Multi-Dimensional Predictive Analytics for Risk Estimation of Extreme Events

(Submitted to the IEEE High-Performance Computing, Data and Analytics Conference `www.hipc.org/hipc2016`)

Laks Raghupathi*, David Randell†, Emma Ross†, Kevin C. Ewans* and Philip Jonathan†
*Shell India Markets Private Ltd., Bangalore, 560048, India
Email: l.raghupathi@shell.com
†Shell Research Ltd., Manchester, M22 0RR, United Kingdom

***Abstract*—Modelling rare or extreme events is critical in many domains, including financial risk, computer security breach, network outage, corrosion and fouling, manufacturing quality and environmental extremes such as floods, snowfalls, heat-waves, seismic hazards and meteorological-oceanographic events like extra-tropical storms, hurricanes and typhoons. Statistical modelling enables us to understand extremes and design mechanisms to prevent their occurrence and manage their impact.**

Extreme events are challenging to characterise as they are, by definition, rare and unusual even in a big data world. The frequency and extent of extreme events is typically driven by both primary attributes (*dependent* variables) and secondary attributes (*independent* variables or *covariates*). Studies have shown that improved inference can be gained from including covariate effects in predictive models but this inclusion comes at a heavy computation cost.

In this paper, we present a framework for risk estimation from extreme events that are non-stationary; i.e., they are dependent on multi-dimensional covariates. The approach is illustrated by estimation of offshore structural design criteria in a storm environment non-stationary with respect to storm direction, season and geographic location. The framework allows consistent assessment of structural reliability with thorough uncertainty quantification. The model facilitates estimation of risk for any combination of covariates, which can be exploited for improved understanding and ultimately optimal marine structural design. The computational burden incurred is large, especially since thorough uncertainty quantification is incorporated, but manageable using slick algorithms for linear algebraic manipulations and high-performance computing.

***Keywords*-large-scale extremes; statistical modelling; covariates; uncertainty quantification;**

I. Background

Extreme events from natural phenomena and human activity are of global concern because they have potentially devastating consequences for society; recent international financial crises and extreme climate events bear testimony. Yet until recently, the literature has tended to focus on understanding the mean attributes of physical systems and their typical variation. Extreme events, by definition, are rare, difficult to study and even harder to predict [1]. Occurrences of extreme events may be inter-dependent in space or time, for example due to common or related underlying causes [2], [3], [4].

Statistical modelling enables us to understand risk due to extremes and design mechanisms to manage their potential impact. *Extreme value theory* (EVT) [5] describes statistical distributions and stochastic processes appropriate for the modelling of extreme phenomena, motivated by asymptotic arguments. A review from an ocean engineering perspective is presented in [6]. According to EVT, for a variable with any underlying *max-stable* distribution, exceedances over a (sufficiently high) threshold follow the *generalised Pareto* (GP) distribution; a result sometimes known as the *law of small numbers*. [7]. The GP distribution is used to model extreme ocean waves [6], extreme corrosion and fouling [8] and extreme financial markets [9]. Inferences motivated by sound combinations of statistical and physical understanding result in better characterisation of extremes [10].

To characterise extreme events, we use data not only for rare values of primary attributes, but also for secondary attributes (or *covariates*), since the characteristics of extremes typically vary as a function of secondary attributes. Studies [11] have shown that inclusion of covariates in a predictive model improves estimation compared with estimation from a model which ignores them [12], [13]. However inclusion of complex covariate dependence complicates inference and introduces a substantial computational burden.

In this paper, we present a framework for risk estimation from extreme events that are dependent on multi-dimensional covariates. The approach is illustrated in application to the estimation of extreme *metocean* (meteorological-oceanographic) storm environments conditional on storm direction, season and geographic location. [13] discusses large-scale marginal spatio-directional extreme value modelling using efficient statistical algorithms and parallel computing environments. Marginal return values for storm severity (measured using significant wave height, H_S) for locations in the Gulf of Mexico (GoM) within a large spatial neighbourhood are estimated, accounting for spatial and storm directional variability of peaks over threshold. Applications to other ocean basins (such as the South China Sea, SCS) require the incorporation of a seasonal covariate, since storm systems exhibit strong *seasonal* variation in those basins. For example, the SCS regional climate is characterised by northeast and southwest *monsoons* and passing *typhoons* [14]. To quantify an environment such as SCS, therefore, a four-dimensional spatio-directional-seasonal model is required. Statistical modelling of directional and seasonal effects over a spatial domain is the focus of the current paper.

Our approach allows consistent assessment of the reliability of marine structures with respect to extreme environments with thorough uncertainty quantification. The model facilitates estimation of risk for any combination of covariates, which can be exploited for improved understanding of the environment and its effects on offshore and coastal structures, and ultimately for optimal structural design. The computational burden incurred is large, but manageable using a combination of slick linear algebraic manipulations [15] and high-performance computing.

II. Approach

A. Data Cube

We describe the multi-dimensional predictive analytic framework using a *data cube*. A data cube model in typical Online Analytical Processing (OLAP) applications is a means of modelling and visualisation in multiple dimensions [16]. We need to consider variables representing extremes modelled with respect to covariates of different dimensions. Figure 1 illustrates this. At the simplest *0D* level, we have a stationary model with no covariates; we expect this model not to describe physical reality well when covariate effects are present; it will provide poor inference in general. On the next level, we have a set of one-dimensional models for which extremes are non-stationary with respect to a single covariate (e.g., *1D-Drc*, *1D-Ssn*). In the metocean context, such models typically provide significant improvements over the stationary model. For example, real improvements in structural reliability for no additional cost are achievable if covariate effects are incorporated in the specification of design criteria in general, e.g. by: a) designing structures with different strengths in different directions; b) locating critical modules in the most benign environments; and c) performing operations of limited duration during the more benign periods, all consistent with other safety requirements for the offshore structure.

Extending to two dimensions, we might consider a directional-seasonal model (*2D Drc-Ssn*) which characterises directional and seasonal variation [17]. Predictions will be different from those using a directional model, if seasonal variation is an important driver of extremes. Characterising this extra variability is particularly useful in assessing the risk of offshore operations of limited duration, for example, where directional variation is also expected.

The models described so far address a single location but we can also seek to improve predictions by exploiting data from a neighbourhood of locations. When limited sample size is a concern, using multiple locations in inference is generally useful. Sample size is generally problematic in extreme value analysis, due to the fact that the inference strategy is motivated by asymptotic arguments. In the oceanographic context, only the largest values in the sample (for example, observations exceeding some threshold) should be used for extreme value modelling, but the sample size must also be sufficient for good empirical modelling. The amount of data available for analysis at a specific location is therefore often limited. Reducing the threshold risks invalidating asymptotic arguments of the underlying the modelling strategy. Increasing the threshold value reduces sample size still further. One pragmatic solution is to aggregate observations from a neighbourhood of spatial locations and analyse the pooled sample using the approach known as *site pooling*. However, neighbouring locations in general have different extreme value characteristics, and observations from neighbouring locations are typically not *statistically independent*. Naive analysis of spatial extremes over spatial neighbourhoods can lead therefore to erroneous inferences. Recent developments in statistical modelling of extreme ocean environments offer alternative solutions to naive site pooling. [18], [13] for example are able to characterise distributions of storm severity over spatial locations giving practitioners a *consistent* approach to extreme value analysis, and consistent design estimates *across* the domain of interest.

