Deep learning joint extremes of metocean variables using the SPAR model

E. Mackay*[1], C.J.R. Murphy-Barltrop[2,3], J. Richards[4], and P. Jonathan[5]

[1]University of Exeter, UK
[2]Technische Universität Dresden, Germany
[3]ScaDS.AI, Germany
[4]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, UK
[5]Lancaster University, UK

June 23, 2025

Abstract

This paper presents a novel deep learning framework for estimating multivariate joint extremes of metocean variables, based on the Semi-Parametric Angular-Radial (SPAR) model. When considered in polar coordinates, the problem of modelling multivariate extremes is transformed to one of modelling an angular density, and the tail of a univariate radial variable conditioned on angle. In the SPAR approach, the tail of the radial variable is modelled using a generalised Pareto (GP) distribution, providing a natural extension of univariate extreme value theory to the multivariate setting. In this work, we show how the method can be applied in higher dimensions, using a case study for five metocean variables: wind speed, wind direction, wave height, wave period, and wave direction. The angular variable is modelled using a kernel density method, while the parameters of the GP model are approximated using fully-connected deep neural networks. Our approach provides great flexibility in the dependence structures that can be represented, together with computationally efficient routines for training the model. Furthermore, the application of the method requires fewer assumptions about the underlying distribution(s) compared to existing approaches, and an asymptotically justified means for extrapolating outside the range of observations. Using various diagnostic plots, we show that the fitted models provide a good description of the joint extremes of the metocean variables considered.

1 Introduction

Many problems in offshore and coastal engineering require estimation of joint extremes for metocean variables. Responses of offshore and coastal structures are dependent on multiple variables, such as wind speed and direction, wave height, period and direction, flow speed and direction. Providing reliable estimates of the joint extremes in this setting is a challenging problem for metocean engineers. Various design standards recommend the use of the environmental contour method [1]. Some types of contour can be estimated without an explicit model for the joint distribution of variables [2, 3]. However, environmental contour methods typically make simplifying assumptions and only give approximate estimates of long-term extreme responses [4, 5]. Full probabilistic analysis of long-term extreme responses requires a model for the joint density of the relevant metocean variables. A wide range of approaches have been proposed for estimating joint densities. In the offshore engineering literature, the two most popular approaches are global hierarchical models and copula models – see, e.g., [6, 7].

Let $\mathbf{X} = (X_1, ..., X_d) \in \mathbb{R}^d$, $d \geq 2$ denote a continuous random vector with joint density function $f_{\mathbf{X}}$, and marginal density and cumulative distribution functions f_{X_j} and F_{X_j}, respectively, for $j = 1, ..., d$. In the global hierarchical approach [8–11], the joint density is written as

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1)\, f_{X_2|X_1}(x_2|x_1) \cdots f_{X_d|(X_1,...,X_{d-1})}(x_d|x_1, ..., x_{d-1}), \tag{1}$$

where $f_{X_j|(X_1,...,X_{j-1})}$ is the density of X_j conditional on $(X_1, ..., X_{j-1})$ for $j \in \{2, \ldots, d\}$. Inference typically involves selecting parametric forms for f_{X_1}, $f_{X_2|X_1}$, ..., $f_{X_d|(X_1,...,X_{d-1})}$ and estimating relations between the parameters of

*email: e.mackay@exeter.ac.uk, ORCID: 0000-0001-7121-4231

the conditional densities and the conditioning variables. There are various problems with this approach. Firstly, there is no a priori reason to suppose that variables follow any particular parametric distribution, and misspecified models can have dramatic consequences when approximating dependence structures. Secondly, a model fit to all of the observations does not guarantee a good fit to the tail, which is the region of interest for extremes. Finally, the models for the parameters of the conditional densities are usually based on ad hoc assumptions, and provide no rationale for extrapolating outside the range of observations. In many cases, it has been shown that such models provide a poor fit to observed data, especially in extreme regions [12].

For copula modelling, the joint density is written as

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) \cdots f_{X_d}(x_d) c(F_{X_1}(x_1), \cdots, F_{X_d}(x_d)), \tag{2}$$

where $c : [0,1]^d \mapsto [0, \infty)$ is the copula density of $\mathbf{X}$ [13]. In this case, inference involves choosing parametric models for both the marginal densities and for c. As with the global hierarchical approach, there are no a priori reasons to choose particular models. Similarly, fitting to all observations does not guarantee a good fit to the tails. Moreover, different copula models have very different behaviours in the joint tail regions, meaning extrapolation can vary substantially for different choices of copula model [13].

There are also a wide range of methods in the statistical literature for modelling joint extremes (e.g., [14–16]). However, many of these approaches make strong assumptions about the dependence structure, or copula, which are often not supported by environmental datasets [17]. The most popular choice for metocean variables is the conditional extremes model [18], which describes the joint distribution of variables conditional on at least one variable being large. The key limitation of this approach is that it only characterises the region of variable space where the conditioning variable is large, and inferences made using different conditioning variables are not necessarily consistent [19]. A further limitation of this method (and other methods in the multivariate extremes literature) is that it requires a transformation of the margins to a standard scale. This requires first estimating the marginal distributions for each variable – a process which is subject to uncertainty. Furthermore, it has been demonstrated that poor marginal estimates greatly affect the quality of the resulting multivariate inference [20].

In this paper, we discuss the application of a new method, introduced in [21], which overcomes the limitations of existing approaches and provides a general, flexible framework for modelling multivariate extremes. The model is referred to as the Semi-Parametric Angular-Radial (SPAR) model. The SPAR model provides a framework for estimating multivariate extremes that does not require strong assumptions about the form of the margins or dependence structure, and provides a justified means of extrapolating outside the range of observations. Moreover, the model is only fitted to extreme observations, meaning that no assumptions are required about the bulk of the distribution. Theoretical aspects of the SPAR model are presented in [22], and an inference approach in a two-dimensional setting is provided in [23, 24]. The purpose of this paper is to extend the modelling method to the general multivariate setting, with two or more dimensions. The two-dimensional inference scheme in [23, 24] utilised cubic regression splines to model the angular dependence of the radial distribution parameters. In this work we take a different approach, and adopt a deep learning scheme, where the radial distribution parameters are modelled using artificial neural networks. As discussed further in Section 3, this offers some computational advantages in higher-dimensional settings.

The paper is organised as follows. Section 2 describes a brief overview of the theoretical aspects of the model. Our deep learning approach for estimating the radial component of the SPAR model is introduced in Section 3. Section 4 presents an example application of the model to a five-dimensional problem: estimating the joint extremes of wind speed, wave direction, wave height, wave period, and wave direction. We discuss the challenges that arise for these particular variables, and how well the model assumptions are satisfied in this setting. Several novel types of diagnostic plots are introduced to assess the fit of the model. We conclude in Section 5 with a discussion and outlook on future work.

2 Theory

2.1 SPAR model definition

The SPAR model can be viewed as an extension of the univariate peaks-over-threshold (POT) method to the multivariate setting. It involves a transformation of variables to angular-radial coordinates, and then models the upper-tail of the radial variable, conditional on angle, using a non-stationary generalised Pareto (GP) model. Suppose that we have a continuous random vector $\mathbf{X} = (X_1, ..., X_d) \in \mathbb{R}^d$ with joint density function $f_{\mathbf{X}}$. We define radial and angular variables as

$$R = \|\mathbf{X}\|_2, \quad \mathbf{W} = \mathbf{X}/R, \tag{3}$$

where $\|\cdot\|_2$ is the L^2 or Euclidean norm, defined by $\|(x_1, ..., x_d)\|_2 = (x_1^2 + \cdots + x_d^2)^{1/2}$. Note that $R \in [0, \infty)$ and $\mathbf{W} \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ is the unit hypersphere in $\mathbb{R}^d$. The joint density function of $(R, \mathbf{W})$ is related to $f_{\mathbf{X}}$ via

$$f_{R,\mathbf{W}}(r, \mathbf{w}) = r^{d-1} f_{\mathbf{X}}(r\mathbf{w}), \quad r \geq 0, \, \mathbf{w} \in \mathbb{S}^{d-1}, \tag{4}$$

where r^{d-1} is the Jacobian determinant for the transformation $\mathbf{X} \to (R, \mathbf{W})$. As for global hierarchical models, the angular-radial joint density can be written in conditional form as:

$$f_{R,\mathbf{W}}(r, \mathbf{w}) = f_{\mathbf{W}}(\mathbf{w}) \, f_{R|\mathbf{W}}(r|\mathbf{w}). \tag{5}$$

Noting that $\mathbf{X} = R\mathbf{W}$, and that $\mathbf{W}$ lies on the surface of the unit hypersphere, we can see that the 'extreme' parts of the distribution of $\mathbf{X}$ correspond to large values of the radial variable at any given angle. Therefore, the problem of modelling multivariate extremes is transformed to that of modelling an angular density $f_{\mathbf{W}}$ and the tail of the conditional radial density $f_{R|\mathbf{W}}$. For a given angle $\mathbf{w}$, the density $f_{R|\mathbf{W}}(r|\mathbf{w})$ is univariate. Univariate extreme value theory suggests that a suitable model for the tail of $f_{R|\mathbf{W}}$ is the GP distribution, with parameters conditional on angle (e.g., [25]). This motivates the SPAR model, whereby parametric and non-parametric models are used to model the conditional radial and angular distributions, respectively. Define a threshold function $u(\mathbf{w}) > 0$ to be the quantile of $R|(\mathbf{W} = \mathbf{w})$ with exceedance probability $\zeta \in (0, 1)$, with ζ close to 0, i.e., the solution of $\zeta = \Pr(R > u(\mathbf{w})|\mathbf{W} = \mathbf{w})$. The SPAR model can be written as

$$f_{R,\mathbf{W}}(r, \mathbf{w}) = \zeta f_{\mathbf{W}}(\mathbf{w}) f_{\mathrm{GP}}(r - u(\mathbf{w}); \xi(\mathbf{w}), \sigma(\mathbf{w})), \quad r > u(\mathbf{w}), \, \mathbf{w} \in \mathbb{S}^{d-1}, \tag{6}$$

where f_{GP} is the GP density function, and $\xi(\mathbf{w}) \in \mathbb{R}$ and $\sigma(\mathbf{w}) > 0$ are shape and scale parameters, respectively, given as functions of the angle $\mathbf{w}$. The GP density function is given by

$$f_{\mathrm{GP}}(r; \xi, \sigma) = \begin{cases} \dfrac{1}{\sigma}\left(1 + \xi\dfrac{r}{\sigma}\right)^{-1-\frac{1}{\xi}}, & \xi \neq 0, \\ \dfrac{1}{\sigma}\exp\left(-\dfrac{r}{\sigma}\right), & \xi = 0, \end{cases} \tag{7}$$

which is supported on $0 \leq r \leq r^F$, where $r^F = \infty$ for $\xi \geq 0$ and $r^F = -\sigma/\xi$ for $\xi < 0$.

Many non-parametric methods for estimation of densities, such as kernel density methods, mixture models or spline-based methods, assume that the density is finite and continuous. Similarly, many representations for non-stationary modelling of parametric distributions assume that the parameter functions are finite and continuous. Therefore, to simplify our inference, we assume that the angular density $f_{\mathbf{W}}$, threshold function $u(\mathbf{w})$, and GP parameter functions, $\xi(\mathbf{w})$ and $\sigma(\mathbf{w})$, are finite and continuous with respect to the angle $\mathbf{w}$.

After estimation of the angular density and GP parameter functions, (4) and (6) can be combined to obtain the SPAR estimate of the joint density in the original variable space for observations satisfying $r > u(\mathbf{w})$, i.e.,

$$f_{\mathbf{X}}(r\mathbf{w}) = \zeta \, r^{1-d} \, f_{\mathbf{W}}(\mathbf{w}) \, f_{\mathrm{GP}}(r - u(\mathbf{w}); \xi(\mathbf{w}), \sigma(\mathbf{w})). \tag{8}$$

Calculating marginal and joint probabilities using the SPAR model then involves either integration of the joint density over specified angular and radial domains, or via Monte Carlo techniques, i.e., by simulating from the estimated model and deriving probability estimates empirically. To simulate from the SPAR model, we first draw an angle $\mathbf{w}$ from $f_{\mathbf{W}}$, then use inversion sampling to generate a corresponding value z from the GP distribution with parameter vector $(\xi(\mathbf{w}), \sigma(\mathbf{w}))$, and finally define a radial value $r = u(\mathbf{w}) + z$. The pair $(r, \mathbf{w})$ is then a random sample from the SPAR model. This can be converted back to the original variable space using the inverse transformation $\mathbf{x} = r\mathbf{w}$. As the SPAR model is only fitted to observations for which $r > u(\mathbf{w})$, one can create a sample (of the original random vector $\mathbf{X}$) of size N by simulating ζN points from the SPAR model, and then resampling $(1 - \zeta)N$ points from observations with $r \leq u(\mathbf{w})$. The rationale for this is that there should be a sufficient number of observations within the body of the distribution to obtain a reasonable estimate from resampling.

