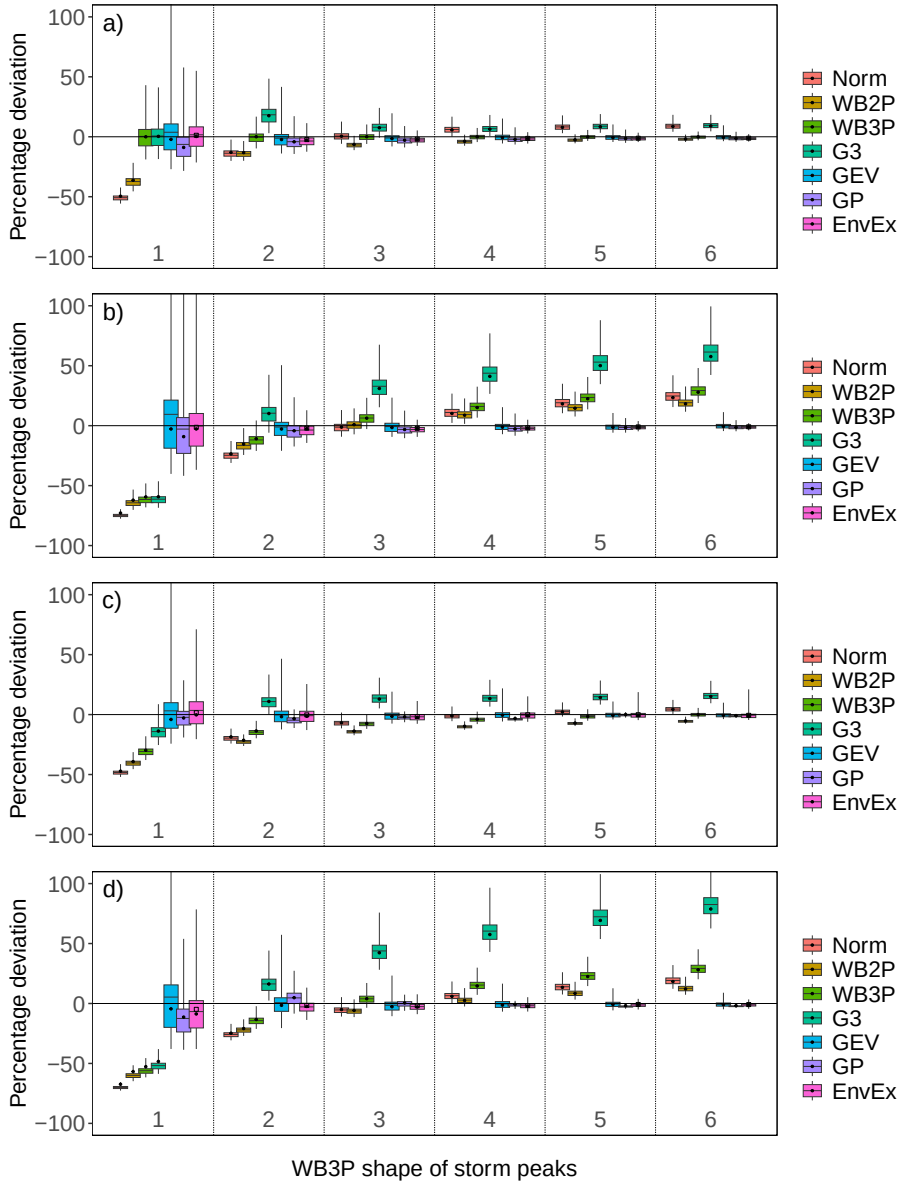


Figure 5: Boxplots of the percentage deviations of the predicted 100 year maxima from the expected value of the true 100 year maxima return value for all methods evaluated for different WB3P storm peak shapes 1-6 for categories 1-4 in the respective panels a)-d), with a) *iid* WB3P values, b) seasonal dependence, c) serial correlation, d) seasonal dependence and serial correlation. The boxes represent the inter-quartile range (25% - 75%) and the whiskers include 99% (0.5% - 99.5%). The dots show the percentage deviation from the predicted return level estimator $\tilde{q}_2'$ compared to the corresponding value from the original data.