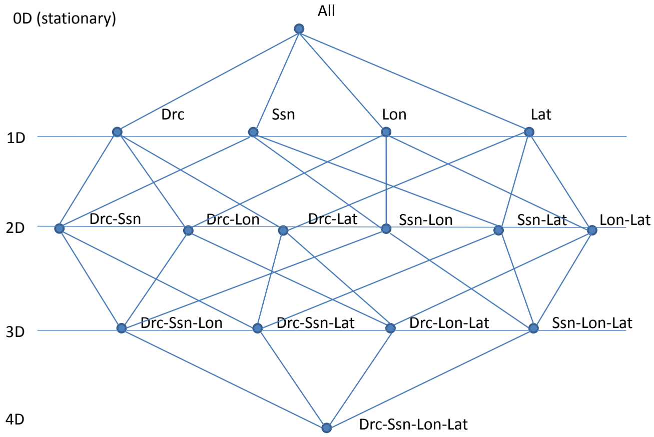


Figure 1. *Data cube* model representation incorporating direction (*Drc*), season (*Ssn*), longitude (*Lon*) and latitude (*Lat*) covariates. Inferences possible using any combination of covariates. The simplest 0D cuboid represents a model stationary with respect to covariates, whereas the 4D cuboid represents a model with all covariate effects incorporated.

In the existing literature on non-stationary marginal extreme value analysis, covariate descriptions of various complexities are used. Early efforts assumed simple linear regression-type relationships between parameters of the extreme value model and one or more covariates. More general descriptions including Fourier representations of periodic covariates [19], non-parametric approaches including splines [20] and Gaussian processes [21] followed. Multi-dimensional covariate descriptions were subsequently proposed using thin-plate splines [22] and tensor products of marginal spline bases, and form the basis of the current approach. These advances provide physically more appropriate models, but demand efficient computational tools for useful application. In general, little effort has been devoted to comparing and relating inferences from models using different covariates to derive the complete picture. The current work is an attempt to address this need. Below, we model the same physical variable as a function of different combinations of covariates, and compare inferences from different models. Notice that though in principle we can model *any* combination covariates shown in Figure 1, only a few subset amongst these may be of practical interest.

B. *Model Components*

Extending the work of [23] and [22], summarised in [13], we model storm peak significant wave height H_S^{SP}, defined as the largest value of significant wave height observed per location during the period of a storm event. At a given location, storm peak events are reasonably assumed to be statistically independent given covariates since they correspond to occurrences of independent atmospheric pressure fields. We assume that each storm event is observed at all locations within the neighbourhood under consideration. Thus for a sample $\{\dot{z}_i\}_{i=1}^{\dot{n}}$ of $\dot{n}$ values of H_S^{SP} observed at locations $\{\dot{x}_i, \dot{y}_i\}_{i=1}^{\dot{n}}$ with dominant wave directions $\{\dot{\theta}_i\}_{i=1}^{\dot{n}}$ and seasons $\{\dot{\phi}_i\}_{i=1}^{\dot{n}}$ at H_S^{SP} (henceforth *storm directions* and *storm seasons*), we proceed using the peaks-over-threshold approach as follows (cf. Figure 2)) on peaks.

We first estimate a *threshold* function ψ above which observations $\dot{z}$ are assumed to be extreme. The threshold varies smoothly as a function of covariates ($\psi \triangleq \psi(\theta, \phi, x, y)$) and is estimated using quantile regression. We retain the set of n threshold exceedances $\{z_i\}_{i=1}^{n}$ observed at locations $\{x_i, y_i\}_{i=1}^{n}$ with storm peak directions $\{\theta_i\}_{i=1}^{n}$ and seasons $\{\phi_i\}_{i=1}^{n}$ for further modelling. We next estimate the *rate* of occurrence ρ of threshold exceedance using a Poisson process model with Poisson rate ρ ($\triangleq \rho(\theta, \phi, x, y)$). Finally we estimate the *size* of occurrence of threshold exceedance using a GP model. The GP shape and scale parameters ξ and σ are also assumed to vary smoothly as functions of covariates, with ξ real and $\sigma > 0$. Positivity of GP scale is ensured throughout in the optimisation scheme. The GP shape parameter is unrestricted in the full optimisation, but limited to the interval (-0.5, +0.2) in the estimation of the GP starting solution.

This approach to extreme value modelling follows that of [20] and is equivalent to direct estimation of a non-homogeneous Poisson point process model [24], [6]. We emphasise that, in common with [20] and [6], we perform marginal non-stationary extreme value analysis across a grid of (dependent) spatial locations, accounting *marginally* for directional, seasonal and spatial variability in extremal characteristics. We further account for the effects of extremal spatial dependence between locations on inferences using a block bootstrapping scheme (cf. §II-E on computational aspects).

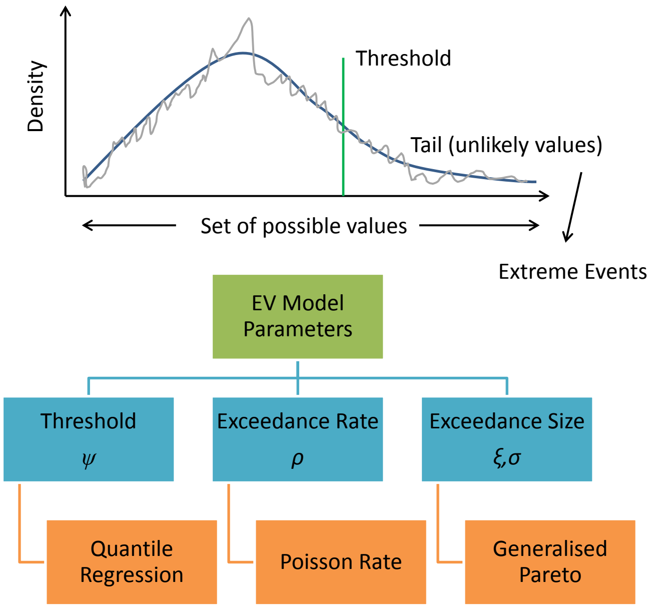


Figure 2. Model components: extreme value quantile threshold ψ, Poisson rate of threshold exceedance ρ, and size of threshold exceedance quantified by GP shape and scale ξ, σ. Each component is modelled as function of covariates θ, ϕ, x, y.

C. *Parameterising Covariates*

Applying the approach to whole-basin applications is methodologically straightforward, but computationally challenging. Physical considerations suggest we should consider parameters ψ, ρ, ξ and σ to be smooth functions of covariates θ, ϕ, x, y [25]. For estimation, this can be achieved by expressing the value of each parameter in terms of a linear combination of suitable basis functions for the domain D of covariates, where $D = D_\theta \times D_\phi \times D_x \times D_y$. Here $D_\theta = [0, 360)$ is the (marginal) domain of storm peak directions, $D_\phi = [0, 360)$ is the (marginal) domain of storm peak seasons (expressed as normalised days of 360-day calendar year), and D_x, D_y are the domains of x- and y-values (e.g. longitudes and latitudes) under consideration.