2.2 Exceedance probability contours

As described in [24], the SPAR model provides an explicit means for calculating a contour with a specified exceedance probability $\beta \in (0, \zeta]$. The radius of the contour at angle $\mathbf{w}$ is simply the quantile of the GP distribution at exceedance probability β/ζ, given by

$$r_\beta(\mathbf{w}) = u(\mathbf{w}) + \frac{\sigma(\mathbf{w})}{\xi(\mathbf{w})}\left(\left(\frac{\beta}{\zeta}\right)^{-\xi(\mathbf{w})} - 1\right). \tag{9}$$

This contour is defined in terms of the probability of an observation falling anywhere outside the contour region, or the 'total exceedance probability'. As such, these contours are more conservative than those defined in terms of marginal exceedance probabilities, such as IFORM contours (or variants thereof), with the conservatism increasing with the number of dimensions [26]. Moreover, if the primary interest of the analysis is to estimate environmental contours, then the use of the SPAR model is not necessary. Instead, we recommend the use of the Direct-IFORM method [2, 3], which does not require a model for the joint density or any assumptions about the dependence structure between the variables.

The contours defined in (9) can be projected into two dimensions. In Cartesian space, points on the contour are given by $\mathbf{x} = r_\beta(\mathbf{w})\,\mathbf{w}$ for $\mathbf{w} \in \mathbb{S}^{d-1}$. The radii $r_\beta \geq 0$ can be computed for a discrete set of points on the sphere (see Appendix A), and the projection into dimensions $i, j \in \{1, ..., d\}$ is just the i and j components of $\mathbf{x}$.

However, in our application it was found that these contours do not provide a useful diagnostic tool. This is because they do not account for the angular density. In regions of low angular density, estimates of the contours can be highly uncertain. Simpson and Tawn [27] proposed to mitigate for this by either setting the radius to zero when the angular density is zero, or adjusting the exceedance level to account for the angular density. However, the latter approach requires an integral over angles that becomes impractical in higher dimensions. We have therefore opted to use alternative diagnostics, described in Section 4.5.

3 Inference

Inference for the SPAR model involves estimating the angular density, and the GP threshold and parameter functions. Since the angular-radial density factorises (see (5)), these problems are separable: inference for $f_{\mathbf{W}}$ can be conducted independently of that for $(u(\mathbf{w}), \sigma(\mathbf{w}), \xi(\mathbf{w}))$. Inference for the angular density is discussed in Section 3.1, and modelling of the conditional radial variable is discussed in Section 3.2. Code for fitting our model is available upon reasonable request.

3.1 Angular modelling

Estimation of densities on the hypersphere $\mathbb{S}^{d-1}$ is part of a discipline known as directional statistics [28, 29]. The key difference from estimation of densities on $\mathbb{R}^d$ is that the surface of the hypersphere is periodic and bounded, and so distributions defined on $\mathbb{S}^{d-1}$ must conserve these constraints. Various parametric and non-parametric approaches have been developed for estimating densities on the hypersphere, which are directly analogous to approaches used in Euclidean space. These include kernel density (KD) estimation [30, 31], mixture models [32–34], and spline-based methods [35]. See [36] for a recent review of non-parametric approaches, and [37] for a comparison of various generative deep learning methods for modelling angular densities.

In keeping with the previous uses of SPAR in two dimensions, in this work we use a KD method for the angular modelling. The advantages of KD estimation are that it is simple to implement and fast to sample from. The main downside is that explicit calculation of the density can be slow for large sample sizes. However, for our purpose, simulation from the model is more important than explicit calculation of the density, so this drawback is less significant.

3.1.1 Kernel density estimation for spherical data

Suppose that $K(\mathbf{w}; \boldsymbol{\mu}, \kappa)$ is a density function on $\mathbb{S}^{d-1}$, with mean direction $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$, bandwidth parameter $\kappa \geq 0$ that controls the concentration of the density around $\boldsymbol{\mu}$. Then the KD estimate of the angular density $f_{\mathbf{W}}$ from a sample of n observations $\{\mathbf{w}_i\}_{i=1,...,n} \subset \mathbb{S}^{d-1}$ is given by

$$\hat{f}_{\mathbf{W}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{w}; \mathbf{w}_i, \kappa), \tag{10}$$

That is, the estimate is effectively a mixture of n kernels, centred at each data point. From a modelling perspective, there are two key choices to make: the type of kernel to use, and the determination of the bandwidth.

3.1.2 Choice of kernel

The choice of kernel function has a relatively minor impact on the accuracy of the density estimate compared to the choice of bandwidth. While different kernels can slightly alter the shape of the density estimate, the bandwidth parameter, which controls the amount of smoothing, is the most crucial factor affecting the accuracy [38]. Here we are mainly guided by computational considerations. The most common choice of kernel for KD models on the sphere is

the von Mises-Fisher (vMF) distribution [39]. However, simulation from this distribution requires rejection sampling, which can be slow. Therefore, we opt to use the *power spherical* (PS) distribution [40]. The PS and vMF distributions are both rotationally symmetric distribution about the mean direction, and have similar shapes. However, simulation from the PS distribution is much faster, as discussed further in Appendix B. For $\mathbf{w} \in \mathbb{S}^{d-1}$ the density of the PS distribution is given by

$$K(\mathbf{w}; \boldsymbol{\mu}, \kappa) = (4\pi)^{-\eta} \frac{\Gamma(2\eta + \kappa)}{\Gamma(\eta + \kappa)} \left[\tfrac{1}{2}(1 + \mathbf{w}^\top \boldsymbol{\mu})\right]^\kappa, \tag{11}$$

where $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is the mean direction, $\kappa \geq 0$ is the bandwidth parameter, $\eta = (d-1)/2$, and $\Gamma(\cdot)$ is the gamma function. The dot product of two unit vectors is the cosine of the angle between them. Therefore, the term $\mathbf{w}^\top \boldsymbol{\mu} \in [-1, 1]$, defines the cosine of the arc length between $\mathbf{w}$ and $\boldsymbol{\mu}$, and plays an analogous role to the Euclidean distance $\|\mathbf{x} - \boldsymbol{\mu}\|_2 \in [0, \infty)$ for isotropic kernels on $\mathbb{R}^d$.

For KD models, the bandwidth parameter κ determines the balance between over-smoothing and over-fitting; it's optimisation is discussed below. The PS distribution with $\kappa = 0$ is the uniform distribution on $\mathbb{S}^{d-1}$. As κ increases, the distribution becomes more concentrated about the mean direction. For larger sample sizes, using higher bandwidth parameters tends to result in a better resolution of complex distributional features. However, this can lead to numerical overflow problems when evaluating the gamma functions in (11), since $\Gamma(z) \sim \sqrt{2\pi z}(z/e)^z$ as $z \to \infty$. A more numerically stable expression is

$$K(\mathbf{w}; \mu, \kappa) = (4\pi)^{-\eta} \exp\left[\kappa \log(z) + \log\left(\Gamma(2\eta + \kappa)\right) - \log\left(\Gamma(\eta + \kappa)\right)\right],$$

where $z = \frac{1}{2}(1 + \mathbf{w}^\top \boldsymbol{\mu}) \in [0, 1]$ and hence $\log(z) \in (-\infty, 0]$. We also have $\log(\Gamma(2\eta + \kappa)) - \log(\Gamma(\eta + \kappa)) \sim \eta \log(\kappa)$ as $\kappa \to \infty$. The log gamma function $\log(\Gamma(\cdot))$ can be computed directly in most software without having to evaluate the gamma function, thus avoiding overflow issues.

3.1.3 Bandwidth optimisation

There are various ways in which the bandwidth parameter for the KD model can be optimised, as discussed in [29]. Here we use a cross-validation scheme to minimise the negative log-likelihood (NLL) for a hold-out sample. This is the most computationally expensive part of the method – for each test point, the kernel must be evaluated for each observation in the training sample. For a leave-one-out cross-validation scheme, this entails $(n-1)^2$ evaluations of the kernel:

$$\text{NLL}(\kappa) = -\sum_{i=1}^{n} \log\left(\hat{f}_{\neq i}(\mathbf{w}_i)\right), \tag{12}$$

where $\hat{f}_{\neq i}(\mathbf{w}_i)$ is the KD estimate (10) of the density from all observations except $\mathbf{w}_i$, evaluated at $\mathbf{w}_i$ (i.e., the predictive likelihood). However, as the NLL is effectively an expectation over $\mathbf{W}$, we can take a stochastic approach and estimate the NLL from a smaller sample of size m, chosen at random from the observations, so that we only need $(n-1)m$ kernel evaluations. This is similar to the approach used in stochastic gradient descent (see below) in machine learning, where a small batch of observations is used to approximate the true gradient of the cost function.

When observations are serially correlated, the estimate of $\hat{f}_{\neq i}(\mathbf{w}_i)$ will be derived, in part, from observations that are correlated with $\mathbf{w}_i$, namely $\{..., \mathbf{w}_{i-2}, \mathbf{w}_{i-1}, \mathbf{w}_{i+1}, \mathbf{w}_{i+2}, ...\}$. To obtain an unbiased estimate of the predictive likelihood using a cross-validation scheme, for each point $\mathbf{w}_i$ at which we want to calculate the predictive likelihood, we need to leave out observations around $\mathbf{w}_i$ which are correlated with it, say $\mathbf{w}_{i-k}, ..., \mathbf{w}_{i+k}$ for some fixed k.

In the example presented in Section 4, we found that using $m = 1000$ prediction points for estimating the NLL gave smooth and repeatable results for our sample size of $n = 271,704$ hourly observations, requiring $\approx 271 \times 10^6$ evaluations of the kernel (a speed up of ≈ 271 times compared to the full cross-validation scheme). We also set $k = 48$, corresponding to a two-day exclusion around each point used to calculate the predictive NLL. Computations of the NLL for 50 logarithmically-spaced test values of $\kappa \in [10^1, 10^4]$ took approximately 5 minutes on a laptop with an Intel Core i7-1355U 1.7GHz processor. The results are shown in Figure 1, which indicates an optimal value of $\kappa \approx 1200$.

3.1.4 Simulation from a kernel density estimate

Once the value of the bandwidth parameter has been optimised, it is straightforward to simulate from the KD model. As mentioned above, the KD model is effectively a mixture model, with one component corresponding to each observation, centred at the location of the observation. Therefore, to generate a random sample from the KD model requires two steps. First, we sample an observed angle $\mathbf{w}_i$ at random (i.e., a random integer $i \in \{1, ..., n\}$). This corresponds to choosing one component from the mixture model at random. Secondly, we generate a random

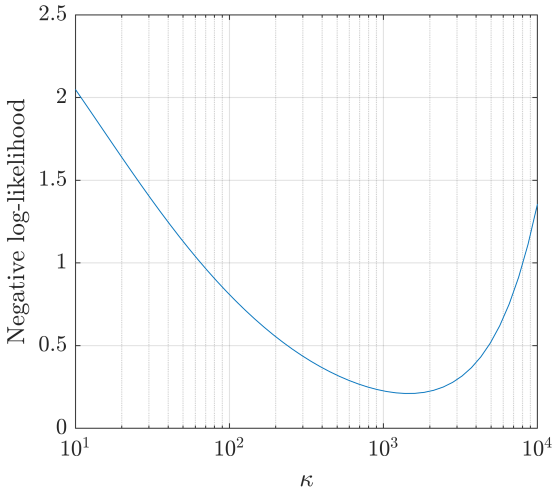

Figure 1: Optimisation of bandwidth parameter κ for kernel density estimate of the angular distribution, in terms of the minimum predictive negative log-likelihood.

value from the kernel with mean direction $\mathbf{w}_i$ and bandwidth κ. As described in Appendix B, simulation from the PS kernel on $\mathbb{S}^{d-1}$ requires generating a beta random variable and $d-1$ standard normal random variables, and applying a rotation. Computationally this is very efficient. For the example presented in Section 4, generating a sample of size 2.7×10^7 (100 times the observed sample size) took approximately 10 s on a laptop with specifications mentioned above.

3.2 Conditional radial modelling

Viewing the angular variable $\mathbf{W}$ as a 'covariate' for the radial variable R, inference for the SPAR model is analogous to a non-stationary univariate POT analysis, for which many parametric and semi-parametric approaches have been proposed [41–45]. Non-stationary POT modelling can be performed using generalised Pareto (GP) regression (i.e., modelling GP parameters as functions of covariates), which proceeds by first estimating the threshold function, $u(\mathbf{w})$, via a quantile regression procedure, and then estimating the GP scale and shape parameter functions, $\sigma(\mathbf{w})$ and $\xi(\mathbf{w})$ respectively, via likelihood-based inference procedures. The choices of the functional forms for $u(\mathbf{w})$, $\sigma(\mathbf{w})$, and $\xi(\mathbf{w})$ determine the flexibility of the overall model. As the dimension d grows, these mappings become increasingly complex, and models that represent the functions via semi-parametric models (such as the splines used in initial work with the SPAR model [23] and other similar angular-radial approaches [46, 47]) become increasingly computationally-demanding to estimate. Consequently, we adopt a deep learning approach, whereby the threshold and GP parameter functions are represented using artificial neural networks (ANN). For details on GP regression with deep learning methods, see [48].