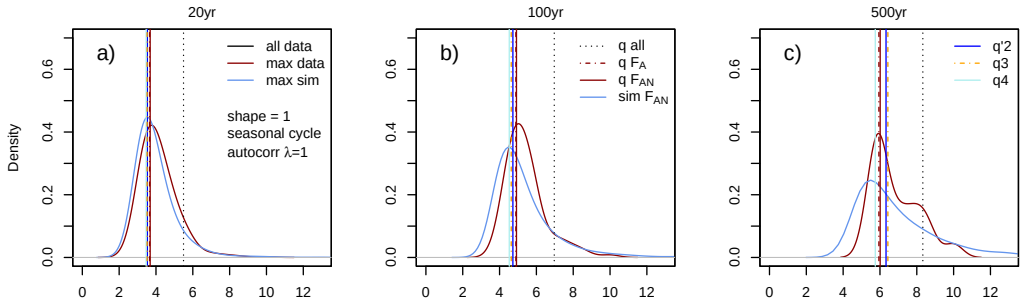


Figure 6: All data (black) and block maxima distributions (dark red) vs predictive distributions simulated from a seasonal `EnvEx` model (light blue). Different return level estimators are applied to the original correlated data (q_{all}, black dotted), the original peak data (q_{F_A} the quantile based in annual maxima and $q_{F_{AN}}$ the quantile based on N-year maxima (here $\mathcal{T}$-year maxima) in dark red) and the predictive distributions (q_2', q_3, q_4).

that is effectively stationary. A more detailed elaboration on the effect of covariates on the global return level can be read in Supplementary Section 1.

Panel c) illustrates the effect of serial correlation on the return value estimates. The stationary EVT approaches and `EnvEx`, that remove the serial correlation by design, produce reasonable estimates. For long tails, WB3P shows large deviations. Apart from G3, all standard distributions display a negative bias for WB3P shapes below $C = 4$. This behavior likely stems from the distribution functions being informed mainly by lower values between the peaks. G3 confidentally overestimates again for any WB3P shape larger than $C = 1$. This result adds some nuances to Mackay et al. [18] who found a positive bias when serial correlation is neglected. Section 6.4 takes a closer look at this feature and a more thorough elaboration can be read in Supplementary Section 1.

The combined effect of serial correlation and a seasonal cycle is visible in (panel d). From all tested methods, reliable estimates can only be retrieved by the stationary EVT distributions or the non-stationary `EnvEx` approach. Relying on a stationary Weibull fit or any other misspecified model can introduce vastly erroneous return values.

6.2. Stability across a range of return periods

Small errors in tail characteristics may result in substantial differences for long return periods. We check the stability of `EnvEx` and validate the predictive distributions of maxima and return level estimators for long return periods. Specifically, we consider the $\mathcal{T}$-year return periods of $\mathcal{T} = 20, 100, 500$ which represent the respective data availability ratios 20/50, 2, and 10 of the provided data of 50 years.

We focus on the most challenging case from the previous section, namely to model extremes of serial correlated and covariate dependent data from a long tailed (WB3P $C_{const} = 1$) distribution (Fig. 5 d)). `EnvEx` manages to predict the 20-year maxima seemingly well (Fig. 6a). For higher return periods the predicted distribution exhibits a slight negative bias which does not necessarily yield a negative bias in the return values (Fig. 6b-c). In our experiments, q_3 exhibits

the largest and q_4 the lowest return value consistent with Jonathan et al. [16]. The biases of all estimators compared to the truth differ from the estimator centered view in Jonathan et al. [16] because the model and the resulting predictive distributions are slightly biased. As evident in Fig. 5, the biases are only a few percent and practically negligible for downstream problems with the strongest bias for very long tails.