For each covariate (and marginal domain) in turn, we first calculate a B–spline basis matrix for an index set of size m ($m << n$) covariate values; potentially we could calculate the basis matrix for each of the n observations, but usually avoid this for computational and statistical efficiency. For instance in the case of D_θ, we calculate an $m_\theta \times p_\theta$ basis matrix B_θ such that the value of any function at each of the m_θ points in the index set for storm direction can be expressed as linear combination $B_\theta \beta_\theta$ for some $p_\theta \times 1$ vector β_θ of basis coefficients. Note that periodic marginal bases can be specified if appropriate (e.g. for D_θ and D_ϕ). Then we define a basis matrix for the four-dimensional domain D

using tensor products of marginal basis matrices. Thus

$$\mathbf{B} = \mathbf{B}_y \otimes \mathbf{B}_x \otimes \mathbf{B}_\phi \otimes \mathbf{B}_\theta \tag{1}$$

provides an $m \times p$ basis matrix (where $m = m_\theta m_\phi m_x m_y$, and $p = p_\theta p_\phi p_x p_y$) for modelling each of ψ, ρ, ξ and σ on the corresponding *spatio-directional-seasonal* index set of size m. Any of ψ, ρ, ξ and σ (η, say, for brevity) can then be expressed in the form $\eta = \mathbf{B}\beta$ for some $p \times 1$ vector β of basis coefficients. Model estimation therefore reduces to estimating appropriate sets of basis coefficients for each of ψ, ρ, ξ and σ. The value of any marginal p_η (i.e. p_θ, p_ϕ, p_x, or p_y) is equal to the number q_η of spline knots specified for periodic domains (e.g D_θ, D_ϕ), and to $q_\eta + d_\eta$ for aperiodic domains (e.g D_x, D_y), where d_η is the order of the B-spline function specified (always 3 in this work, so that spline functions are twice differentiable).

The roughness R of any function η defined on the support of the spline basis can be easily evaluated on the index set (at which $\eta = \mathbf{B}\beta$). For a one-dimensional (e.g. directional) spline basis, following [26], writing the vector of differences of consecutive values of β as $\Delta\beta$, and vectors of second and higher order differences using $\Delta^k\beta = \Delta(\Delta^{k-1}\beta)$, $k = 2, 3, ...$, the roughness R of η is given by

$$R = \beta' \mathbf{P} \beta \tag{2}$$

where $\mathbf{P} = (\Delta^k)'(\Delta^k)$ for differences of order $k = 1, 2, 3, ...$ (with appropriate modifications to preserve periodicity as necessary). For a spatio-directional spline basis, the penalty matrix P can be similarly defined using

$$\mathbf{P} = \mathbf{P}_y \otimes \mathbf{P}_x \otimes \mathbf{P}_\phi \otimes \mathbf{P}_\theta \tag{3}$$

We manage the considerable computational challenges of basin-wide extreme value modelling using a combination of generalised linear array methods (GLAM) and parallel computing as described in [13].

D. Simulation Steps

For clarity, we next briefly describe key modelling steps.

1) *Identification of peaks over threshold*: Storm peak significant wave heights H_S^{SP} are isolated from time-series of sea state H_S using the procedure described in [27]. Contiguous intervals of H_S above a low peak-picking threshold are identified, each interval now assumed to correspond to a storm event. The peak-picking threshold corresponds to a covariate-dependent (i.e., directional for *1D-Drc* case, directional-seasonal *2D-Drc-Ssn*) quantile of H_S with specified non-exceedance probability (referred to as *PP-NEP*), estimated using quantile regression. The maximum of H_S during the storm interval is taken as the storm peak significant wave height H_S^{SP} for subsequent extreme value modelling (cf. Figure 3). The values of other covariates θ, ϕ, x, y at the time of the storm peak significant wave height are referred to as storm peak values of those variables.
2) *Extreme value threshold estimation:* Threshold estimation in general is a difficult problem in extreme value analysis with recent emphasis not just on the threshold but also the uncertainty quantification on subsequent inferences such as return levels [28]. Accordingly, our approach avoids the selection of a specific threshold by estimating an *ensemble* [17] of non-stationary extreme value models corresponding to different threshold choices. Each of these in turn corresponds to a plausible choice of threshold non-exceedance probabilities (referred to as *EV-NEP*).
3) *Model estimation and computation of return value distributions*: We estimate an ensemble of Poisson models for the rate of occurrences of threshold exceedances, and the corresponding ensemble of GP models for the size of exceedances. We then simulate under the ensemble of models to produce realisations of storm peak characteristics corresponding to any return period of interest, and accumulate return value distributions.

Table I lists the peak-picking threshold *PP-NEPs* used for different model types in this work, and the sizes of resulting H_S^{SP} samples. Extreme value analysis was performed subsequently using exceedances of the non-stationary quantile with non-exceedance probability *EV-NEP* from the interval shown.

E. Complexity and Uncertainty Quantification

1) *Dimensionality*: As described above, the number of parameters to be estimated is the product of the number of parameters per marginal spline basis. In principle, inference therefore requires manipulation and inversion of very large matrices. Fortunately, as demonstrated in [13], generalised linear array methods (GLAMs) offer significant computational advantages.
2) *Roughness penalisation*: We estimate parameters using *penalised* maximum likelihood estimation, to avoid potential over-fitting of non-parametric models. The (negative log) likelihood function is penalised using a linear combination of parameter roughnesses with roughness coefficients for each model parameter ψ, ρ, ξ and σ. The choice of optimal penalty coefficients is determined by *block cross-validation*. Since each storm is observed as spatially-dependent event at all spatial locations, we define a storm block to be the set of occurrences of a particular storm at all locations for the purposes of both cross-validation (in estimation of model smoothness) and bootstrapping for uncertainty quantification (see below).
3) *Bootstrapping*: It is critical to quantify the uncertainty with which extreme value models are estimated. Resampling techniques such as *bootstrapping* can be used to estimate the uncertainty of model parameters

and estimates of return values and other structure variables [20], applicable when dependent data from neighbouring locations are used. In the current context, bootstrapping involves resampling the original sample with replacement to create a bootstrap resample. The whole modelling procedure, including parameter and return value estimation is then executed for the resampled data. By repeating this scheme for a large number of bootstrap resamples, we can quantify sampling uncertainty on parameter and return value estimates. Model fitting for each bootstrap resample is independent of fitting for all others. Computationally, bootstrapping is embarrassingly parallel, allowing efficient parallel implementation.