Although the representation of the GP parameter functions via ANNs differs from previous approaches in the SPAR framework, the 'loss function' used for optimisation of the model is the same. That is, we estimate the threshold and GP parameter functions that minimise the relevant loss functions evaluated on a hold-out dataset (in the case of the GP parameters, the loss function is the negative log-likelihood), as described in Section 3.2.2. This work builds upon the approach of [49], who use deep learning to estimate the extremal dependence structure of random vectors on standardised marginal scales via a similar angular-radial decomposition. In contrast, our approach does not require marginal transformation, and is thus not subject to marginal estimation uncertainty.

3.2.1 Neural network representation of conditional radial parameters

Several recent approaches have used deep learning for GP regression; see, e.g., [50–53]. In these approaches, ANNs are used to model the relationships between covariates and GP parameter and threshold functions. Our approach is analogous, with the covariates taken to be angles on the hypersphere. Note that the hypersphere is compact (i.e., closed and bounded), and that this is a desirable property for extrapolation in deep learning.

Here, we design two neural network models; one for the threshold function $u(\mathbf{w})$ and one for the parameter vector $(\nu(\mathbf{w}), \xi(\mathbf{w}))$, where $\nu(\mathbf{w}) = \sigma(\mathbf{w})(\xi(\mathbf{w})+1)$ is the modified scale parameter. Unlike $\sigma(\mathbf{w})$, the modified scale parameter

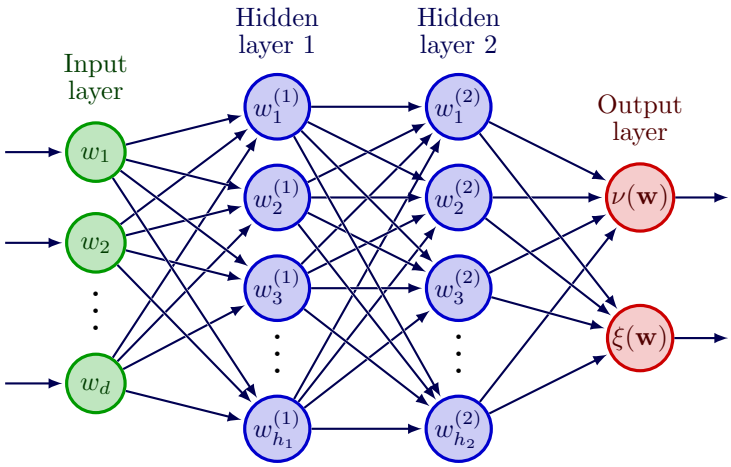

Figure 2: Example schematic of a multi-layered perceptron (MLP) model for the GP parameters, with $L = 2$ hidden layers. The inputs are the components of the angle $\mathbf{w} = (w_1, w_2, ..., w_d)$ and the outputs are the GP parameter functions, $(\nu(\mathbf{w}), \xi(\mathbf{w}))$.

$\nu(\mathbf{w})$ is orthogonal to $\xi(\mathbf{w})$, which helps to mitigate numerical instabilities during model fitting [41, 54, 55]; [53] show that this is particularly helpful when fitting deep GP regression models. Both $u(\mathbf{w})$ and $(\nu(\mathbf{w}), \xi(\mathbf{w}))$ are modelled by multi-layer perceptrons (MLPs), which are a standard class of fully-connected ANN that compose multiple layers of 'neurons' [56]. Each neuron passes a linear combination of input variables through a nonlinear 'activation function', and the output is then passed to the subsequent layer; detailed discussions and illustrative figures can be found in [48], and an example schematic for an MLP representation of $(\nu(\mathbf{w}), \xi(\mathbf{w}))$ is presented in Figure 2. Inference for the parameter functions involves estimating the linear coefficients (the 'weights' and 'biases') in each neuron of the corresponding MLP.

Prior to inference, the architecture of the MLP must be defined, i.e., the number of hidden layers, denoted by L, the number of neurons in each hidden layer, denoted by $h_1, \ldots, h_L$, and the type(s) of activation function(s). The resulting set of estimable parameters for the MLP contains all of the weights and biases in each hidden layer, as well as the final $(L+1)$-th layer; we denote this by $\mathcal{W} := \left\{ \left(a^l, b^l\right); l = 1, \ldots, L+1 \right\}$, with weights $a^l \in \mathbb{R}^{h_l \times h_{l-1}}$ and biases $b^l \in \mathbb{R}^{h_l}$. Note that the estimable sets of parameters differ between the MLPs for $u(\mathbf{w})$ and $(\nu(\mathbf{w}), \xi(\mathbf{w}))$; we denote their respective parameter sets by $\mathcal{W}_u$ and $\mathcal{W}_{(\nu,\xi)}$. For both MLPs, we take all hidden layer activation functions to be the rectified linear unit function, $\text{ReLU}(\mathbf{x}) = (\max\{x_1, 0\}, \max\{x_2, 0\}, \ldots)$. The final layers of the MLPs make use of an exponential transformation to ensure that the scale $\nu(\mathbf{w})$ and threshold $u(\mathbf{w})$ are strictly positive, i.e., $\nu(\mathbf{w}) > 0, u(\mathbf{w}) > 0$ for all $\mathbf{w}$. For numerical stability, we also ensure that $\xi(\mathbf{w})$ satisfies $\xi(\mathbf{w}) \in (-0.5, 0.1)$ for all angles $\mathbf{w}$. Selection of the remaining tuning parameters is discussed in Section 3.2.3.

3.2.2 Estimating the neural network parameters

To obtain estimates of the MLP parameter sets, $\mathcal{W}_u$ and $\mathcal{W}_{(\nu,\xi)}$, we optimise specified loss functions. Suppose that we have a set of radial and angular observations $\{(r_i, \mathbf{w}_i); i = 1, ..., n\}$. Recall from Section 2.1 that the threshold $u(\mathbf{w})$ is taken to be the quantile of $R|(\mathbf{W} = \mathbf{w})$ at exceedance probability ζ. We thus can estimate $u(\mathbf{w})$ (and its corresponding parameter set $\mathcal{W}_u$) using quantile regression techniques [57]. In this case, an appropriate loss function for $u(\mathbf{w})$ is the tilted loss,

$$\mathcal{L}_u(\mathcal{W}_u) := \sum_{i=1}^{n} \rho_{1-\zeta} \left\{ r_i - u\left(\mathbf{w}_i\right) \right\}, \tag{13}$$

where $\rho_\alpha(t) := t(\alpha - \mathbb{1}\{t < 0\})$ for indicator function $\mathbb{1}$ and where dependency of $u(\mathbf{w})$ on $\mathcal{W}_u$ has been suppressed from notation.

After estimation of $u(\mathbf{w})$, we define $I_u := \{i \in \{1, ..., n\} : r_i > u(\mathbf{w}_i)\}$ as the set of indices of radial threshold exceedances. The MLP that defines the GP parameter functions can be considered as a 'conditional density estimation network', with the negative log-likelihood function used for optimisation; see, e.g., [58, 59]. In this case, we perform

maximum likelihood estimation, with loss

$$\mathcal{L}_{GP}(\mathcal{W}_{(\nu,\xi)}) := -\sum_{i \in I_u} \log \left[f_{\text{GP}} \left(r_i - u(\mathbf{w}_i); \xi(\mathbf{w}_i), \frac{\nu(\mathbf{w}_i)}{\xi(\mathbf{w}_i) + 1} \right) \right]. \tag{14}$$

Optimisation of both losses, (13) and (14), proceeds via stochastic gradient descent and the ADAM algorithm [60]. To mitigate overfitting, data are split into training (80%) and validation (20%) sets, with the latter used to check for parameter convergence. We refer the reader to [49] for a more detailed overview of the fitting procedure.

As noted in Section 2, when the GP shape parameter $\xi(\mathbf{w})$ is negative, the distribution of $R|(\mathbf{W} = \mathbf{w})$ has a finite upper endpoint. Training of a deep GP regression model which permits negative shape parameter values can be computationally troublesome; see discussion by [61]. At a given angle $\mathbf{w}_i$, if $\xi(\mathbf{w}_i) < 0$ and the radial observation r_i exceeds the upper endpoint, i.e., $r_i > u(\mathbf{w}_i) - \sigma(\mathbf{w}_i)/\xi(\mathbf{w}_i)$, then the loss function in (14) will evaluate to a non-finite value. Consequently, the loss surface over which we optimise $\mathcal{W}_{(\nu,\xi)}$ is highly irregular, and iterative gradient descent methods (like ADAM) may have trouble finding global maxima, or may predict out-of-sample parameter estimates that are infeasible, i.e., the loss is non-finite. To circumvent these issues during training, we initialise the MLP to ensure that the shape parameter function $\xi(\mathbf{w})$ is non-negative for all angles $\mathbf{w}$; in this way, at the outset of the training procedure, the loss function is guaranteed to be finite for all $w_i, i = 1, \ldots, n$. Then, if the gradient descent optimisation algorithm produces non-finite loss values during training, we restart training (from the last iteration with finite loss values) with a smaller learning rate. In our application we found that the fully-trained MLP provided negative values of $\xi(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{S}^{d-1}$. We note that the issue of non-finite loss occurs in other likelihood-based inference procedures for GP regression, and is not a problem specific to MLPs.

3.2.3 Selecting an architecture

An important choice when fitting a neural network model is the choice of architecture: this corresponds to the set of hyperparameters introduced in Section 3.2.1. There is no 'best practice' for this selection within the deep GP regression literature [48], and the appropriate architecture is likely to be domain specific; see [56]. Selecting a model with more hidden layers and more neurons results in higher flexibility, but at the cost of increased parameter variability and computational expense. In the spirit of parsimony, we wish to select the simplest model possible while still capturing the observed variability in the threshold and GP parameter functions over the angular domain.

To select our 'optimal' architecture for the application detailed in Section 4, we perform a grid-search over architecture choices. For each configuration, we estimate a range of model fit diagnostics; these are discussed below. The optimal architecture is then chosen as the configuration which visually provides the best model diagnostics. We found that, for both MLP models, a simple architecture is preferable: $L = 3$ hidden layers, with $h_l = 16$ neurons per layer. This results in two MLPs, each comprising approximately 650 estimable parameters; inference for these models is not computationally demanding, and can be conducted on a standard laptop.

As with univariate POT models, selecting a suitable threshold for our model is critical. Too low a threshold will result in the asymptotic arguments motivating the use of the GP model not being applicable, causing bias; whereas too high a threshold will result in too few observations for fitting, resulting in higher variance. The process of threshold selection for our model is directly analogous to that in univariate problems. That is, we fit the model for a range of threshold exceedance probabilities ζ, and check for stability of inferences and goodness of fit. The threshold is then selected as the largest value of ζ for which inferences are approximately stable for $\zeta_0 < \zeta$. In our application, we found that $\zeta = 0.1$ was suitable.

We remark that while our selected architecture works well for our application, we do not advocate the general use of these hyperparameters. Instead, we recommend that practitioners who apply our framework perform a similar grid-search, and use post-fit diagnostics to select the optimal architecture.

4 Application to five-dimensional problem

4.1 Dataset

In this section, we consider the application of the SPAR model to a hindcast dataset consisting of 31-years of wind and wave variables from 01/01/1990 to 31/12/2020, for a site in the Celtic Sea, off the south-west coast of the UK, which has been identified for development of floating wind farm projects. The dataset consists of hourly values of significant wave height (H_s), mean wave period (T_m), mean wave direction (θ_{wave}), hourly mean wind speed at 10 m above sea level (U_{10}), and wind direction (θ_{wind}). Both wind and wave directions are defined as the directions in which the winds or waves are 'going to'. These variables all influence the motion and loading of floating wind turbines,

and understanding their joint extremes is important for design. Since directional variables are periodic, it does not make sense to talk about 'extreme directions'. Instead, we work with the x- and y-components of wave height and wind speed, defined as $H_x = H_s \cos(\theta_{wave})$, $H_y = H_s \sin(\theta_{wave})$, $U_x = U_{10} \cos(\theta_{wind})$, and $U_y = U_{10} \sin(\theta_{wind})$.

Unlike many classical modelling approaches, the SPAR approach extrapolates in all directions, allowing one to perform inference in any extreme region of interest. When variables have a defined lower bound at zero, as is the case for many physical quantities, the SPAR model should be able to infer this directly from the data, and these physical limits should correspond to the upper bounds of the GP model for the radial variable at the relevant angles. However, inferences at endpoints are highly uncertain since they correspond to zero exceedance probability, and consequently the model may infer a slightly negative bound in some directions. It is therefore safer to work with variables that do not have hard lower bounds at zero, so that uncertainties in the radial endpoint do not result in estimates that are not physically possible. The directional variables H_x, H_y, U_x, and U_y are all defined on $(-\infty, \infty)$, although clearly they will have some upper and lower bounds due to physical constraints (in general, we would expect the region of variable space in which the density is non-zero to be bounded, due to physical limits). For the period variable, we address the problem of the lower bound by defining $L_T = \log(T_m) \in (-\infty, \infty)$ and using this variable in the model.

