Contents lists available at ScienceDirect

Coastal Engineering

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

Uncertainty quantification in estimation of extreme environments

Matthew Jones[a], Hans Fabricius Hansen[b], Allan Rod Zeeberg[c], David Randell[a], Philip Jonathan[d,e,*]

[a] Shell Global Solutions International B.V., 1031, HW Amsterdam, The Netherlands
[b] Danish Hydraulics Institute, Agern Alle 5, 2970, Horsholm, Denmark
[c] TOTAL E&P Danmark A/S, Britanniavej 10, 6700 Esbjerg, Denmark
[d] Shell Research Ltd., London, SE1 7NA, United Kingdom
[e] Department of Mathematics and Statistics, Lancaster University, LA1 4YW, United Kingdom

ARTICLE INFO

Keywords:
Bayesian uncertainty analysis
Emulation
Discrepancy
Extreme
Significant wave height
Non-stationary

ABSTRACT

We estimate uncertainties in ocean engineering design values due to imperfect knowledge of the ocean environment from physical models and observations, using Bayesian uncertainty analysis. Statistical emulators provide computationally efficient approximations to physical wind–wave environment (i.e. "hindcast") simulators and characterise simulator uncertainty. Discrepancy models describe differences between hindcast simulator outputs and the true wave environment, where the only measurements available are subject to measurement error. System models (consisting of emulator–discrepancy model combinations) are used to estimate storm peak significant wave height (henceforth H_S), spectral peak period and storm length jointly in the Danish sector of the North Sea. Using non-stationary extreme value analysis of system output H_S, we estimate its 100-year maximum distribution from two different system models, the first based on 37 years of wind–wave simulation, the second on 1200 years; estimates of distributions of 100-year maxima are found to be in good general agreement, but the influence of different sources of uncertainty is nevertheless clear. We also estimate the distribution of 100-year maximum H_S using non-stationary extreme value analysis of storm peak *wind speed*, propagating simulated extreme winds through a system model for H_S; we find estimates to be in reasonable agreement with those based on extreme value analysis of H_S itself.

1. Introduction

Estimation of characteristics of extreme ocean environmental variables is critical in marine and coastal structural design. This typically requires extreme value analysis of historical data from measurements and hindcasts, characterising the environment over some period of time, typically of the order of 30–100 years. The use of extreme value analysis is motivated by asymptotic arguments concerning the forms of tails of probability distributions (e.g. Beirlant et al., 2004). Inference involves estimating the maximum value that might be observed in a time period considerably longer than that of the historical sample, typically of the order of 1000 or 10000 years (or analogous extreme quantiles of the distribution of the annual maximum). Estimation is complicated by numerous sources of systematic and random variation, including temporal (e.g. Chavez-Demoulin and Davison, 2012) and spatial (e.g. Davison et al., 2012) dependence of both typical and extreme values, non-stationarity with respect to multiple covariates (e.g. Mendez et al., 2006; Mendez et al., 2008; Mackay et al., 2010; Vanem, 2015), and measurement scale or convergence uncertainty (Guedes-Soares and Scotto, 2001; Wadsworth et al., 2010; Reeve et al., 2012; Papastathopoulos and Tawn, 2013). There are also sources of procedural uncertainty, such as threshold selection in peaks over threshold analysis (e.g. Scarrott and MacDonald, 2012; Northrop et al., 2017), and block length for block maxima analysis. Bayesian inference provides a natural framework for bias and uncertainty quantification in extreme value analysis; recent articles (e.g. Cooley et al., 2006; Tancredi et al., 2006; Sanchez-Archilla et al., 2008; MacDonald et al., 2011; Reich and Shaby, 2012; Randell et al., 2016) illustrate the propagation of uncertainty from threshold choice to distributions of extreme values such as the *N*-year maximum.

There is a large literature investigating different sources of uncertainties related to estimation of extreme values; Wada and Waseda (2018) provides a recent discussion. These sources can be classified as natural (inherent or aleatory) or sampling (model or epistemic) uncertainties. Aleatory uncertainty refers to the fundamental natural randomness of the phenomenon being considered and cannot be

[*] Corresponding author. Shell Research Ltd., London, SE1 7NA, United Kingdom.
E-mail address: philip.jonathan@shell.com (P. Jonathan).

https://doi.org/10.1016/j.coastaleng.2018.07.002
Received 21 February 2018; Received in revised form 8 June 2018; Accepted 8 July 2018
Available online 12 July 2018

avoided or reduced by more measurements; an example is the variability of wave height within a sea state. Epistemic uncertainty refers to lack of information due to limitations in the size and quality of data, inadequate models and so on. Epistemic uncertainty can be reduced by increasing the sample size, improving the accuracy of the measurements, and improving our models. Coles and Simiu (2003) suggests that quantifying uncertainty is an essential component, arguably the most important component, of any extreme value analysis.