III. Application

We consider estimation of extreme value models for H_S^{SP} in the South China Sea using historical data from the recent SEAFINE database [29]. The storm characteristics are discussed in §I and e.g. in [14] . Both directional and seasonal covariates influence the rate of occurrence and size of extreme events (as can be seen in Figure 3); spatial variation is also anticipated. We outline 1D directional (*Drc*), 2D directional-seasonal (*Drc-Ssn*), 3D spatio-seasonal (*Ssn-Lon-Lat*) and 4D spatio-directional-seasonal (*Drc-Ssn-Lon-Lat*) models. To the best of our knowledge, this is the first demonstration of extreme value modelling with *nested multi-dimensional* covariates up to 4D.

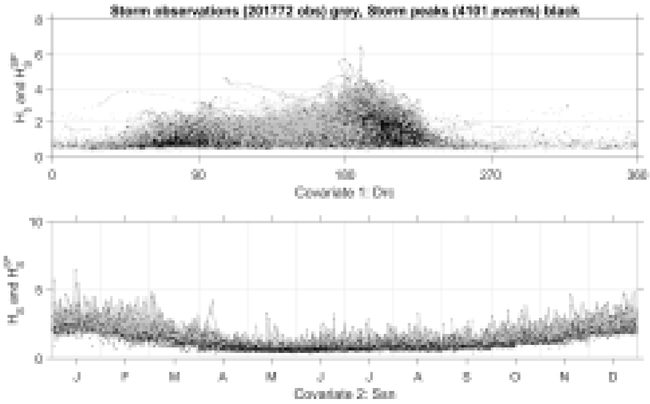


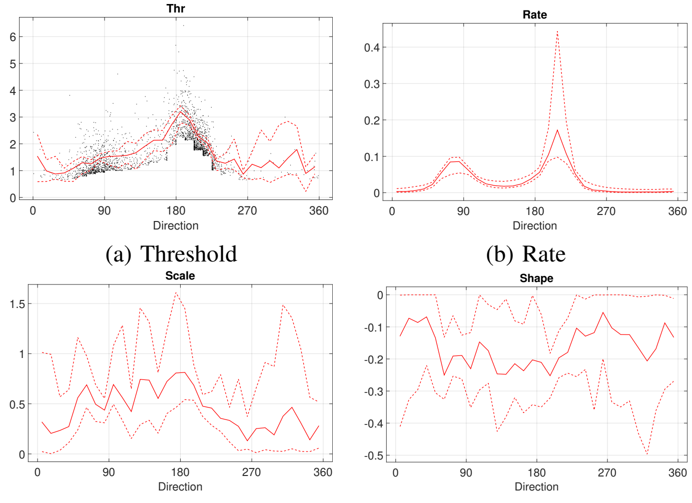
Figure 3. H_S^{SP} (black) and sea state H_S (grey) with direction θ (top) and season ϕ (bottom).

A. 1D directional model

H_S^{SP} occurrences were isolated as a function of the storm direction (θ) using the procedure described in §II-D. The most severe storms occur in the South (around 180^o, with North being 0^o clock-wise orientation; see upper panel of Figure 3). Parameters estimates (with bootstrap 95% uncertainty bands) of the multi-dimensional covariate model are shown in Figure 4. Threshold estimate and rate of occurrence reflect storm peak characteristics. GP shape is negative throughout. Model diagnostics are essential to demonstrate adequate model fit. Of primary concern is that the estimated extreme value model generates directional distributions consistent with observed storm peak data. Accordingly, Figure 5 compares return values from the model with those from the original sample, indicating a good agreement across directional sectors. Figure 6 shows the 100-year return value for H_S^{SP} directionally (black) or omni-directionally (red).

Model	*PP-NEP*	*#Str Peaks*	*EV-NEP*
1D-Drc	0.80	2355	[0.5,0.9]
2D-Drc-Ssn	0.60	4101	[0.5,0.9]
3D-Ssn-Lon-Lat	0.65	3156	[0.5,0.9]
3D-Drc-Lon-Lat	0.80	2916	[0.5,0.9]
4D-Drc-Ssn-Lon-Lat	0.65	3791	[0.5,0.9]

Table I
Parameters used for the extraction of storm peaks for different models.



(a) Threshold
(b) Rate
(c) GP scale
(d) GP shape

Figure 4. *1D Drc*: Directional model parameters showing median estimates (solid) and 95% uncertainty bands (dashed). (a) Threshold ψ, (b) rate ρ, (c) GP scale σ and (d) GP shape ξ.

B. 2D directional-seasonal model

Parameter estimates for a directional-seasonal model are shown in Figure 7. Threshold and rate of occurrence estimates suggest more extreme H_S^{SP} in the southern sector during winter months. Uncertainty bands from the ensemble model indicate reasonable estimates. Diagnostic plots for the 2D and all subsequent higher-dimensional models similar to Figure 5 indicate good model fit. Figure 8 shows directional and seasonal return values corresponding to a return period of 100 years; comparing the second panel of this figure with

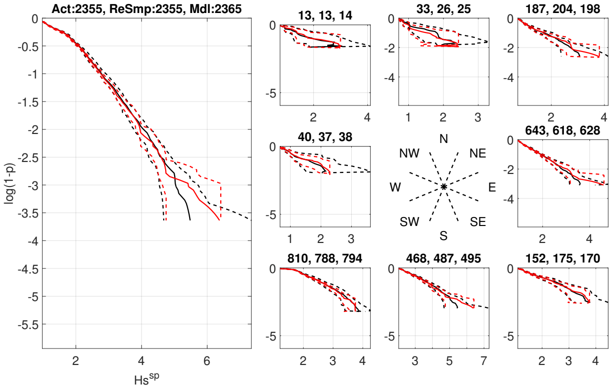


Figure 5. *1D Drc*: Validation of return value estimates for H_S^{SP}. The red curves represent the median (solid) and 95% bootstrap uncertainty band (dashed) for the empirical quantile function of the original sample. The black curves represent the corresponding quantile function estimated under the ensemble model. The left hand plot corresponds to the omni-directional, and the 8 right hand panels to directional estimates. Plot titles give the numbers of actual, sampled and simulated events.