4.2 Normalisation and choice of origin

As different physical variables have different scales, we normalise each variable by its standard deviation. Without this, angles would tend to be clustered in the plane of whichever variable has the largest scale (wind speed in this case). We also need to define an origin in order to transform to polar coordinates. In previous work using SPAR [23, 24], the origin was defined at the mean of each variable. In our application, more care must be taken when defining an appropriate origin. For the model to provide a useful description of the extremes at all angles, the support of the density, $\text{supp}(f_{\mathbf{X}}) := \{\mathbf{x} \in \mathbb{R}^d : f_{\mathbf{X}}(\mathbf{x}) > 0\}$, must be star-shaped with respect to the chosen origin [62]. That is, given an origin $\mathbf{x}_0 \in \text{supp}(f_{\mathbf{X}})$ and any point $\mathbf{x} \in \text{supp}(f_{\mathbf{X}})$, the line segment from $\mathbf{x}_0$ to $\mathbf{x}$ is contained in $\text{supp}(f_{\mathbf{X}})$. Under this assumption, all rays from the origin reach the 'edges' of the distribution without passing through regions which have zero density. In this way, the data-cloud has a well-defined 'inside' and 'outside', with the 'outer' region considered 'extreme', and the GP-based representation of the radial component in this region is reasonable. This assumption can be verified by checking plots of the density of observations along various rays from the origin, e.g., histograms of the observed radial variable within small angular ranges. However, as discussed further below, in higher dimensional spaces, a large number of angles are required to obtain a reasonable coverage of the surface of $\mathbb{S}^{d-1}$, and visual inspection of plots of the radial density over each angular range is time consuming. Ultimately a data-driven approach for selecting an optimal choice of origin would be best.

In this example, we use knowledge of the physical processes to define an appropriate origin. Firstly, wave breaking limits the maximum possible wave height for a given wave period, with the limit related to wave steepness given by $s = 2\pi H_s/(gT_m^2)$. In the three-dimensional subspace containing the variables (H_x, H_y, T_m), this results in a conical-shaped bound on the data, centred along the axis $H_x = H_y = 0$, with the radius of the cone given by $H_{s,max} = s_{max} g T_m^2/(2\pi)$, where s_{max} is the limiting steepness. The value of s_{max} depends on the water depth and wind speed (among other factors), but the maximum value was found to be around $s_{max} \approx 0.08$ for our dataset. This conical shape to the distribution (see Figures 3 and 7) suggests that an appropriate choice of origin should be somewhere on the axis $H_x = H_y = 0$. Due to the physical dependence of wave height on wind speed, we also locate the origin at $U_x = U_y = 0$. The choice of origin for L_T is somewhat arbitrary, but experimentation showed that using the mean value of L_T gave satisfactory results.

The angular and radial variables are therefore defined with respect to the normalised variables given by $X_1 = U_x/\text{STD}(U_x)$, $X_2 = U_y/\text{STD}(U_y)$, $X_3 = H_x/\text{STD}(H_x)$, $X_4 = H_y/\text{STD}(H_y)$, and $X_5 = (L_T - \text{mean}(L_T))/\text{STD}(L_T)$, where $\text{STD}(\cdot)$ denotes the standard deviation function. The pairwise relations between these normalised variables are shown in Figure 3. A radial grid has been overlaid to illustrate that these two-dimensional projections are approximately star-shaped with respect to this origin. Lines of bounding steepness $s = 0.08$ are shown in the plots of (X_3, X_5) and (X_4, X_5) as dashed lines. It can be seen that there is far less scatter in the variables close to these bounds due to the physical limitations. Another feature of the data that is evident is the strong positive correlation between the x- and y-components of the wave height and wind speed (i.e., the pairs (X_1, X_3) and (X_2, X_4)).

4.3 Exploratory data analysis

Before assessing the model fit, it is useful to consider various visualisations of the data, in order to understand how it is distributed over the five-dimensional space. The plots below the diagonal in Figure 3 show the empirical densities of pairs of angular components (W_i, W_j), $i, j \in \{1, ..., d\}$. By definition, these variables must fall within the unit circle. Any large gaps in the observed values indicate that it is not possible to fit the model at these angles. The objective

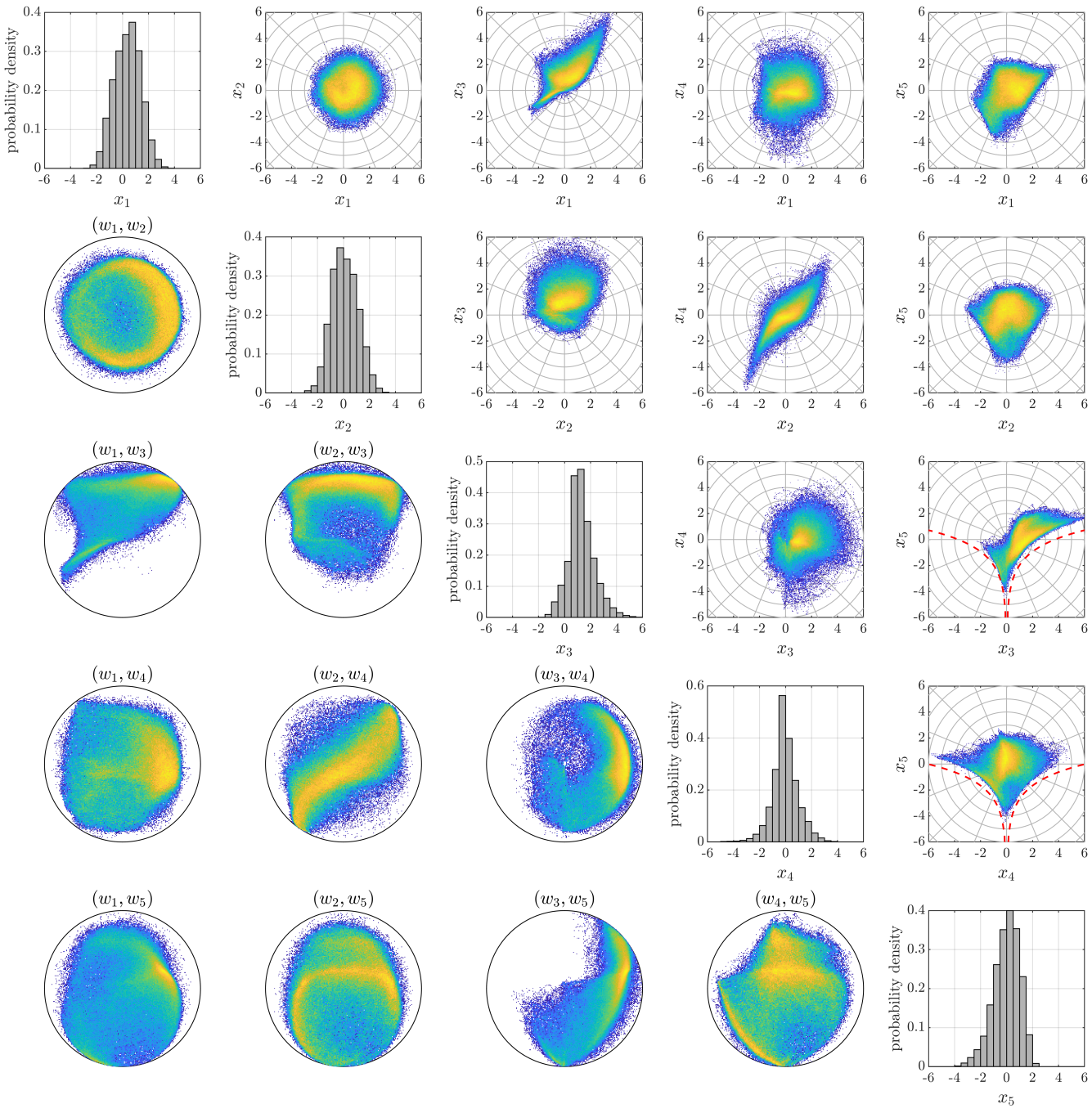

Figure 3: Upper right plots: Empirical joint densities of pairs of normalised observations. Dashed red lines in plots of (x_3, x_5) and (x_4, x_5) are lines of constant wave steepness $s = 0.08$. Plots on diagonal are histograms of each normalised variable. Lower left plots: Empirical density of pairs of angular components of observed data (yellow = high density, blue = low density). The normalised variables $(X_1, ..., X_5)$ correspond to normalised $(U_x, U_y, H_x, H_y, \log(T_m))$.

of using the MLP model of the angular variation of the radial distribution is to estimate a relatively smooth variation with angle. The model should therefore be able to smooth over small angular ranges with no observations. However, the model is unlikely to be able to accurately estimate the behaviour of the radial component over large angular ranges where there is little or no data.

Consider the joint occurrence of wave direction and wind direction, illustrated in Figure 4. Note that the wind direction is $\theta_{wind} = \text{atan2}(W_1, W_2)$ and $\theta_{wave} = \text{atan2}(W_3, W_4)$, where $\text{atan2}(x, y)$ is the four-quadrant inverse tangent function. So the angles shown in Figure 4 are a subset of $\mathbb{S}^3$ where $w_1^2 + w_2^2 = w_3^2 + w_4^2 = 1/2$ (known as the Clifford torus). As discussed above, wind and wave directions tend to be roughly aligned, although there is some scatter. However, there are large areas of the variable space where observations are sparse. So attempting to model the conditional joint distribution of three other variables (wind speed, wave height, wave period), let alone their joint extremes, is likely to be very challenging in these regions.

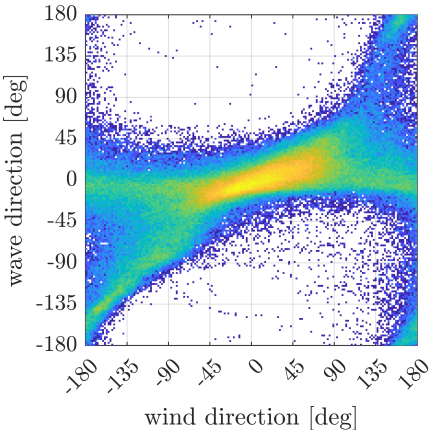

Figure 4: Empirical joint density of wind direction and wave direction. (Blue = low density, yellow = high density).

To assess the variation in the density of angles around a circle, we could plot the number of observations within discrete angular ranges as a histogram. In higher dimensions, visualisation becomes more difficult, but a similar approach can be taken. We count the number of observations within a small angular range of a pseudo-regularly spaced grid of points on the sphere. It is not possible to define evenly-spaced points on the surface of a (hyper) sphere in three or more dimensions. To address this issue we define a set of points U that are approximately regularly-spaced points on $\mathbb{S}^{d-1}$, using the method described in Appendix A. For each $\mathbf{u} \in U$ we count the number of observed angles with $\arccos(\mathbf{w}_i^\top \mathbf{u}) < \theta_{max}$, where θ_{max} is some prescribed range. Although there may be some overlap between the ranges defined above, this analysis still gives an indication of how the density of angles varies over the sphere.

Figure 5 shows the empirical cumulative distribution function (CDF) of the number of observations in a cell chosen at random, for a cell radius of $\theta_{max} = 15°$ and a set of 1002 direction vectors (generated using $m = 5$ points along each dimension – see Appendix A). One feature that is evident is that over 50% of angular cells contain no observations. This is due to the particular choice of origin, which was selected to meet the assumption of a star-shaped distribution. Figure 5 also shows the number of observations in each of the $2^5 = 32$ orthants in $\mathbb{R}^5$ (an orthant is the d-dimensional analogue of a quadrant of the plane). There are five orthants which contain no observations, and a further five which contain fewer than 100 observations (out of a total of 271,704 observations). The SPAR model estimates of the extremes in the orthants with little data will therefore be highly uncertain. However, Figure 6 shows scatter plots of the maximum observed values of H_s and U_{10} in each of the 1002 angular cells against the corresponding number of observations in the cell, indicating that larger values of these variables tend to coincide with higher angular densities. The higher uncertainties associated with the lower occurrence regions should therefore have less effect on the global extremes.

Figure 7 shows a scatter plot of T_m against H_x and H_y. (A plot of the normalised variables (X_3, X_4, X_5) would look similar, but the non-normalised variables are shown here to aid physical interpretation). The conical shaped bound imposed by the limiting wave steepness is evident. Another feature that is apparent is that the data cloud is hollow on the side $H_x < 0$. This is because of fetch limitations in this direction, meaning that waves propagating towards the west are steeper wind-driven waves, so that H_s and T_m are strongly correlated in this region. This violates the assumption of the distribution being star-shaped with respect to the choice of origin. However, the 'edge region' that is not modelled corresponds to low values of H_s at a given T_m and wave direction, which is less critical for extreme responses. For the present choice of origin at $(H_x, H_y, T_m) = (0, 0, 6.2)$, it appears that there are some sharp changes

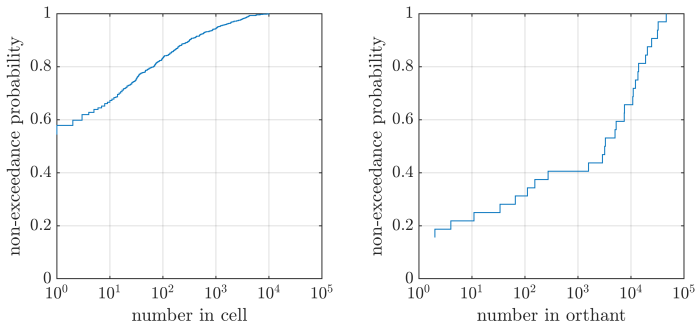

Figure 5: Left: Empirical distribution of the number of points within a 15° radius of each direction vector. Right: empirical distribution of the number of points in each of the 32 orthants of $\mathbb{R}^5$.