6.3. Effect of comparing apples and pears

Computing $q_{\mathcal{T}}$ applying $F^{-1}_{X^A}(1-\frac{1}{\mathcal{T}})$, $F^{-1}_{X^{SP}}(1-\frac{1}{\lambda\mathcal{T}})$, and $F^{-1}_{X}(1-\frac{1}{N\mathcal{T}})$ for *iid* data is equally valid and considers the same quantity (*the apple*). In various offshore standards, however, often methods like the initial distribution method are allowed which fits a model to all data and attempts to estimate $q_{\mathcal{T}}$ based on $F^{-1}_{X}(1-\frac{1}{N\mathcal{T}})$ with non-*iid*, correlated data (*the pear*). All models that do not take into account serial correlation are in principal designed to model the pears rather than the apples. A model independent effect of this can be observed in Fig. 6. The black dotted line indicates the quantile q_{all} for each return period $\mathcal{T}$ taken from all original available serially correlated data. It is significantly higher than the quantile based on only all the independent annual peaks ($q_{F_A} = F^{-1}_{X^A}(1-\frac{1}{\mathcal{T}})$). In the absence of serial correlation, the two quantiles are exactly the same. However, serial correlation reduces the number of independent values and hence the effective sample size. If treated as *iid*, this can lead to an over-representation of values in the tail and, hence, to a positive bias, as also noted by Mackay et al. [18]. Assuming a population could be approximated perfectly using any distribution, not taking into account serial correlation will still lead to a potentially massively biased return level estimate (more in Supplementary Section 1).

Taking into account the model-fitting process and whether a specific quantile is obtained from the fitted model or from a simulated predictive distribution, the results differ. In our experiments using the WB3P stationary model, we observe rather an underestimation than an overestimation when serial correlation is introduced (5c). The same is true when simply retrieving a quantile from the fitted distribution. If we were in fact aiming for q_{all} (example in Fig.6), the underestimation would be even more severe.

6.4. The effect of varying the degree of serial correlation

We investigate how the degree of serial correlation affects the previously discussed outcomes. We chose $\lambda = 0.5$ and $\lambda = 2$ comparable to Mackay et al. [18]. The return level and the deviation from the truth are handled in the same way as in the previous chapter, i.e. simulating maxima for the given return period and comparing the results to the truth known from the synthetic data.

Interestingly, the simulated return level estimates from the standard distributions (Norm, WB2P, WB3P, G3) move closer to the truth with increasing λ (Fig. 7). This is due to the design of the power storm model. A greater λ automatically means an increase of mass in the density of simulated values toward the right, an increased serial correlation for high values with more independent low values. A lower value of λ means a shift of mass toward the left with increasing serial correlation for low values and more independent peaks. Thus a high λ leads to a more populated tail and improved estimates which can partly compensate the effect of the misspecification of the models especially for long tailed data. `EnvEx` and the stationary EVT distributions are unaffected by the change in serial correlation (not shown), as they by definition only use declustered and uncorrelated peaks. If GP is applied to unclustered data a positive bias