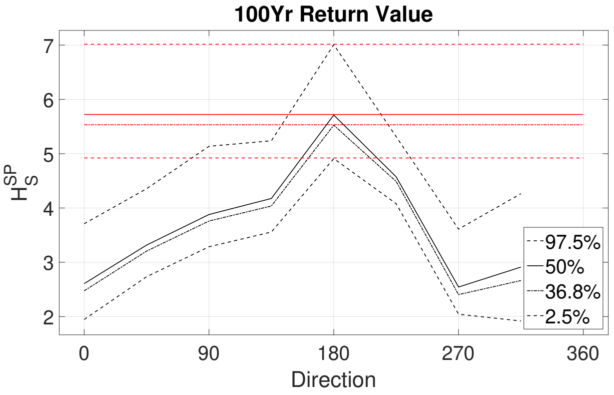
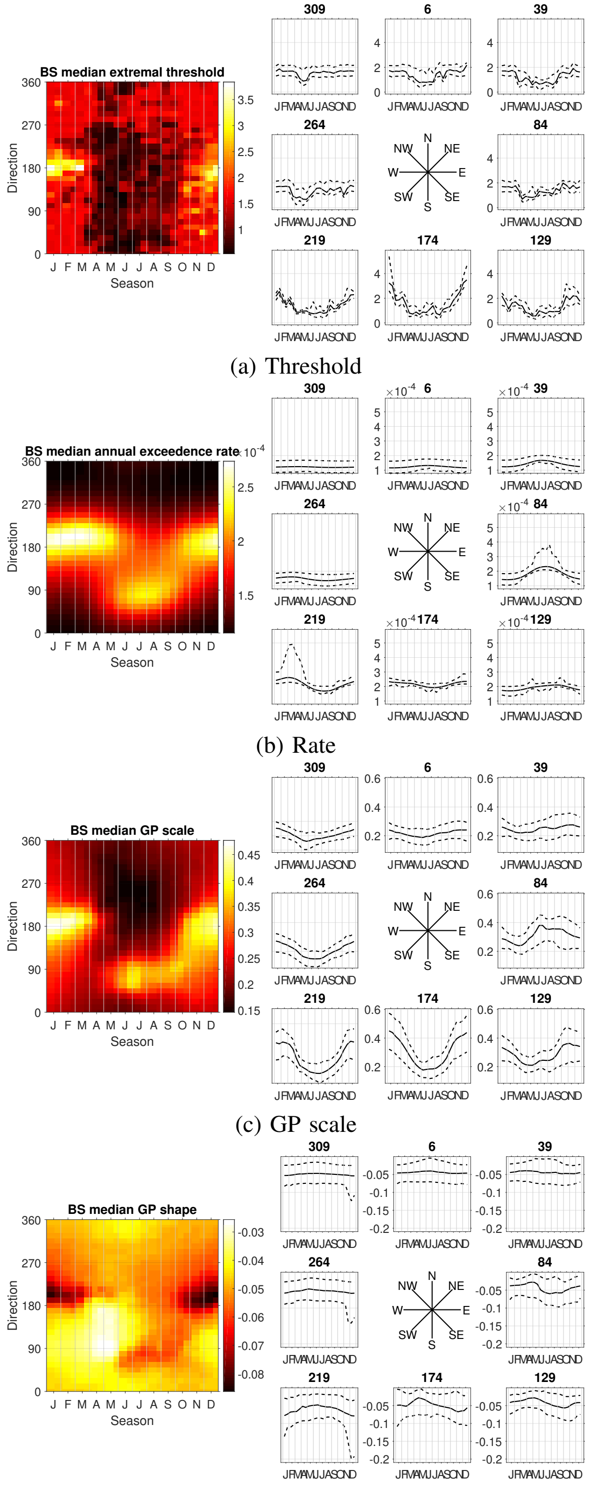


Figure 6. *1D Drc*: 100-year H_S^{SP} return value (metres) by direction. Quantiles of non-stationary (black) and stationary (red) return values illustrated, corresponding to different quantile levels.

Figure 6 suggests reasonable agreement between the models for estimation of directional return values.

C. 3D spatio-seasonal, 3D spatio-directional and 4D spatio-directional-seasonal models

We now extend the directional-seasonal model to include spatial variation, using a 11 × 11 spatial grid centred at the location for which the directional and directional-seasonal models were previously estimated. Grid spacing is approximately 11km. For illustration, we also compare estimates of return values from a 3D spatio-seasonal model (ignoring directional effects) with those from a 4D model. Figures 9 and 10 show median parameter estimates from the ensemble model by season and direction respectively. Figure 11 shows the 100-year return value as a function of direction, season and location; for comparison, Figure 12 illustrates spatio-seasonal return values from a 3D spatio-



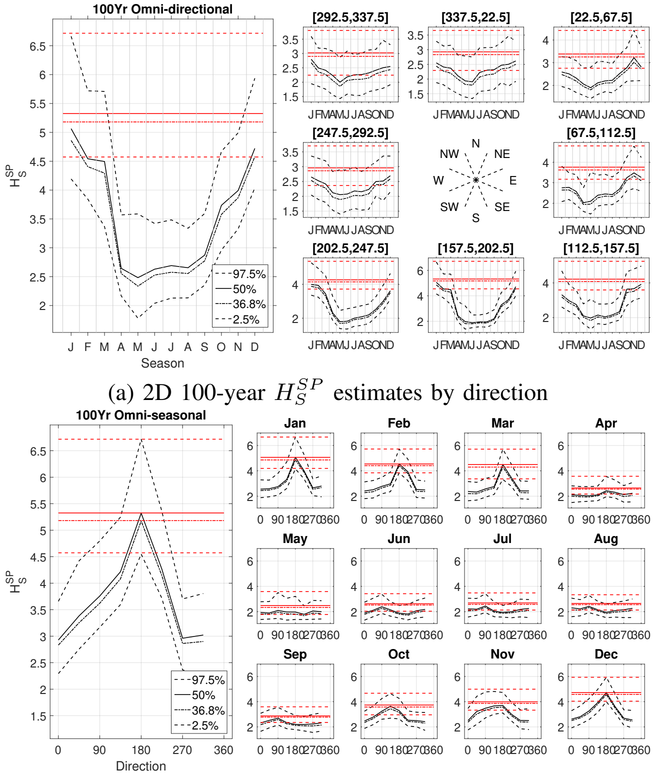
(a) Threshold

(b) Rate

(c) GP scale

(d) GP shape

Figure 7. *2D Drc-Ssn*: Directional-seasonal model. (a) Threshold ψ, (b) rate ρ, (c) GP scale σ and (d) GP shape ξ. Left panel shows median estimates and 8 right-hand panels the median (solid) and 95% uncertainty bands (dashed) for seasonal variation in 8 directional octants. Titles give average number of events per sector.



(a) 2D 100-year H_S^{SP} estimates by direction

(b) 2D 100-year H_S^{SP} estimates by season

Figure 8. *2D Drc-Ssn*: 100-year H_S^{SP} return value (metres) by (a) direction and (b) season. In each plot the left-hand panel shows the omni-covariate return values with the right hand panel showing covariate dependence. Quantiles of non-stationary (black) and stationary (red) return values illustrated, corresponding to different quantile levels.

seasonal model. There is good agreement, suggesting that characterising directional dependence is not important for estimation of (omni-directional) spatio-seasonal return values. Of course, quantifying directional dependence is essential for estimation of directional return values, and impossible with a 3D spatio-seasonal model.

Figure 13 compares estimates for directional return values from 1D directional, 2D directional-seasonal, 3D spatio-directional and 4D spatio-directional-seasonal models. To our knowledge, this is the first time that such a comparative analysis of the estimates from multidimensional extreme value models has been presented. Figure 14 compares seasonal return values in a similar way. In this case, there is good agreement between estimates for directional and seasonal return values from models incorporating at least those covariates. Asymptotically, we can argue that this might be the case when inferences are dominated by covariate combinations yielding the most extreme events; in practice however for finite samples, this may not be expected [30]. In summary, our framework enables an unified approach to deriving consistent estimates of extreme value by effectively modelling covariate effects.