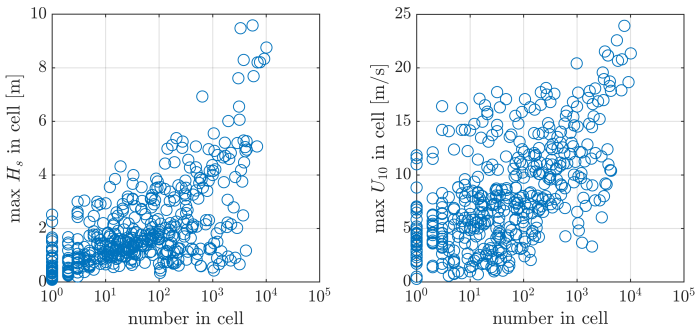

Figure 6: Maximum observed H_s and U_{10} in each local angular cell vs. number in cell.

in the distribution with the vertical angle, for rays towards the negative x direction. This might be hard to capture with an ANN. Nevertheless, we can assess how well the model performs, given these challenges.

4.4 Assessment of angular density model

The assessments of the models for both the angular and conditional radial distributions are derived from a sample simulated from the fitted model, of 100 times the size of the original sample (i.e., 2.71×10^7 samples, or approx. 3000 years of hourly data). The larger sample size reduces the uncertainty in estimates in the lower-probability regions, compared to the observed sample, and allows us to assess how well the model extrapolates from the observations. Figure 8 shows a comparison of the empirical joint densities of paired angular components (w_i, w_j), $i, j \in \{1, ..., 5\}$, with contours of the corresponding joint densities from the KD model. For the empirical density of the observations, the colour resolution has been restricted so that the boundaries of the coloured regions correspond to the contour levels from the KD model. That is, contours from the KD model have been computed at density levels $10^{-3}, 10^{-2.5}, ..., 10^{-0.5}$, and for the observations, all bins where the estimated density is in the interval, say $[10^{-2.5}, 10^{-2})$, have been assigned the same colour. Overall, the KD model compares very well with the observed densities, with the contours following the complex patterns in the observations. The smoothing effect of the model on the observations is evident, especially at low density levels.

Whilst Figure 8 shows that the KD model provides a good representation of dependence between pairs of angles, it does not consider how the model performs locally on small regions of $\mathbb{S}^4$. To motivate how this can be assessed in five dimensions, first consider the one-dimensional case. Figure 9 shows the empirical marginal densities of each angular component depicted as histograms, compared with the marginal estimates from the KD model. When we assess the fit of the model in this way, we are making a visual assessment of the whether the number of observations in each bin agrees with the number predicted by the model. This idea can readily be extended to higher dimensions, by partitioning the space into discrete regions and comparing the number of observations in a region, with the expected number in that region from the model. In this case, we use a Voronoi partition relative to a set of pseudo-regularly spaced direction vectors, generated using the method described in Appendix A. A Voronoi partition is a partitioning of a space, where each partition is the set of points closest to a given reference point. Figure 10 shows an example partitioning of the sphere in $\mathbb{R}^3$ relative to a set of 102 pseudo-regularly spaced direction vectors.

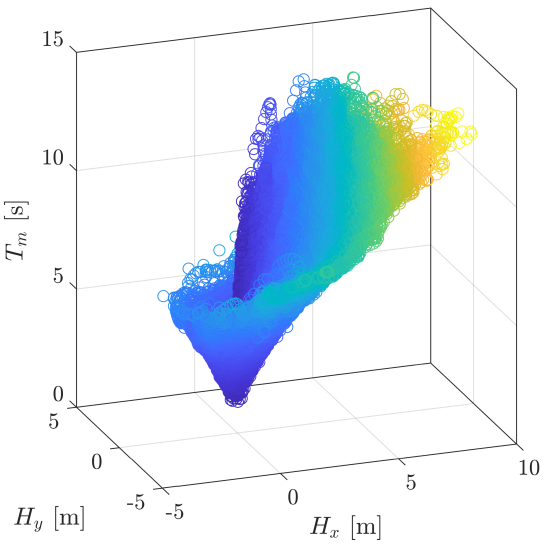

Figure 7: Scatter plot of observations of x-y components of H_s and T_m. Colours indicate the value of $H_s = \sqrt{H_x^2 + H_y^2}$ (blue=low, yellow=high).

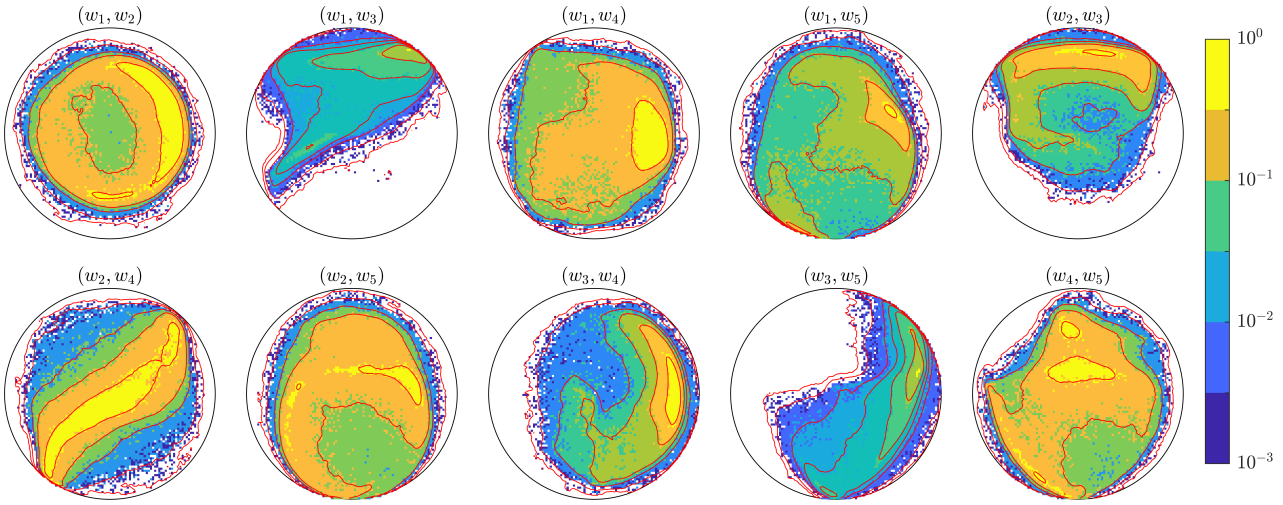

Figure 8: Comparison of empirical joint densities of pairs of angular components (coloured plots) with contours of the joint densities from a sample from the KD model (red lines). Note that the colour resolution has been restricted so that the boundaries of the coloured regions correspond to the contour levels from the KD model.

Suppose that we have a partitioning of $\mathbb{S}^{d-1}$ into non-overlapping cells, together with a model from which we can calculate the probability, p_i, that an observation chosen at random falls into cell i. If the observations are independent, the number of observations, N_i, falling into cell i follows a binomial distribution with CDF

$$\Pr(N_i \leq k) = \sum_{j=0}^{k} \binom{n}{j} p_i^j (1-p_i)^{n-j}, \quad k \in \mathbb{N}, \tag{15}$$

where n is the number of independent observations. In our application, the data are serially correlated, and hence violate the above assumption of independence. To mitigate for this, we down-sample the data to take one observation per day, resulting in a sample size of $n = 11321$ independent observations.

Equation (15) can be used to add confidence bounds to scatter plots of the observed number in a cell against the expected number predicted by the model. Figure 11 shows a scatter plot of the number of observations in each cell

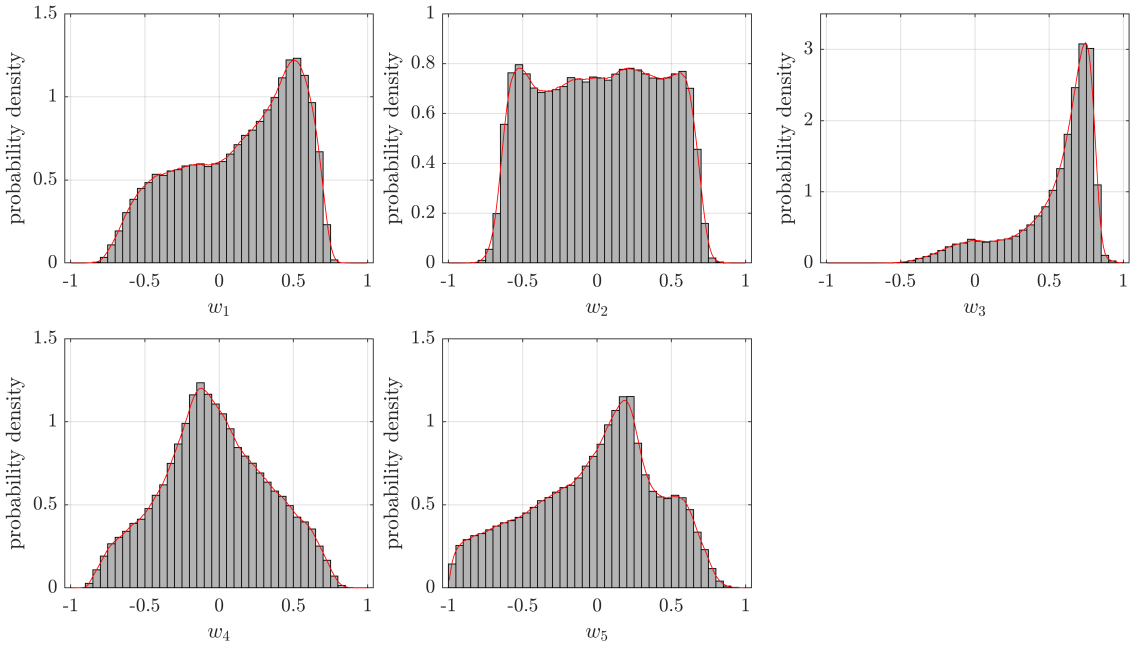

Figure 9: Comparison of empirical marginal densities of angular components (histograms) with marginal densities from the KD model (red lines).

of a Voronoi partition of $\mathbb{S}^4$ against the expected number from the KD model. The expected number in each cell is calculated as $\mathrm{E}(N_i) = np_i$, where n is the observed (down-sampled) sample size and p_i is the fraction of the simulated sample falling into cell i. The Voronoi partition is based on a set of 5890 direction vectors, generated using $m = 8$ in the method described in Appendix A. However, only 1001 of the resulting cells contained observations, due to the non-uniformity of the angular distribution, discussed above. Overall, the model does very well at replicating the local characteristics of the density, both in the high and low density regions, with differences attributable to sampling variability.

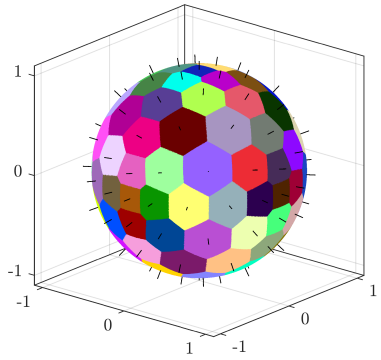

Figure 10: Example of Voronoi partitioning of the sphere in $\mathbb{R}^3$ into non-overlapping cells, corresponding to the set of points closest to each direction vector (black lines).

4.5 Assessment of conditional radial model

We assess the quality of the conditional radial model in several ways, analogous to the diagnostics for the angular model. The ability of the model to recreate the pairwise relations is considered in Figure 12, which shows empirical joint densities of pairs of the observed variables (on the orignal scale) compared with contours of the joint density from the fitted model. Although the SPAR model gives an explicit representation of the joint density, this cannot be projected into lower dimensions without computationally-expensive integration. Instead, the joint densities from the

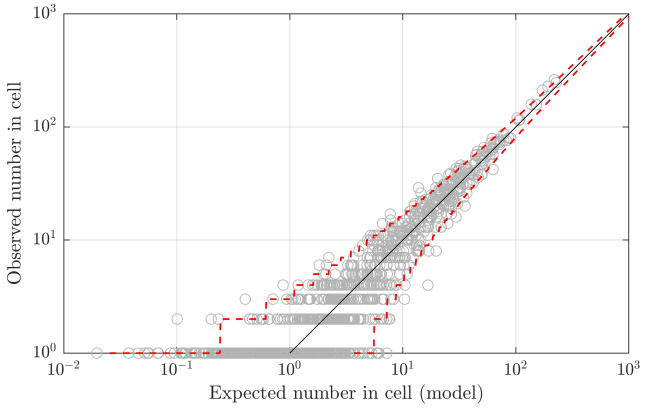

Figure 11: Comparison of observed number of data points in a local angular cell against the expected number in the cell from the KD model, based on a Voronoi partition of $\mathbb{S}^4$. Solid black line shows 1:1 relation, red dashed lines show 95% confidence interval on observed number based on sampling variability.

model are estimated empirically from a simulated sample from the model. As such, the density shown is subject to sampling uncertainty in the low density regions, but less than that for the observations, due to the larger simulated sample size. Within the body of the data (i.e., below the threshold level) the observations are resampled rather than modelled. Good agreement between the model contours and observations is therefore to be expected in the bodies of the data clouds. However, the density contours in the outer 'extreme' regions correspond to points which are simulated from the non-stationary GP model. Overall, the simulated relationships between variables closely follow those of the observations. The model predicts some larger values of steep waves than observed. However, these occur at lower wave heights, so are less critical for design purposes.