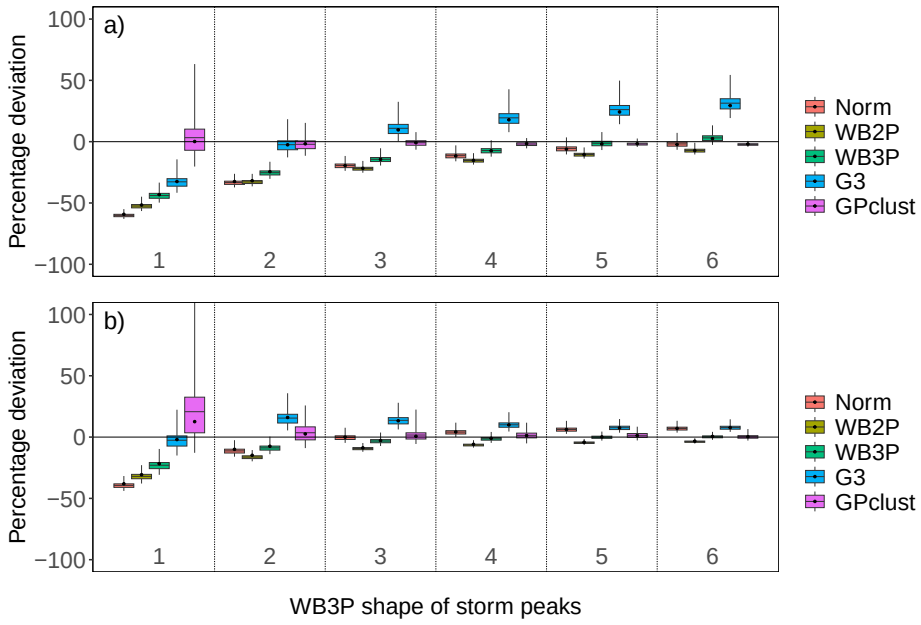


Figure 7: Same as Fig. 5 c) but for $\lambda = .5$ in panel a) and for $\lambda = 2$ in panel b). GPclust applies GP to serially correlated data above a threshold.

becomes evident when modelling long tailed data, which is similar to the observations in Mackay et al. [18].

7. `EnvEx` applied to hindcast data

We finally apply `EnvEx` to hindcast data at a location in the central North Sea demonstrating the full workflow as illustrated by the equations Eq. (1)-(3). First, we pick the primary variable of interest to be H_s to choose "storms" and their peaks. The threshold model is defined to be depending on two covariates, "day of the year" (doy) and "peak direction" (Pdir). Dependence between those covariates is assumed and we choose tensor product smoothers (Sec. 2.1) to compensate. The varying threshold is clearly visible and effectively allows for storms in all seasons and all directions according to their relative magnitude (Fig. 8). Storms that are violating the decorrelation time scale (24h), are combined to one storm. The storms are multivariate datasets consisting of multiple contemporaneous response variables and covariates.

The dataset of storms is resampled with replacement, resulting in 100 storm datasets, each containing the same number of storms as the original NORA3 dataset. We decide on the response variables $\mathbf{X} = [Hs, U10]$ and define the non-stationary extreme threshold, occurrence, and exceedence models for each variable to focus only on the extreme peaks and storms. We perform a cross-validation according to Eq. 16 for each variable to determine the most suitable threshold range for the subsequent model fitting procedure. An example of the cross-validation result for *Hs* is shown in Supplementary Figure 6. A hockey stick like behavior is often observed where the placement of the minimum seems to be related to the tail characteristics, i.e. a long

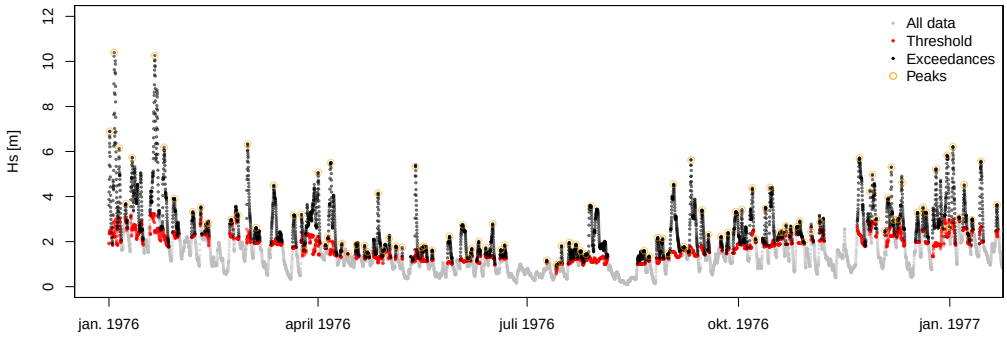


Figure 8: Subset of a NORA3 hindcast *Hs* timeseries in the Central North Sea illustrating a non-stationary threshold with storm and storm peak picking. All data above the storm-picking threshold are black, data below is depicted in grey. The threshold is red and storm peak values are indicated with circles.

tail requires a higher threshold than a shorter tail.