IV. Discussion

Understanding rare or extreme events is important in many fields of human activity, yet predicting their characteristics is challenging exactly because extreme events occur infrequently. Moreover, there is strong evidence that the characteristics of extreme events are dependent in general on covariates; estimating this non-stationarity is vital to reliable extreme value predictions. Prediction is generally problematic, since it corresponds to estimating points far out in the tail of a probability distribution; predictions exhibit large inherent natural (*aleatory*) uncertainty and large sampling (*epistemic*) uncertainty.

If an extreme value model is to be used reliably, it must describe aleatory uncertainty adequately; in the context of metocean extremes, this means that covariate effects must be captured within the model. In this work, we explore a set of nested non-stationary marginal extreme value models with increasingly sophisticated representations of non-stationarity. We show that directional and seasonal variability is present in samples of storm severity in the South China Sea, and that this variability is captured in extreme value model parameters, and reflected in estimates for return values under the model. If an extreme value model is to be used reliably, we should also seek to reduce its epistemic uncertainty as much as possible; this can be achieved by using the largest possible sample consistent with the extreme value model, since statistical efficiency increases with effective sample size. In a metocean context, this requires modelling a heterogeneous sample of threshold exceedances exhibiting covariate dependencies; capturing these adequately is essential. We demonstrate how storm severity in the South China Sea varies with storm direction, season and location in general. Directional and seasonal effects are clear in parameter estimates and estimates for return values. However, we also find that estimates, integrated over covariates as necessary, for spatio-seasonal return values from 3D spatio-seasonal and 4D spatio-directional-seasonal models are consistent. This is also true of estimates of directional (or seasonal) return values from all of 1D, 2D, 3D and 4D models incorporating directional (or seasonal) covariate are consistent. More generally, extreme value inferences for any combination of directional, seasonal and spatial covariates can be made consistently from the 4D model.

We recommend that the potential influence of covariates always be considered in extreme value estimation. The approach outlined in this article provides a general framework to achieve this.

Acknowledgements

The authors are grateful for the advice and encouragement of Shell colleagues Vadim Anokhin, Bertwim van Beest, Shejun Fan, Graham Feld and Vianney Koelman, and academic colleagues at Lancaster University, UK.

References

[1] M. Ghil, P. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. W. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin, "Extreme Events: Dynamics, Statistics and Prediction ," *Nonlin. Processes Geophys.*, vol. 18, pp. 295–350, 2011.

[2] A. S. Sharma, D. N. Baker, A. Bhattacharyya, A. Bunde, V. P. Dimri, H. K. Gupta, V. K. Gupta, S. Lovejoy, I. G. Main, D. Schertzer, H. von Storch, and N. W. Watkins, "Complexity and Extreme Events in Geosciences: An Overview," *Extreme Events and Natural Hazards: The Complexity Perspective*, vol. 196, 2012.

[3] P. J. Mailier, D. B. Stephenson, C. A. T. Ferro, and K. I. Hodges, "Serial clustering of extratropical cyclones," *Mon. Weather Rev.*, vol. 134, pp. 2224–2240, 2006.

[4] J. Z. Y. Ogata and D. Vere-Jones, "Stochastic Declustering of Space-Time Earthquake Occurrences," *J. Am. Statist. Soc.*, vol. 97, pp. 369–380, 2002.

[5] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[6] P. Jonathan and K. C. Ewans, "Statistical Modelling of Extreme Ocean Environments with Implications for Marine Design : A Review," *Ocean Engineering*, vol. 62, pp. 91–109, 2013.

[7] M. Falk, J. Husler, and R. D. Reiss, *Laws of small numbers: extreme and rare events*. Basel, Switzerland: Birkhauser, 2004.

[8] J. Paik and D. Kim, "Advanced Method for the Development of an Empirical Model to Predict Time-Dependent Corrosion Wastage," *Corrosion Sci.*, vol. 63, pp. 51–58, 2012.

[9] Y. Bensalah, "Steps in Applying Extreme Value Theory to Finance: A Review ," *Bank of Canada*, 2000.

[10] C. Taylor, "Corrosion Informatics: An Integrated Approach to Modelling Corrosion," *Corr. Engg. Sci. Tech.*, vol. 50, pp. 490–508, 2015.

[11] P. Woodworth, "Trends in U.K. Mean Sea Level," *Marine Geodesy*, vol. 11, pp. 57–87, 1987.

[12] A. Davison and R. L. Smith, "Models for Exceedances Over High Thresholds," *J. R. Statist. Soc. B*, vol. 52, p. 393, 1990.

[13] L. Raghupathi, D. Randell, P. Jonathan, and K. Ewans, "Fast Computation of Large Scale Marginal Spatio-Directional Extremes," *Comp. Stat. Dat. Anal.*, vol. 95, pp. 243–258, 2016.

[14] L. Raghupathi, D. Randell, P. Jonathan, and K. C. Ewans, "Consistent Design Criteria for South China Sea With a Large-Scale Extreme Value Model ," in *Proc. Offshore. Tech. Conf. Asia*, 2016.

[15] I. Currie, M. Durban, and P. Eilers, "Generalized Linear Array Models with Applications to Multidimensional Smoothing," *J. R. Stat. Soc. B*, 2006.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2012.

[17] D. Randell, G. Feld, K. C. Ewans, and P. Jonathan, "Distributions of Return Values for Ocean Wave Characteristics in the South China Sea using Directional-seasonal Extreme Value Analysis," *Environmetrics*, vol. 26, no. 6, pp. 442–450, 2015.

[18] A. C. Davison, S. A. Padoan, and M. Ribatet, "Statistical Modelling of Spatial Extremes," *Statistical Science*, vol. 27, pp. 161–186, 2012.

[19] P. Jonathan and K. C. Ewans, "Modelling the seasonality of extreme waves in the Gulf of Mexico," *ASME J. Offshore Mech. Arct. Eng.*, vol. 133:021104, 2011.

[20] V. Chavez-Demoulin and A. Davison, "Generalized Additive Modelling of Sample Extremes," *J. Roy. Statist. Soc. Series C: Applied Statistics*, vol. 54, p. 207, 2005.

[21] M. Jones, D. Randell, K. Ewans, and P. Jonathan, "Statistics of extreme ocean environments: non-stationary inference for directionality and other covariate effects," *Ocean Eng.*, vol. 119, pp. 30–46, 2016.

[22] P. Jonathan and K. C. Ewans, "A spatiodirectional Model for Extreme Waves in the Gulf of Mexico," *ASME J. Offshore Mech. Arct. Eng.*, vol. 133:011601, 2011.

[23] ——, "On Modelling Seasonality of Extreme Waves," in *Proc. 27th International Conf. on Offshore Mechanics and Arctic Engineering, 4-8 June, Estoril, Portugal*, 2008.

[24] J. M. Dixon, J. A. Tawn, and J. M. Vassie, "Spatial Modelling of Extreme Sea-levels," *Environmetrics*, vol. 9, pp. 283–301, 1998.