As the SPAR model does not model the tails of the marginal variable directly, it is also useful to assess how well the simulated data matches the observed marginal tails. Figure 13 shows exceedance and non-exceedance probabilities on a logarithmic scale for the five variables used in the model, as well as the derived variables H_s, U_{10}, and s. The model predicted some unrealistically steep conditions, so we have removed data points where steepness exceeded a value of 0.1. This corresponded to approximately 0.01% of the simulated sample and small wave heights ($H_s < 3$ m). With these points removed, overall, the model performs well and provides a good match for both the upper and lower observed tails. There is a slight tendency to underestimate the upper tail of H_x, which results in a slight underestimation of the upper tail of H_s.

Finally, to give an indication of the variation in model performance over the angular domain, Figure 14 shows QQ plots of simulated against observed threshold exceedances, binned over small angular ranges. As above, we have used a grid of 1002 fixed angles and a radius of 15° to define the angular bins. Only bins with 200 or more observations have been used, so that there are approximately 20 or more threshold exceedances per bin. To account for the variation of the threshold level over the bins, the threshold is taken as the empirical quantile at exceedance level ζ from the simulated dataset (this approximates an average of the non-stationary threshold over the bin). The simulated data are then interpolated to the same probability levels as the observations. There is some scatter between the observed and simulated values. However, the aggregated trend shows good agreement, although with a slight tendency to underestimate at larger values. Although the total sample size is relatively large, the number of observations in local angular regions can be quite small. The underestimation from the model may therefore be related to the use of maximum likelihood estimation, which is known to produce a slight negative bias in quantile estimates for small sample sizes [63].

5 Discussion and conclusions

In this work, we have introduced a deep learning framework for inference with the SPAR model. The computational scalability and robustness of neural networks result in a modelling approach which requires very few assumptions, offers a high degree of flexibility, and can be applied in higher-dimensional settings compared to existing techniques. We use our approach to approximate the complex joint tail behaviour of a five dimensional metocean dataset, with diagnostics indicating our model is able to accurately represent the observed dependence structure. Given the complex dependence observed in the data, the MLP model for the angular variation of the GP parameters performs very well

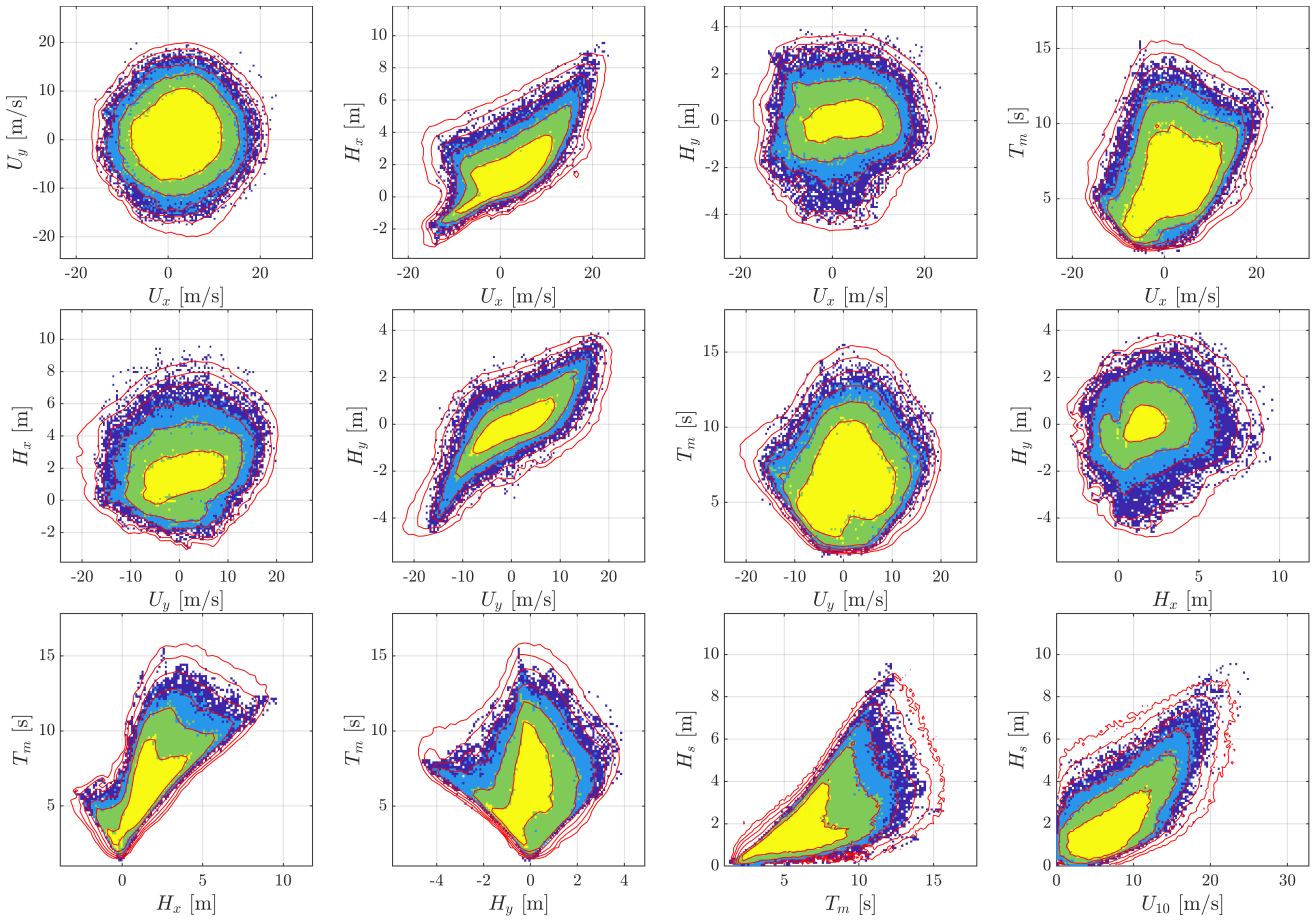

Figure 12: Comparison of empirical joint densities of pairs of variables (coloured plots) with contours of the joint densities from a sample from the SPAR model of 100 times the original sample size (red lines). Note that the colour resolution has been restricted so that the boundaries of the coloured regions correspond to the contour levels from the SPAR model. Contours at equal logarithmic increments.

in representing the 'extremes' of the dataset. By 'extremes', we are referring not just to the largest and smallest values of each observed variable, but anywhere on the outer part of the data cloud. Moreover, the SPAR model provides an asymptotically justified basis for extrapolating outside the range of observations. Simulation from the fitted SPAR model subsequently allows ones to generate large, physically-realistic event sets, allowing practitioners to easily perform robust risk assessments and estimate probabilities of structural failure.

We note that selecting an architecture for any neural network is non-trivial, and care must be taken to ensure the resulting model offers sufficient flexibility without overfitting. We also remark that, in general, neural networks require a large amount of data for accurate model fitting [56]; this is generally a challenge for modelling extremes since, by definition, very little data are available. It is currently not clear what sample sizes, tuning parameters, or architectures are required to accurately fit the SPAR model via deep learning, and this should be explored in further work. In the present approach, we have used 80% of the data for training and 20% for validation. The validation data are used to avoid overfitting. It is possible that using only 20% of observations for validation is insufficient to force a sufficiently smooth solution in regions of sparse angular observations. An alternative is to use a full cross-validation scheme, in which the model results are averaged over, e.g., five fits using a different 20% of the data for testing. This would make better use of the limited observations, although with increased computational cost. Averaging over multiple plausible candidate models, provides a general means of accommodating uncertainty due to modelling choices that are somewhat arbitrary, like the location of the polar origin, the threshold level, and neural network architecture.

One notable observation from Section 4 was that the choice of origin for defining the SPAR model was non-trivial. Initially, we naively assumed that the componentwise mean was a suitable origin, but this resulted in issues when fitting the model, as the data cloud was not star-shaped with respect to this initial choice of origin. Selecting an appropriate origin for the SPAR model beyond the lower dimensional ($d \leq 3$) setting remains a challenge as full data

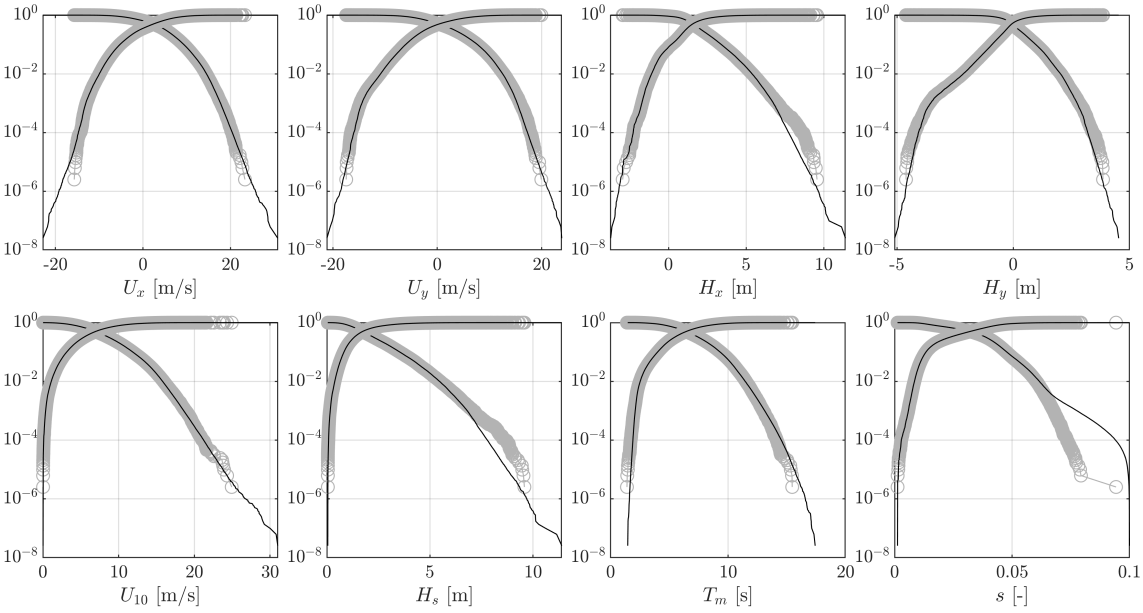

Figure 13: Exceedance and non-exceedance probabilities of marginal and related variables for observed (grey circles) and simulated (black lines) data.

visualisation is not possible, and inappropriate choices can invalidate the modelling assumptions. Future work could explore the robust selection of the origin in a data-driven manner, removing the need for domain-specific knowledge.

In the current work, we have included the wind and wave directions in the model by considering the x- and y-components of the wind speed and wave height. Another way to model the five considered variables is to treat the wind and wave directions as covariates, and then model the joint distribution of wind speed, wave height, and wave period conditional on these two covariates. Initial investigations suggest that this could resolve some of the issues in the current approach caused by the complex shape of the data cloud. If the origin of the polar coordinate system is allowed to vary with wind and wave direction, then the remaining three-dimensional data clouds appear more star-shaped, potentially resulting in a better fit for the model. This will be investigated in future work. Note that if we were to model a two-dimensional case of say wave height and direction, using the SPAR approach, then this is identical to fitting a non-stationary model to H_s with wave direction as a covariate. Indeed, this was partially the motivation for the SPAR approach. However, when more variables are involved, treating directions as covariates or considering the joint distribution of x- and y-components does represent different sets of modelling assumptions.

From Section 4, it was also clear that as the dimension of the data increases, so too does the sparsity. This is particularly apparent in Figure 5, where we observed some orthants containing just a handful of observations. Inference within such regions is problematic, since there is very little data to train the neural network model. We also noted that in sparse regions, the resulting parameter estimates did not always respect the physical features of the data, e.g., upper bounds of the variables. Such issues are not unique to the SPAR model, and one would expect to encounter the same problems with alternative modelling approaches when applied in high dimensions. The effect of dimensionality on data sparsity is not necessarily intuitive to understand. For example, generating pseudo-regularly spaced points on $\mathbb{S}^4$ in the manner described in Appendix A, with an average spacing of approximately 4.5° (corresponding to $m = 20$ points along the positive half of each dimension), results in 216,002 direction vectors. This corresponds to $216,002/(24 \times 365.25) \approx 24.6$ years of hourly observations – a similar order of magnitude to the size of dataset considered here (although, as discussed above, serial correlation in the data reduces the effective sample size). Compare this to the spacing of the same number of direction vectors in two dimensions, which would be $360/216022 \approx 0.0017°$. So, a dataset that gives a high directional coverage in two dimensions, gives a much more sparse directional coverage in five dimensions. Moreover, the sparsity increases exponentially with the number of dimensions, meaning that this type of analysis (or any multivariate density modelling method) will require an exponential increase in dataset length to maintain a similar level of mean spacing between observed angles. Future work could explore whether notions of sparsity can be incorporated into the SPAR framework to improve the robustness and efficiency of the model.