Choosing a threshold range, all models are fit to all response variables to arrive at a complete set of non-stationary marginal models from which we predict the $q2$, $q3$, $q4$ return level estimators. We estimate a complete set of HT2004 models and investigate the stability of the HT2004 parameters for a range of thresholds for each variable (Supplementary Figure 3), as well as the model fit (Supplementary Figure 4). In this example, we fit our multivariate extremes models using a subset of the marginal model predictions, i.e. 100 year return levels, to subsequently predict their multivariate counterparts from the other response variable. This creates an information rich set of predictions that can be dissected (Fig. 9). The results can be further filtered or the modelling tailored to the question at hand. It becomes obvious that the choice of the subregion evidently changes the predicted distribution of joint extremes (Fig. 9).

We choose the set of 200 storms from the region B+C (Fig. 9c) from which we consider *Hs*, *U10*, *Pdir*, and *doy* for the matching procedure. The simulated storms are scaled and rotated (Supplementary Figure 5). The resulting storm trajectories together with their historic counterpart are displayed in Fig. 10. Most of the matched historic peaks coincide with very high waves. The steepness based scaling fulfills its purpose so that the trajectories seem to seamlessly continue the observational point cloud. Alternatively, one could also predict *Tm02* with the HT2004 model and use it together in the matching procedure. Examples in the affiliated `github` repository show how to do this.

Finally, we integrate along all chosen storms to arrive at a distribution for the individual maximum wave height and the respective return values. Since we propagated all necessary covariates we can investigate covariate sub-spaces, i.e. directional sectors (Fig. 11a). The return level estimates for the sector with the lowest waves is smaller than the estimate for the sector with the highest waves (Fig. 11b). The omni-directional estimate is highest. In this framework, this can never be violated by design because all estimates are taken from the same model and the same

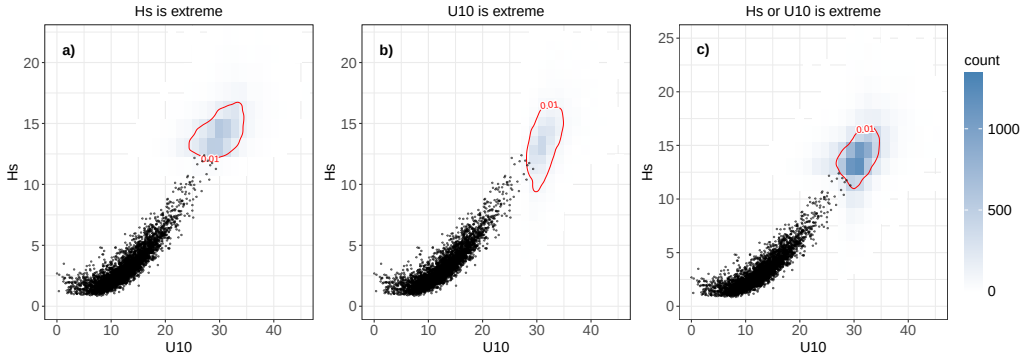


Figure 9: HT2004 multivariate predictions for 100 year maxima from *Hs* and *U10*. Joint maxima included in the original NORA3 dataset are represented by black dots, predictions for different HT2004 regions are displayed in blue. a) *Hs* is large = region C, b) *U10* is large = region B, and c) *Hs* or *U10* are large = region B+C. For the definition of the regions please refer to Fig. 3. Other subregions can also be chosen if desired. The red contour encompasses 99% of the simulations.

set of simulations.