[25] D. Randell, Y. Yu, P. Jonathan, and K. C. Ewans, "Modelling Covariate Effects in Extremes of Storm Severity on the Australian North West Shelf," in *Proc. 32nd Intl. Conf. Offshore Mech. Arc. Eng., Nantes, France*, 2013.

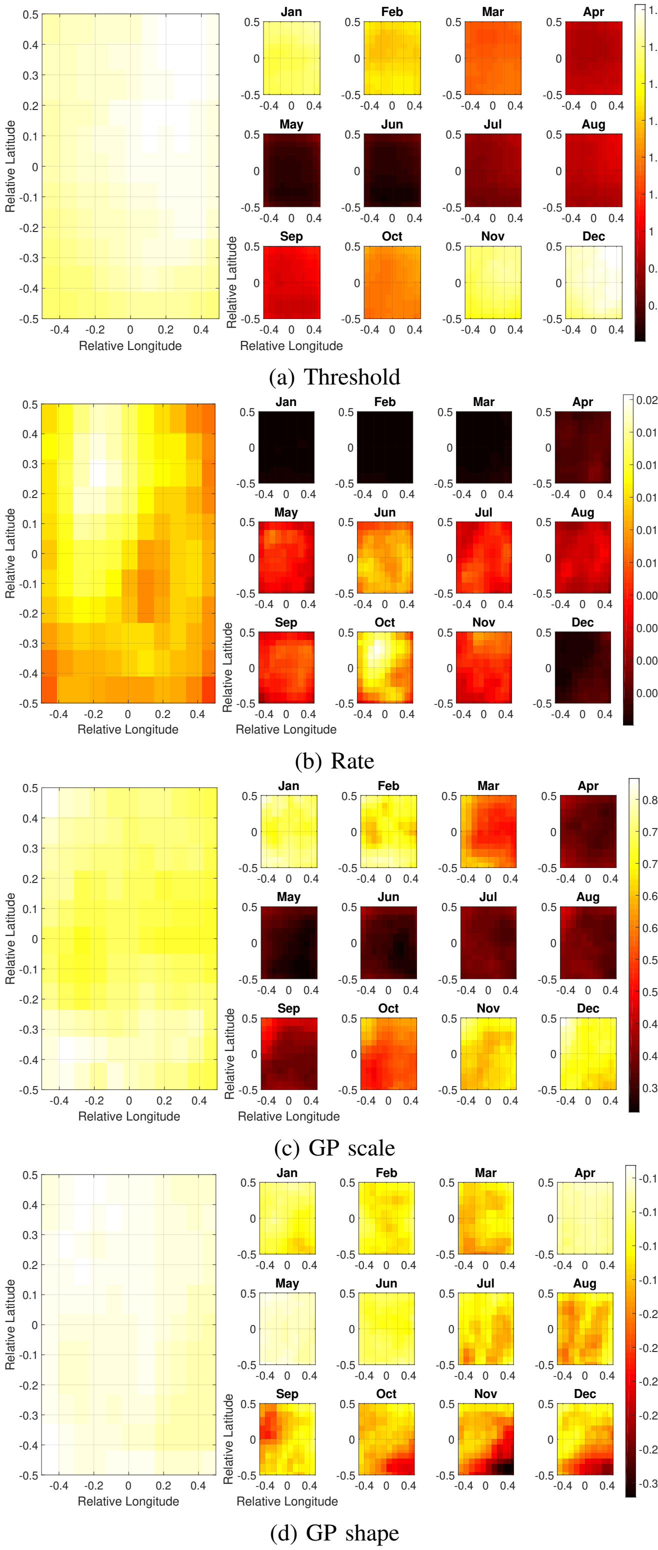
[26] P. H. C. Eilers and B. D. Marx, "Splines, Knots and Penalties," *Wiley Interscience Reviews: Computational Statistics*, vol. 2, pp. 637–653, 2010.

[27] K. C. Ewans and P. Jonathan, "The Effect of Directionality on Northern North Sea Extreme Wave Design Criteria," *J. Offshore Mech. Arc. Engg.*, vol. 130, no. 10, 2008.

[28] C. Scarrott and A. MacDonald, "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *Revstat*, vol. 10, pp. 33–60, 2012.

[29] Oceanweather, "SEAFINE: SEAMOS South FINEgrid Hindcast Study," http://www.oceanweather.com/metocean/seafine/, 2014, accessed: 2016-02-24.

[30] P. Jonathan, K. C. Ewans, and G. Z. Forristall, "Statistical estimation of extreme ocean environments: The requirement for modelling directionality and other covariate effects," *Ocean Eng.*, vol. 35, pp. 1211–1225, 2008.