In the present implementation, the model is only estimated for observations above the threshold level, with the threshold exceedance probability, ζ, chosen by testing models fitted with different values of ζ. A possible extension,

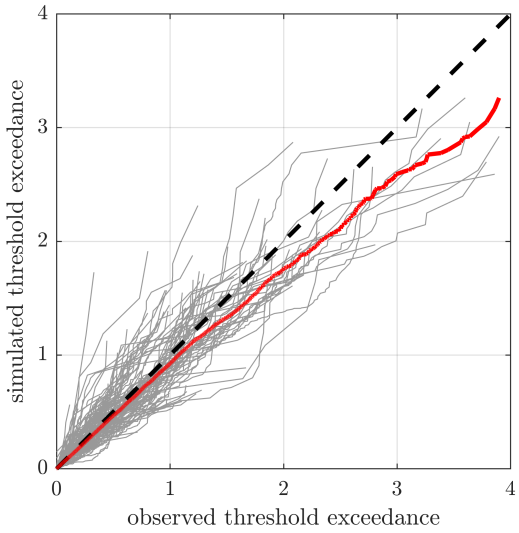

Figure 14: QQ plots of simulated versus observed threshold exceedances in bins of 15° radius from a grid of points on $\mathbb{S}^4$. Grey lines are individual bins; red line is aggregated value over all bins; black dashed line is equality.

would be to model the entire radial distribution with a blended model that takes a semi-parametric form in the bulk of the distribution and a parametric GP form in the tail, similar to [64]. This has the advantage that the uncertainty in threshold estimation is incorporated in the inference.

Finally, we note that although we have accounted for serial correlation in the data in some of the model fitting and diagnostics, the resulting SPAR model is a model for all observations. As such, when we simulate from the model, we do not account for the serial correlation. This could potentially be addressed by using a two-step simulation procedure, where we first simulate a point from the model, then simulate a section of time series which passes through the simulated point. The time series simulation could be accomplished using a resampling approach [65] or by modelling the temporal evolution of variables conditional on a peak value in a similar manner to [66]. This will be investigated in future work.

Acknowledgment

EM was funded by the EPSRC Supergen Offshore Renewable Energy Hub, United Kingdom [grant no: EP/Y016297/1]. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/)

Appendix A Pseudo-regularly spaced points on the hypersphere

To generate a grid of pseudo-regularly spaced points on the hypersphere, we take the approach proposed in [3], and define a regular grid of points on the L^1 sphere and project this onto the surface of the L^2 sphere. This is computed by creating a regular grid of points in the cube $[-1,1]^d$, with $2m+1$ points along each dimension: $U_{grid} = \{(i_1/m, ..., i_d/m) : i_j \in \{-m, ..., m\}, j = 1, ..., d\}$. Then, we define $U_1 = \{\mathbf{u} \in U_{grid} : \|\mathbf{u}\|_1 = 1\}$, where $\|\cdot\|_1$ is the L^1 norm, given by $\|\mathbf{u}\|_1 = |u_1| + \cdots + |u_d|$ for $\mathbf{u} = (u_1, ..., u_d)$. Finally, we compute a set of direction vectors $U \subset \mathbb{S}^{d-1}$ by $U = \{\mathbf{u}/\|\mathbf{u}\|_2 : \mathbf{u} \in U_1\}$. This is illustrated in Figure 15 for the case $m = 5$ and $d = 3$.

Appendix B Simulation from the power spherical distribution

A method for simulating from the PS distribution was presented in [40, Algorithm 1]. Here we present some background details on simulating from rotationally-symmetric distributions on $\mathbb{S}^{d-1}$, to illustrate why the PS distribution offers computational advantages over the more commonly-used von Mises-Fisher (vMF) distribution. Further details can be found in standard texts on directional statistics, e.g., [28, 29]. An essential tool for modelling random vectors on the

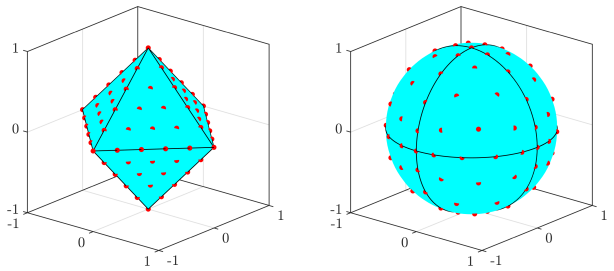

Figure 15: Illustration of the mapping of regularly spaced vectors on the surface of the L^1 unit sphere (left) onto the L^2 unit sphere (right).

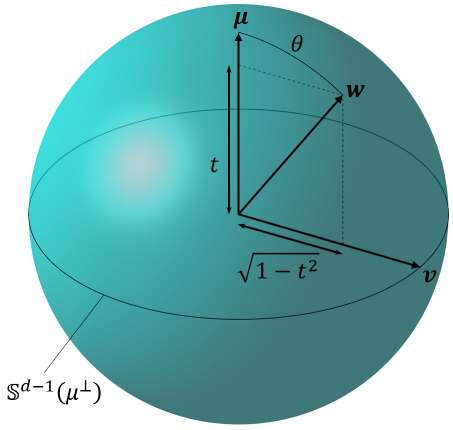

Figure 16: Illustration of tangent-normal decomposition of a vector $\mathbf{w} \in \mathbb{S}^2$ relative to $\boldsymbol{\mu} \in \mathbb{S}^2$.

sphere is the *tangent-normal decomposition*, illustrated in Figure 16. For any vectors $\mathbf{w}, \boldsymbol{\mu} \in \mathbb{S}^{d-1}$ we can express $\mathbf{w}$ in terms of components that are aligned to $\boldsymbol{\mu}$ and tangential to it:

$$\begin{aligned} \mathbf{w} &= t\boldsymbol{\mu} + \sqrt{1-t^2}\mathbf{v}, \\ &= \cos(\theta)\boldsymbol{\mu} + \sin(\theta)\mathbf{v}, \end{aligned}$$

where $t = \cos(\theta) = \mathbf{w}^\top \boldsymbol{\mu}$ and $\mathbf{v} = (\mathbf{w} - t\boldsymbol{\mu})/\|\mathbf{w} - t\boldsymbol{\mu}\|_2$. The unit vector $\mathbf{v}$ lies in the tangent space to $\boldsymbol{\mu}$, $\mathbb{S}^{d-1}(\boldsymbol{\mu}^\perp) \coloneqq \{\boldsymbol{x} \in \mathbb{S}^{d-1} : \boldsymbol{x}^\top \boldsymbol{\mu} = 0\}$.

A density $f : \mathbb{S}^{d-1} \to [0, \infty)$ that is rotationally symmetric about its location $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ can be expressed in terms of an angular function $f_a : [-1, 1] \to [0, \infty)$, as $f(\mathbf{w}) = f_a(\mathbf{w}^\top \boldsymbol{\mu})$ for all $\mathbf{w} \in \mathbb{S}^{d-1}$. That is, the density at $\mathbf{w}$ just depends on the distance from $\mathbf{w}$ to $\boldsymbol{\mu}$ along the surface of the sphere. The tangent-normal decomposition of a rotationally symmetric random vector leads to the following useful result [67]. Suppose that random vector $\mathbf{W} \in \mathbb{S}^{d-1}$ has density f that is rotationally symmetric about $\boldsymbol{\mu}$, with corresponding angular function f_a. Define the tangent and normal components of $\mathbf{W}$ as above, with $T = \mathbf{W}^\top \boldsymbol{\mu}$ and $\mathbf{V} = (\mathbf{W} - T\boldsymbol{\mu})/\|\mathbf{W} - T\boldsymbol{\mu}\|_2$. Then $T \in [-1, 1]$ and $\mathbf{V} \in \mathbb{S}^{d-1}(\boldsymbol{\mu}^\perp)$ are independent and $\mathbf{V}$ is uniformly distributed in $\mathbb{S}^{d-1}(\boldsymbol{\mu}^\perp)$. Moreover, T has density

$$f_T(t) = \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} f_a(t)(1-t^2)^{(d-3)/2}. \tag{16}$$

The distribution of T is referred to as the marginal distribution of $\mathbf{W}$. In words, this result states that we can represent any rotationally symmetric distribution, in terms of one component that describes the distance from the mean direction, and an orientation relative to the mean direction. Due to the rotational symmetry, the distance and orientation are independent.

The tangent-normal decomposition gives a computationally efficient method for simulating from rotationally symmetric distributions [68]. First note that if $\boldsymbol{e}_d = (0, ..., 0, 1)^\top$, then $\mathbb{S}^{d-1}(\boldsymbol{e}_d^\perp)$ is the embedding of $\mathbb{S}^{d-2}$ into $\mathbb{R}^d$ (see Figure 16 - in this example $\mathbb{S}^2(\boldsymbol{e}_3^\perp)$ is the embedding of the unit circle into $\mathbb{R}^3$). Therefore, if we generate a random

T from the marginal distribution of f and a uniform random $\mathbf{V} \in \mathbb{S}^{d-2}$, then $\boldsymbol{Y} = (T; \sqrt{1-T^2}\mathbf{V})$ has density f with location vector $\boldsymbol{e}_1 = (1, 0, ..., 0)^\top$. In then remains to rotate $\boldsymbol{Y}$ to the appropriate orientation $\boldsymbol{\mu}$. This can be achieved using a Householder transformation [69]. The Householder transformation describes a reflection about a (hyper)plane containing the origin with with unit normal vector $\boldsymbol{u}$. The transformation can be expressed in terms of the Householder matrix, defined as $\mathbf{P} = \mathbf{I}_d - \boldsymbol{u}\boldsymbol{u}^T$, where $\mathbf{I}_d$ is the identity matrix and $\boldsymbol{u}\boldsymbol{u}^\top$ is the outer product. For our application, the plane we wish to reflect about has normal vector $\boldsymbol{u} = (\boldsymbol{e}_1 - \boldsymbol{\mu})/\|\boldsymbol{e}_1 - \boldsymbol{\mu}\|_2$. Finally, a random vector with mean direction $\boldsymbol{\mu}$ is obtained as $\mathbf{W} = \mathbf{P}\boldsymbol{Y}$.

Simulation of uniformly distributed points on $\mathbb{S}^{d-2}$ is straightforward. If $N_1, ..., N_{d-1}$ are independent standard normal variables, then $\mathbf{V} = (N_1, ..., N_{d-1})/\|(N_1, ..., N_{d-1})\|_2$ is uniformly distributed on $\mathbb{S}^{d-2}$ [70]. The key feature of the PS distribution that makes it efficient to simulate from, is that the marginal variable T is defined in terms of an affine transformation of a Beta-distributed variable [40]. Namely, $T = 2Z - 1$, where $Z \sim \text{Beta}(\alpha, \beta)$ with $\alpha = \kappa + (d-1)/2$ and $\beta = (d-1)/2$, where κ is the bandwidth parameter (this follows from substituting the expression for the density of the PS distribution (11) into (16)). In contrast, the marginal distribution of the vMF distribution is not related to common distribution types, meaning that rejection sampling is necessary to sample from the marginal, which is computationally inefficient.

References

[1] S. Haver and S. R. Winterstein, "Environmental contour lines: A method for estimating long term extremes by a short term analysis," in *SNAME Maritime Convention, Paper Number: SNAME-SMC-2008-067*, 2008. DOI: 10.5957/SMC-2008-067.

[2] Q. Derbanne and G. de Hauteclocque, "A new approach for environmental contour and multivariate de-clustering," in *38th International Conference on Ocean, Offshore and Arctic Engineering*, Glasgow, 2019, OMAE2019/95993. DOI: 10.1115/OMAE2019-95993.

[3] E. Mackay and G. de Hauteclocque, "Model-free environmental contours in higher dimensions," *Ocean Engineering*, vol. 273, p. 113 959, 2023. DOI: 10.1016/j.oceaneng.2023.113959.

[4] A. F. Haselsteiner, M. Frieling, E. Mackay, A. Sander, and K.-D. Thoben, "Long-term extreme response of an offshore turbine: how accurate are contour-based estimates?" *Renewable Energy*, vol. 181, pp. 945–965, 2022. DOI: 10.1016/j.renene.2021.09.077.

[5] 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*, vol. 311, p. 118 754, 2024. DOI: 10.1016/j.oceaneng.2024.118754.

[6] E. Ross *et al.*, "On environmental contours for marine and coastal design," *Ocean Engineering*, vol. 195, p. 106 194, 2020. DOI: 10.1016/j.oceaneng.2019.106194.

[7] A. F. Haselsteiner *et al.*, "A benchmarking exercise for environmental contours," *Ocean Engineering*, vol. 236, p. 109 504, 2021. DOI: 10.1016/j.oceaneng.2021.109504.

[8] E. M. Bitner-Gregersen, "Joint met-ocean description for design and operations of marine structures," *Applied Ocean Research*, vol. 51, pp. 279–292, 2015. DOI: 10.1016/j.apor.2015.01.007.