8. Summary and conclusion

We developed a framework called `EnvEx` to model non-stationary and joint extreme events. The workflow and validity was demonstrated against synthetic data and in a metocean context using wave hindcast data. `EnvEx` is built on recent developments in extreme value statistics. Additionally, we investigated the effect of serial correlation, tail properties, and covariate effects on the prediction of return levels. Not including these effects can lead to severe negative biases for very long tails and positive biases for short tails. Simple, vanilla EVT distributions manage these challenges when applied correctly but do not consider extremes in less severe seasons or other covariate sub-spaces. `EnvEx` includes non-stationarity and performs well for all tested cases. We demonstrate that computing $q_{\mathcal{T}}$ applying $\mathbb{P}(X \leq q_{\mathcal{T}})$ to correlated data as allowed in offshore standards will lead to positive biases, especially for long tails, even if the distribution of the underlying population is known perfectly.

For `EnvEx`, we chose a modular HT2004 framework because it makes extensions straight forward. In the future, we plan to include other covariates such as climate trends as well as testing various templates for non-stationary distributions for the bulk of the data. The framework and libraries we applied allow parametric models and Generalised Additive Mixed Models (GAMMs) which is a great benefit in terms of flexibility, particularly when smoothness is not an appropriate assumption. By making the implementation openly available, we hope to facilitate the broader adoption of these statistical methods for the benefit of both the scientific community and practitioners in industry and the public sector.

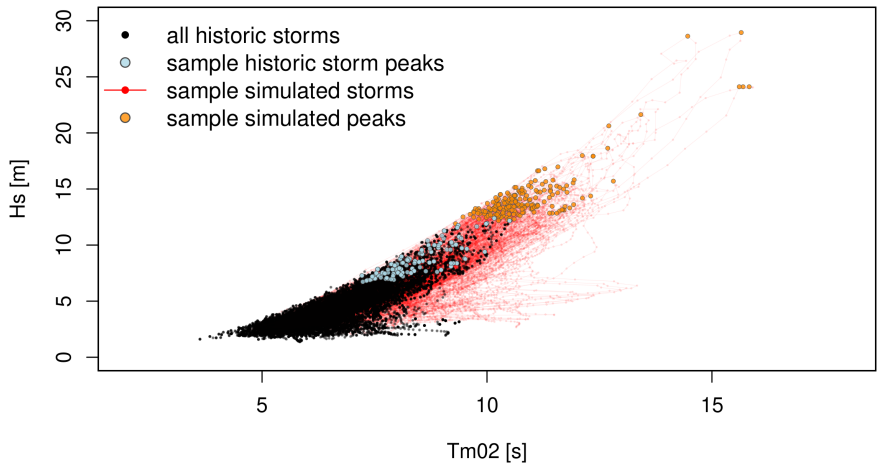


Figure 10: Simulated storm peaks (grey circles filled with orange) and storm trajectories (red lines) together with all historic storms. The matched historic storm peaks used for the storm simulations are lightblue with grey edge.

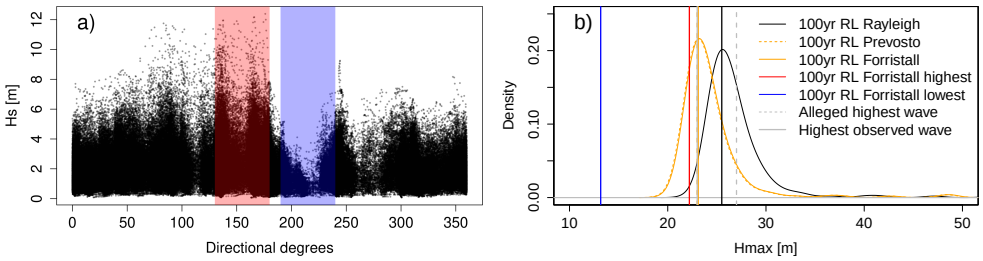


Figure 11: a) Directional sectors prevalent for high (red) and low (blue) wave events. b) Densities of simulated 100 year maxima H_{max} events for different individual wave height distributions. Vertical lines in grey indicate observed individual maxima wave events from an 18 year observational record processed in Malila et al. [19] and used in Bohlinger et al. [3] for modelling extreme waves. Other vertical lines and 100 year return levels for all sectors and only sectorial return levels from a) in red and blue.

