Computational Statistics and Data Analysis 95 (2016) 243–258

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fast computation of large scale marginal extremes with multi-dimensional covariates

Laks Raghupathi [a,*], David Randell [b], Kevin Ewans [c], Philip Jonathan [b]

[a] Shell India Markets Pvt. Ltd., Bangalore, 560048, India
[b] Shell Global Solutions (UK), Manchester M22 0RR, United Kingdom
[c] Sarawak Shell Bhd., 50450 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 7 September 2015
Accepted 12 September 2015
Available online 8 October 2015

Keywords:
Extremes
Splines
Non-stationary effects
Fast-computation
High-performance computation

ABSTRACT

Safe and reliable design and operation of fixed and floating marine structures often located in remote and hostile environments is challenging. Rigorous extreme value analysis of meteorological and oceanographic data can greatly aid the design of such structures. Extreme value analysis is typically undertaken for single spatial locations or for small neighbourhoods; moreover, non-stationary effects of covariates on extreme values are typically accommodated in an ad-hoc manner. The objective of the work summarised here is to improve design practice by estimating environmental design conditions (such as return values for extreme waves, winds and currents) for a whole ocean basin, including additional covariate effects (such as storm direction) as necessary, in a consistent manner. Whole-basin non-stationary extreme value modelling is computationally complex, requiring inter-alia the estimation of tail functions, the parameters of which vary with respect to multi-dimensional covariates characterised by us using tensor products of penalised B-splines. We outline two technical contributions which make whole-basin non-stationary analysis feasible. Firstly, we adopt generalised linear array methods to reduce the computational burden of matrix manipulations. Secondly, using high-performance computing, we develop a parallel implementation of maximum likelihood estimation for the generalised Pareto distribution. Together, these innovations allow estimation of rigorous whole-basin extreme value models in reasonable time. We evaluate the new approach in application to marginal extreme value modelling of storm peak significant wave heights in two ocean basins, accommodating spatial and directional covariate effects.

1. Introduction

1.1. Background

Safe exploration and production of oil and gas from offshore locations requires the design and construction of marine structures able to withstand severe environmental conditions, more extreme than the worst foreseeable during the required structural lifetime, associated with an annual probability of failure typically less than 1 in 10,000. Robust modelling of extreme environmental conditions is therefore critical for safe day-to-day operation and long-term structural reliability. This requires the specification of return values for oceanographic phenomena such as waves, winds, currents and associated

* Corresponding author.
E-mail address: l.raghupathi@shell.com (L. Raghupathi).

http://dx.doi.org/10.1016/j.csda.2015.09.006
0167-9473/

variables. The metocean engineer faces the challenge of analysing huge samples of measured or hindcast meteorological and oceanographic data exhibiting complex dependences, including non-stationarity with respect to multiple covariates and spatial dependence of extremes, in order to estimate both marginal and joint return values for a large number of wave-, wind- and current-related variables.

Extreme value analysis is motivated by asymptotic arguments. In the metocean context, only the largest values in the sample (for example, observations exceeding some threshold) should be used for extreme value modelling, but the sample size must also be sufficient for good empirical modelling. Therefore, the amount of data available for analysis at a specific location is often limited. Reducing the threshold value runs the risk of invalidating asymptotic arguments underlying the modelling strategy. Increasing the threshold value reduces sample size still further. One pragmatic solution to the sample size issue is to aggregate observations from a neighbourhood of spatial locations and analyse the pooled sample. However, neighbouring locations in general have different extreme value characteristic, and observations from neighbouring locations are typically not statistically independent. Naive analysis of spatial extremes over spatial neighbourhoods can lead therefore to erroneous inferences. Fortunately, recent developments in statistical modelling of extreme ocean environments offer solutions other than pooling; a recent review is presented in Jonathan and Ewans (2013). Jonathan et al. (2013, 2014) present recent progress in the spline-based modelling of covariates effect. Here, we further develop these methods for basin-wide marginal modelling of extremes, incorporating non-stationarity due to direction, longitude and latitude, and adjusting for spatial dependence between locations in all uncertainty analyses. An outline of the modelling procedure is given at the start of Section 2.

At present, preliminary design conditions for a specific location are typically estimated by the metocean engineer based on data for (the neighbourhood of) that location. This process is typical time-consuming and technically challenging. Moreover, industry design procedures are sufficiently vague that there could be inconsistency between estimated design conditions for the same location made by different metocean engineers, using different data sources, modelling assumptions and procedures. With a basin-wide extreme value model in place, spatio-directional return values for all locations in the ocean basin of interest can be estimated. Outline directional design conditions for locations of interest in the basin can be inferred based on rigorous, consistent statistical analysis, allowing fair comparison of environmental risk for those locations. The estimates might further be pre-computed, stored in a database and accessed via a graphical user interface to a device of choice querying the database. In practice, it may further be necessary to calibrate hindcast data to measurements at specific locations.

The paper is organised as follows. Section 2 presents the model formulation for extreme values subject to spatio-directional non-stationarity, outlining computational challenges to its implementation basin-wide. Section 3 describes generalised linear array methods and reports their performance by incorporation within the extreme value model. Section 4 describes parallel adaption of the generalised Pareto algorithm and quantifies computational improvements achieved. Section 5 describes the practical application of the above methodology to estimation of basin-wide return values for ocean storm severity in two ocean basins. Discussion, conclusions and opportunities for further development are summarised in Section 6.

2. A spatio-directional marginal extreme value model

The objective is the estimation of *marginal* return values for storm severity (quantified using significant wave height) for locations within a spatial neighbourhood, accounting for spatial and storm directional variability of extremal characteristics.

2.1. Model Components

Following the work of Jonathan and Ewans (2008, 2011), summarised in Jonathan et al. (2014), we model storm peak significant wave height, namely the largest value of significant wave height observed at each location during the period of a storm event. At a given location, storm peak events are reasonably assumed to be statistically independent given covariates since they correspond to occurrences of independent atmospheric pressure fields. We assume that each storm event is observed at all locations within the neighbourhood under consideration. For a sample $\{\dot{z}_i\}_{i=1}^{\dot{n}}$ of $\dot{n}$ storm peak significant wave heights (henceforth H_S) observed at locations $\{\dot{x}_i, \dot{y}_i\}_{i=1}^{\dot{n}}$ with dominant wave directions $\{\dot{\theta}_i\}_{i=1}^{\dot{n}}$ at storm peak H_S (henceforth "storm directions"), we proceed using the peaks-over-threshold approach as follows.

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