[9] J.-T. Horn, E. Bitner-Gregersen, J. R. Krokstad, B. J. Leira, and J. Amdahl, "A new combination of conditional environmental distributions," *Applied Ocean Research*, vol. 73, pp. 17–26, 2018. DOI: 10.1016/j.apor.2018.01.010.

[10] Z. Cheng, E. Svangstu, T. Moan, and Z. Gao, "Long-term joint distribution of environmental conditions in a Norwegian fjord for design of floating bridges," *Ocean Engineering*, vol. 191, p. 106 472, 2019. DOI: 10.1016/j.oceaneng.2019.106472.

[11] M. L. Simão, L. V. S. Sagrilo, and P. M. Videiro, "A multi-dimensional long-term joint probability model for environmental parameters," *Ocean Engineering*, vol. 255, p. 111 470, 2022. DOI: 10.1016/j.oceaneng.2022.111470.

[12] G. de Hauteclocque, E. Mackay, and E. Vanem, "Quantitative comparison of environmental contour approaches," *Ocean Engineering*, vol. 245, p. 110 374, 2022. DOI: 10.1016/j.oceaneng.2021.110374.

[13] H. Joe, *Dependence modeling with copulas*. CRC Press, 2015, pp. 1–457.

[14] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of Extremes: Theory and Applications*. Wiley, Chichester, UK, 2004.

[15] A. W. Ledford and J. A. Tawn, "Modelling dependence within joint tail regions," *Journal of the Royal Statistical Society: Series B (Methodology)*, vol. 59, no. 2, pp. 475–499, 1997. DOI: 10.1111/1467-9868.00080.

[16] J. L. Wadsworth and J. A. Tawn, "A new representation for multivariate tail probabilities," *Bernoulli*, vol. 19, no. 5 B, pp. 2689–2714, 2013. DOI: 10.3150/12-BEJ471.

[17] R. Huser, T. Opitz, and J. L. Wadsworth, "Modeling of spatial extremes in environmental data science: Time to move away from max-stable processes," *Environmental Data Science*, vol. 4, e3, 2025. DOI: 10.1017/eds.2024.54.

[18] J. E. Heffernan and J. A. Tawn, "A conditional approach for multivariate extreme values," *Journal of the Royal Statistical Society: Series B (Methodology)*, vol. 66, pp. 497–546, 2004. DOI: 10.1111/j.1467-9868.2004.02050.x.

[19] Y. Liu and J. A. Tawn, "Self-consistent estimation of conditional multivariate extreme value distributions," *Journal of Multivariate Analysis*, vol. 127, pp. 19–35, 2014. DOI: 10.1016/j.jmva.2014.02.003.

[20] R. Towe, D. Randell, J. Kensler, G. Feld, and P. Jonathan, "Estimation of associated values from conditional extreme value models," *Ocean Engineering*, vol. 272, p. 113 808, 2023.

[21] E. Mackay, "Improved models for multivariate metocean extremes," Supergen ORE Hub, Tech. Rep., Jan. 2022. Available: https://supergen-ore.net/uploads/resources/IMEX_final_project_summary.pdf.

[22] E. Mackay and P. Jonathan, "Modelling multivariate extremes through angular-radial decomposition of the density function," *arXiv preprint arXiv:2310.12711*, Oct. 2023.

[23] C. J. R. Murphy-Barltrop, E. Mackay, and P. Jonathan, "Inference for bivariate extremes via a semi-parametric angular-radial model," *Extremes*, vol. 28, pp. 209–238, 2024. DOI: 10.1007/s10687-024-00492-2.

[24] E. Mackay, C. Murphy-Barltrop, and P. Jonathan, "The SPAR model: A new paradigm for multivariate extremes. Application to joint distributions of metocean variables," *Journal of Offshore Mechanics and Arctic Engineering*, vol. 147, no. 1, p. 011 205, 2025. DOI: 10.1115/1.4065968.

[25] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001. DOI: 10.1198/tech.2002.s73.

[26] E. Mackay and A. F. Haselsteiner, "Marginal and total exceedance probabilities of environmental contours," *Marine Structures*, vol. 75, p. 102 863, 2021. DOI: 10.1016/j.marstruc.2020.102863.

[27] E. S. Simpson and J. A. Tawn, "Inference for new environmental contours using extreme value analysis," *Journal of Agricultural, Biological and Environmental Statistics*, pp. 1–25, 2024. DOI: 10.1007/s13253-024-00612-2.

[28] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, 2000.

[29] C. Ley and T. Verdebout, *Modern Directional Statistics*. Chapman and Hall/CRC, 2017. DOI: 10.1201/9781315119472.

[30] P. Hall, G. Watson, and J. Cabrera, "Kernel density estimation with spherical data," *Biometrika*, vol. 74, no. 4, pp. 751–762, 1987. DOI: 10.1093/biomet/74.4.751.

[31] Z. Bai, C. R. Rao, and L. Zhao, "Kernel estimators of density function of directional data," in *Multivariate Statistics and Probability*, Elsevier, 1989, pp. 24–39. DOI: 10.1016/B978-0-12-580205-5.50008-2.

[32] J. A. Mooney, P. J. Helms, and I. T. Jolliffe, "Fitting mixtures of von Mises distributions: A case study involving sudden infant death syndrome," *Computational Statistics & Data Analysis*, vol. 41, no. 3-4, pp. 505–513, 2003. DOI: 10.1016/S0167-9473(02)00181-0.

[33] Y. Fu, J. Chen, and P. Li, "Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions," *Journal of Statistical Planning and Inference*, vol. 138, no. 3, pp. 667–681, 2008. DOI: 10.1016/j.jspi.2007.01.003.

[34] K. Hornik and B. Grün, "movMF: An R package for fitting mixtures of von Mises-Fisher distributions.," *Journal of Statistical Software*, vol. 58, no. 10, pp. 1–31, 2014. DOI: 10.18637/jss.v058.i10.

[35] J. T. Ferreira, M. A. Juárez, and M. F. Steel, "Directional log-spline distributions," *Bayesian Analysis*, vol. 3, pp. 297–316, 2008. DOI: 10.1214/08-BA311.

[36] A. Pewsey and E. García-Portugués, "Recent advances in directional statistics," *TEST*, vol. 30, pp. 1–58, 1 Mar. 2021. DOI: 10.1007/s11749-021-00759-x.

[37] J. B. Wessel, C. J. Murphy-Barltrop, and E. S. Simpson, "A comparison of generative deep learning methods for multivariate angular simulation," *arXiv preprint arXiv:2504.21505*, 2025.

[38] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[39] R. A. Fisher, "Dispersion on a sphere," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 217, no. 1130, pp. 295–305, 1953. DOI: 10.1098/rspa.1953.0064.

[40] N. De Cao and W. Aziz, "The power spherical distribution," in *Proceedings of the 37th International Conference on Machine Learning INNF+ 2020 Workshop*, Vienna, Austria, 2020. DOI: 10.48550/arXiv.2006.04437.

[41] V. Chavez-Demoulin and A. C. Davison, "Generalized additive modelling of sample extremes," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 1, pp. 207–222, 2005. DOI: 10.1111/j.1467-9876.2005.00479.x.

[42] D. Randell, K. Turnbull, K. Ewans, and P. Jonathan, "Bayesian inference for nonstationary marginal extremes," *Environmetrics*, vol. 27, no. 7, pp. 439–450, 2016. DOI: 10.1002/ENV.2403.

[43] 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*, vol. 114, no. 528, pp. 1865–1879, 2019. DOI: 10.1080/01621459.2018.1529596.

[44] E. Zanini, E. Eastoe, M. Jones, D. Randell, and P. Jonathan, "Flexible covariate representations for extremes," *Environmetrics*, vol. 31, e2624, 2020. DOI: 10.1002/env.2624.

[45] A. M. Barlow, E. Mackay, E. Eastoe, and P. Jonathan, "A penalised piecewise-linear model for non-stationary extreme value analysis of peaks over threshold," *Ocean Engineering*, vol. 267, p. 113 265, 2023. DOI: 10.1016/j.oceaneng.2022.113265.

[46] E. S. Simpson and J. A. Tawn, "Estimating the limiting shape of bivariate scaled sample clouds: With additional benefits of self-consistent inference for existing extremal dependence properties," *Electronic Journal of Statistics*, vol. 18, no. 2, pp. 4582–4611, 2024. DOI: 10.1214/24-EJS2300.

[47] R. Majumder, B. A. Shaby, B. J. Reich, and D. Cooley, "Semiparametric estimation of the shape of the limiting multivariate point cloud," *Bayesian Analysis*, 2025, To appear. DOI: 10.1214/25-BA1514.

[48] J. Richards and R. Huser, "Extreme Quantile Regression with Deep Learning," in *Handbook on Statistics of Extremes*, M. de Carvalho, R. Huser, P. Naveau, and B. J. Reich, Eds., Chapman & Hall / CRC, 2024.

[49] C. J. Murphy-Barltrop, R. Majumder, and J. Richards, "Deep learning of multivariate extremes via a geometric representation," *arXiv preprint arXiv:2406.19936*, 2024.

[50] T. Wilson, P.-N. Tan, and L. Luo, "DeepGPD: A deep learning approach for modeling geospatio-temporal extreme events," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 4245–4253. DOI: 10.1609/aaai.v36i4.20344.

[51] J. Richards, R. Huser, E. Bevacqua, and J. Zscheischler, "Insights into the drivers and spatiotemporal trends of extreme Mediterranean wildfires with statistical deep learning," *Artificial Intelligence for the Earth Systems*, vol. 2, no. 4, e220095, 2023. DOI: 10.1175/AIES-D-22-0095.1.

[52] D. Cisneros, J. Richards, A. Dahal, L. Lombardo, and R. Huser, "Deep graphical regression for jointly moderate and extreme Australian wildfires," *Spatial Statistics*, vol. 59, p. 100 811, 2024. DOI: 10.1016/j.spasta.2024.100811.

[53] O. C. Pasche and S. Engelke, "Neural networks for extreme quantile regression with an application to forecasting of flood risk," *The Annals of Applied Statistics*, vol. 18, no. 4, pp. 2818–2839, 2024. DOI: 10.1214/24-AOAS1907.

[54] D. R. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *Journal of the Royal Statistical Society: Series B (Methodology)*, vol. 49, no. 1, pp. 1–18, 1987. DOI: 10.1111/j.2517-6161.1987.tb01422.x.

[55] S. Tendijck, D. Randell, G. Feld, and P. Jonathan, "Practical non-stationary extreme value analysis of peaks over threshold using the generalised Pareto distribution: Estimating uncertainties in return values," *Ocean Engineering*, vol. 312, p. 119 247, 2024. DOI: 10.1016/j.oceaneng.2024.119247.

[56] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[57] R. Koenker, V. Chernozhukov, X. He, and L. Peng, *Handbook of Quantile Regression*. CRC Press, 2017.

[58] R. Neuneier, F. Hergert, W. Finnoff, and D. Ormoneit, "Estimation of conditional densities: A comparison of neural network approaches," in *Proceedings of the International Conference on Artificial Neural Networks, 1994*, 1994, pp. 689–692. DOI: 10.1007/978-1-4471-2097-1_162.

[59] J. Rothfuss, F. Ferreira, S. Walther, and M. Ulrich, "Conditional density estimation with neural networks: Best practices and benchmarks," *arXiv preprint arXiv:1903.00954*, 2019.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] J. Richards and R. Huser, "Regression modelling of spatiotemporal extreme US wildfires via partially-interpretable neural networks," *arXiv preprint arXiv:2208.07581*, 2022.

[62] G. Hansen, I. Herburt, H. Martini, and M. Moszyńska, "Starshaped sets," *Aequationes mathematicae*, vol. 94, pp. 1001–1092, 2020. DOI: 10.1007/s00010-020-00720-7.

[63] J. R. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized Pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987. DOI: 10.2307/1269343.

[64] R. Majumder and J. Richards, "Semi-parametric bulk and tail regression using spline-based neural networks," *arXiv preprint arXiv:2504.19994*, 2025. DOI: 10.48550/arXiv.2504.19994.

[65] E. B. Mackay and P. Jonathan, "Estimation of environmental contours using a block resampling method," in *International Conference on Offshore Mechanics and Arctic Engineering*, 2020. DOI: 10.1115/OMAE2020-18308.

[66] S. Tendijck, P. Jonathan, D. Randell, and J. Tawn, "Temporal evolution of the extreme excursions of multivariate k th order Markov processes with application to oceanographic data," *Environmetrics*, vol. 35, e2834, 3 2023. DOI: 10.1002/env.2834.

[67] G. S. Watson, *Statistics on spheres*. Wiley, 1983.

[68] G. Ulrich, "Computer generation of distributions on the m-sphere," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 33, no. 2, pp. 158–163, 1984. DOI: 10.2307/2347441.

[69] A. S. Householder, "Unitary triangularization of a nonsymmetric matrix," *Journal of the Association for Computing Machinery (JACM)*, vol. 5, no. 4, pp. 339–342, 1958. DOI: A10.1145/320941.320947.

[70] M. E. Muller, "A note on a method for generating points uniformly on n-dimensional spheres," *Communications of the ACM*, vol. 2, no. 4, pp. 19–20, 1959. DOI: 10.1145/377939.377946.