CRediT authorship contribution statement

PB: Conceptualization, Methodology, Code, Investigation, Formal analysis, Data curation, Visualization, Original draft, Writing – review & editing, Funding acquisition. **PJ:** Conceptualization, Methodology, Theoretical analysis, Original draft, Writing – review & editing. **BY:** Code, Methodology, Writing – review & editing. **TE:** Methodology, Writing – review & editing. **JVB:** Methodology, Writing – review & editing. **ØB:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The NORA3 wave hindcast can be accessed at `https://thredds.met.no/thredds/myocean/ARC-MFC/hindcast-mywave3.html`. The EnvEx code including a README file demonstrating some examples can be accessed at `https://github.com/bohlinger/envex`.

Acknowledgements

We gratefully acknowledge the funding support from the project FORESEE (project no. 334188 from the Research Council of Norway).

References

[1] C. W. Anderson, D. J. T. Carter, and P. D. Cotton. Wave climate variability and impact on offshore design extremes. Technical report, Shell International, 2001.

[2] E.M. Bitner-Gregersen and S. Haver. Joint Environmental Model for Reliability Calculations. *Proceedings of the 1st International Offshore and Polar Engineering Conference, Edinburgh, United Kingdom, August 11-15, 1991*, 1991.

[3] P. Bohlinger, T. Economou, O.J. Aarnes, M. Malila, and Ø. Breivik. A general framework to obtain seamless seasonal–directional extreme individual wave heights—showcase ekofisk. *Ocean Engineering*, 270:113535, 2023. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2022.113535.

[4] Ø. Breivik, A. Carrasco, H. Haakenstad, O.J. Aarnes, A. Behrens, J.-R. Bidlot, J.-V. Björkqvist, P. Bohlinger, B.R. Furevik, J. Staneva, and M. Reistad. The impact of a reduced high-wind charnock parameter on wave growth with application to the north sea, the norwegian sea, and the arctic ocean. *Journal of Geophysical Research: Oceans*, 127(3): e2021JC018196, 2022. doi: https://doi.org/10.1029/2021JC018196.

[5] L. Brett, C. J. White, D. I. V. Domeisen, B. van den Hurk, P. Ward, and J. Zscheischler. Review article: The growth in compound weather and climate event research in the decade since srex. *Natural Hazards and Earth System Sciences*, 25(8):2591–2611, 2025. doi: 10.5194/nhess-25-2591-2025.

[6] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5): 1190–1208, 1995. doi: 10.1137/0916069.

[7] S. Coles. An introduction to statistical modeling of extreme values, 2001.

[8] K. Ewans and P. Jonathan. Evaluating environmental joint extremes for the offshore industry using the conditional extremes model. *Journal of Marine Systems*, 130:124–130, 2013. doi: 10.1016/j.jmarsys.2013.03.007.

[9] G. Feld, D. Randell, Y. Wu, K. Ewans, and P. Jonathan. Estimation of storm peak and intrastorm directional–seasonal design conditions in the north sea. *Journal of Offshore Mechanics and Arctic Engineering*, 137(2):021102, 04 2015. ISSN 0892-7219. doi: 10.1115/1.4029639.

[10] G.Z. Forristall. How should we combine long and short term wave height distributions? *Proceedings of OMAE08, 27th International Conference on Offshore Mechanics and Arctic Engineering, June 15-20, 2008, Estoril, Portugal*, pages 987–994, 2008.

[11] H.F. Hansen, D. Randell, A.R. Zeeberg, and P. Jonathan. Directional–seasonal extreme value analysis of North Sea storm conditions. *Ocean Engineering*, 195:106665, 2020. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2019.106665.