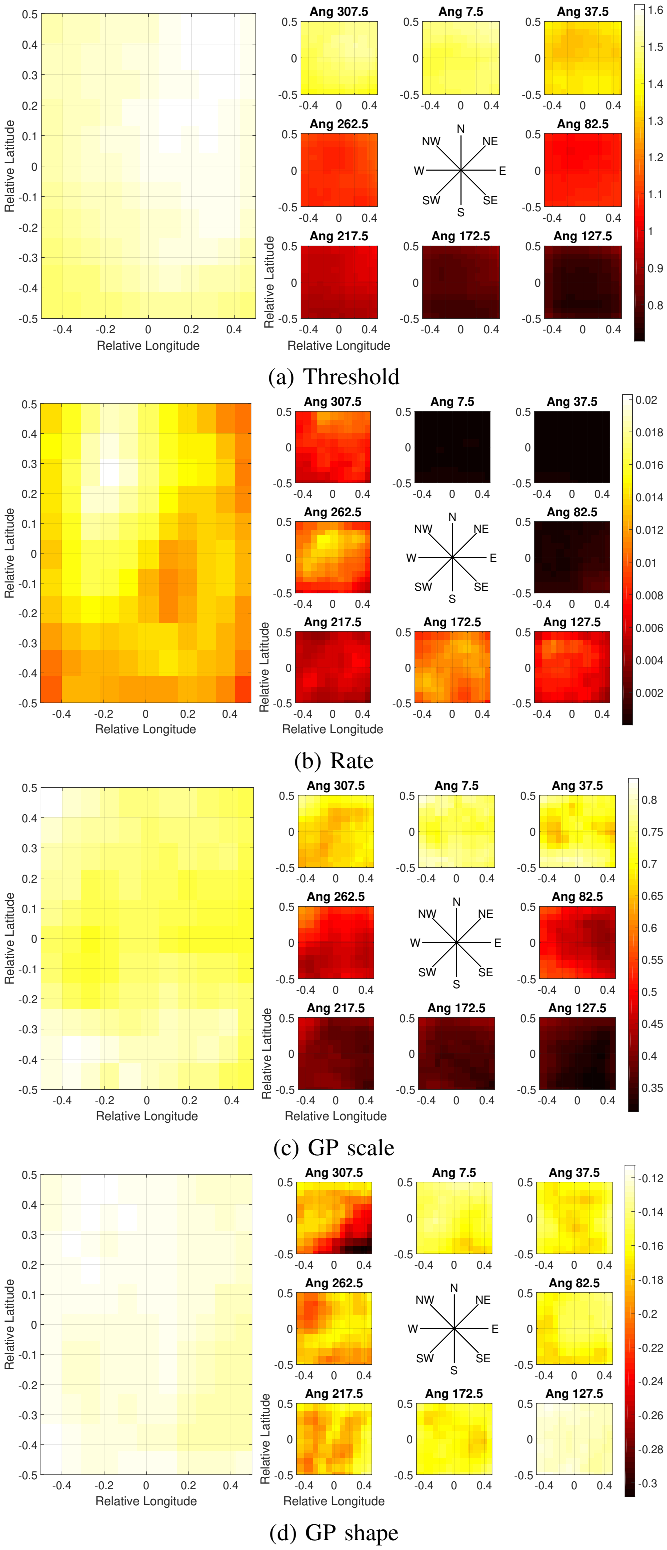
(a) Threshold

(b) Rate

(c) GP scale

(d) GP shape

Figure 9. *4D Drc-Ssn-Lon-Lat*: Extreme value model parameters for directional-seasonal-spatial covariate showing median estimates from the ensemble model by *season*. Left-hand panels indicate omni-seasonal estimates and 12 right-hand panels monthly estimates. (a) Threshold ψ, (b) rate ρ, (c) GP scale σ and (d) GP shape ξ.



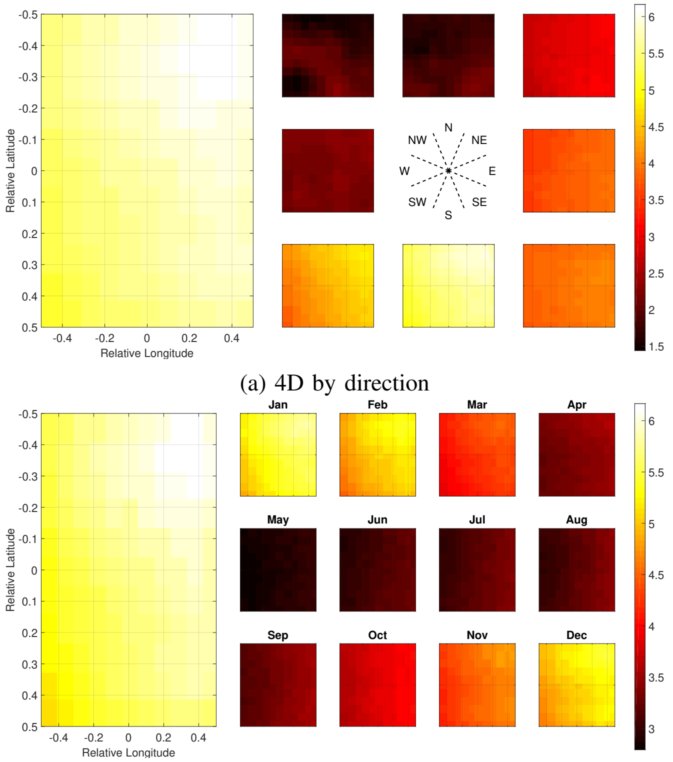
(a) Threshold

(b) Rate

(c) GP scale

(d) GP shape

Figure 10. *4D Drc-Ssn-Lon-Lat*: Extreme value model parameters for directional-seasonal-spatial covariate showing median estimates from the ensemble model by *direction*. Left-hand panels indicate omni-directional estimates and 8 right-hand panels directional octant estimates. (a) Threshold ψ, (b) rate ρ, (c) GP scale σ and (d) GP shape ξ.



(a) 4D by direction

(b) 4D by season

Figure 11. *4D Drc-Ssn-Lon-Lat*: 100-year H_S^{SP} return value (metres) for spatio-directional-seasonal covariate by (a) direction and (b) season. In each plot, the left-hand panel indicates the omni-covariate estimate from the threshold ensemble model with the right-hand panel indicating the directional and seasonal estimates.

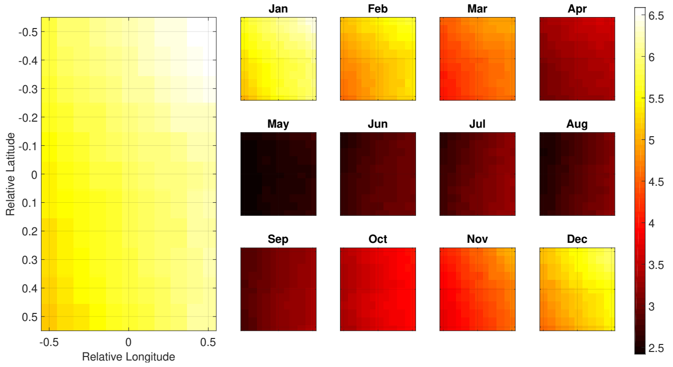


Figure 12. *3D Ssn-Lon-Lat*: 100-year H_S^{SP} return value (metres) for spatio-seasonal covariate. The left-hand panel indicates the omni-seasonal estimate from the threshold ensemble model with the 12 right-hand panel indicating the 12 monthly estimates.

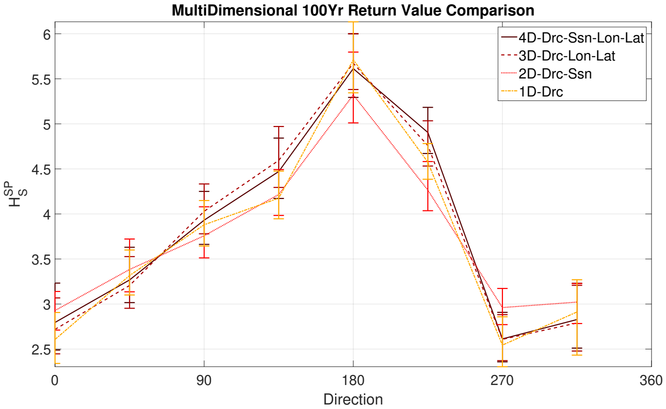


Figure 13. Comparison of 100-year directional return values for H_S^{SP} from 1D, 2D, 3D and 4D models incorporating directional covariate.

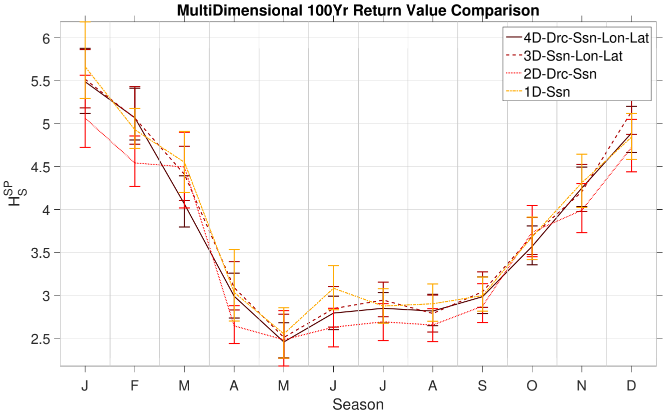


Figure 14. Comparison of 100-year seasonal return values for H_S^{SP} from 1D, 2D, 3D and 4D models incorporating seasonal covariate.