[12] J.E. Heffernan and J.A. Tawn. A Conditional Approach for Multivariate Extreme Values (with Discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(3):497–546, 07 2004. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2004.02050.x.

[13] L.H. Holthuijsen. *Statistics*, page 56–105. Cambridge University Press, 2007.

[14] P. Jonathan and K. Ewans. Statistical modelling of extreme ocean environments for marine design: A review. *Ocean Engineering*, 62:91–109, 2013. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2013.01.004. URL `https://www.sciencedirect.com/science/article/pii/S002980181300019X`.

[15] P. Jonathan, K. Ewans, and D. Randell. Joint modelling of extreme ocean environments incorporating covariate effects. *Coastal Engineering*, 79:22–31, 2013. doi: 10.1016/j.coastaleng.2013.04.005.

[16] P. Jonathan, D. Randell, J. Wadsworth, and J. Tawn. Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised pareto distribution. *Ocean Engineering*, 220:107725, 2021. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2020.107725.

[17] E. Mackay and P. Jonathan. Assessment of return value estimates from stationary and non-stationary extreme value models. *Ocean Engineering*, 207:107406, 2020. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2020.107406.

[18] E. Mackay, G. de Hauteclocque, E. Vanem, and P. Jonathan. The effect of serial correlation in environmental conditions on estimates of extreme events. *Ocean Engineering*, 242: 110092, 2021. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2021.110092.

[19] M.P. Malila, P. Bohlinger, S. Støle-Hentschel, Ø. Breivik, G. Hope, and A.K. Magnusson. A nonparametric, data-driven approach to despiking ocean surface wave time series. *Journal of Atmospheric and Oceanic Technology*, 39(1):71 – 90, 2022. doi: 10.1175/JTECH-D-21-0067.1.

[20] D. Randell, K. Turnbull, K. Ewans, and P. Jonathan. Bayesian inference for nonstationary marginal extremes. *Environmetrics*, 27(7):439–450, 2016. doi: https://doi.org/10.1002/env.2403.

[21] M. Speers, D. Randell, J. Tawn, and P. Jonathan. Estimating metocean environments associated with extreme structural response to demonstrate the dangers of environmental contour methods. *Ocean Engineering*, 311:118754, 2024. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2024.118754.

[22] R. Towe, E. Zanini, D. Randell, G. Feld, and P. Jonathan. Efficient estimation of distributional properties of extreme seas from a hierarchical description applied to calculation of un-manning and other weather-related operational windows. *Ocean Engineering*, 238:109642, 2021. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2021.109642.

[23] R. Towe, D. Randell, J. Kensler, G. Feld, and P. Jonathan. Estimation of associated values from conditional extreme value models. *Ocean Engineering*, 272:113808, 2023. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2023.113808. URL `https://www.sciencedirect.com/science/article/pii/S0029801823001920`.

[24] R. Towe, E. Ross, D. Randell, and P. Jonathan. covxtreme : Matlab software for non-stationary penalised piecewise constant marginal and conditional extreme value models. *Environmental Modelling & Software*, 177:106035, 2024. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2024.106035.

[25] S.N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B*, 73(1):3–36, jan 2011. doi: 10.1111/j.1467-9868.2010.00749.x.

[26] S.N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall / CRC texts in statistical science. Chapman and Hall/CRC, second edition edition, 2017. ISBN 978-1-4987-2833-1.

[27] S.N. Wood, N. Pya, and B. Säfken. Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516):1548–1563, oct 2016. doi: 10.1080/01621459.2016.1180986.

[28] B.D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019. doi: 10.1080/01621459.2018.1529596.

[29] B.D. Youngman. evgam: An r package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26, 2022. doi: 10.18637/jss.v103.i03.

[30] B.D. Youngman and T. Economou. Generalised additive point process models for natural hazard occurrence. *Environmetrics*, 28(4):e2444, 2017.