Model and measurement inadequacy is considered by some authors. Coles and Simiu (2003) and Brooker et al. (2004) consider the effect of hindcast uncertainty on estimation of extreme value distributions from hindcast data. Jonathan and Ewans (2007) considers the effect of measurement error in extreme value estimation. Perhaps because current practice in extreme value analysis is already problematic, relatively little effort has been devoted to quantifying fully the effects of uncertainties in the underlying historical sample on extreme value inferences. Limited sample size is noted by some authors as a key source of epistemic uncertainty. For example, Forristall et al. (1996) suggests that estimates for storm peak significant wave heights made from measurements with finite record lengths can exhibit positive biases due to epistemic sample variability, and that consequently unbiased hindcast-based estimates might be judged to be biased low from comparison with measurements. Orimolade et al. (2016) considers estimation of extreme significant wave heights and associated uncertainties for data from the NORA10 hindcast for the Barents Sea. They explicitly seek to quantify aleatory uncertainty, and epistemic uncertainty due to sample size using bootstrap re-sampling, for estimates based on analysis of peaks over threshold and annual maxima. Bootstrap re-sampling (e.g. Davison and Hinkley, 1997) is probably the most popular and useful technique for quantification of epistemic uncertainty when Bayesian inference is not preferred. When Bayesian inference is used, however, it automatically provides a balanced, comprehensive and transparent framework for specification and estimation of uncertainty.

The fundamental motivation for this work is the acknowledgement that safety- and economically-critical decisions regarding the design and reassessment of marine structures are vunerable to multiple sources of uncertainty. Any rational attempt to make these decisions well must accommodate effects of uncertainties from all components of the decision process in a mathematically coherent fashion. Historically, the mathematical and physical modelling, and computational tools to achieve optimal decisions were not available; for this reason, the basis for design was a combination of observations, simple statistical and physical modelling approaches, safety factors and good engineering judgement. Specifically, uncertainties were not and could not be accommodated systematically and coherently. Now, however, the situation is changing rapidly: workable statistical and computational methodologies exist so that thorough, coherent uncertainty analysis can be performed in marine design. The most prominent methodology in the statistics literature to achieve this is Bayesian uncertainty analysis (see e.g. Berger, 1985). This article seeks to show how Bayesian uncertainty analysis can be applied to practical marine design situation of considerable current interest. Bayesian uncertainty analysis offers the practising marine engineer greater confidence in the process of estimation of marine risk. Estimates for the characteristics of extreme environments from a Bayesian uncertainty analysis may happen to be similar to those obtained using more conventional approaches. This is good news, providing the engineer with confidence that conventional approaches are appropriate in particular cases. However, in general, Bayesian uncertainty analysis provides a better approach to structuring and quantifying uncertainty and hence risk, and should therefore be the *preferred framework for estimating uncertainty well.*

In ocean engineering applications, the data for analysis typically corresponds to output from a hindcast simulator for the ocean environment at the location of interest over some historical time period. The hindcast simulator is a physical model for the environment, calibrated by some procedure to observations from that environment. (The term "simulator" is used in the statistical literature on uncertainty quantification to refer to a numerical model of a physical system; we will use it here also for clarity.) We do not know in general (a) to what extent the hindcast simulator adequately represents the physics of the environment, and its extremes in particular, (b) the bias and uncertainty in offshore measurements used for simulator calibration, and (c) whether calibration of the hindcast simulator to these observations is made reasonably, for subsequent extreme value inferences in particular.

The need to incorporate uncertainties due to underlying simulator assumptions and imperfect data for simulator calibration is well understood in many fields (e.g. Vernon et al., 2010 for galaxy formation, Oyebamiji et al., 2017 for evolution of microbial communities). It would seem rational and desirable to consider effects of such sources of uncertainty on estimates for extreme quantiles of distributions used in ocean engineering also. This paper seeks to achieve this in application to extreme quantile inference in the Danish sector of the North Sea. Specifically, we seek to construct a joint model for the significant wave height and spectral peak period for the sea state corresponding to the peak of a storm, and the number of waves in a storm. This full "system model" can be thought of as consisting of two parts, as described below.

The first part of the system model is a statistical model known as a "hindcast emulator" (or simply "emulator") for the hindcast simulator, with which to predict hindcast simulator outputs for any combination of hindcast inputs. Hindcast inputs include (a) physical covariates (such as wind field variables, geographic location and water depth), (b) hindcast simulator tuning parameters (which quantify effects which cannot be represented adequately in the physical model, such as the extent of bottom friction), and (c) hindcast set-up parameters (such as algorithmic parameters required to run the hindcast simulator, for instance the choice of hindcast spatial and temporal grid resolutions). The purpose of the emulator is to provide computationally rapid estimates for hindcast outputs given specified inputs. Emulators are typically estimated using regression; here we use Gaussian process regression and Bayes Linear analysis as described in Section 3. Emulators are already used in a coastal engineering context e.g. to provide computationally-efficient approximations for near-shore wave transformation (Malde et al., 2018).

The second part of the system model is a statistical model known as a "discrepancy model" which predicts the difference between hindcast simulator outputs and the true wave environment, as a function of physical covariates (such as those described under (a) in the previous paragraph). Discrepancy models are estimated in a similar manner to emulators.

The full system model provides a means for rapid joint estimation of system outputs, reflecting uncertainty regarding imperfect knowledge of the physical environment and its description using the hindcast simulator.

Outputs of system models can be used as inputs to extreme value models, motivating a variety of different estimators for the distribution of 100-year maximum H_S. In this work we consider estimates based on extreme value analysis of measured and hindcast H_S as base cases. But given the importance of good wind field characterisation in wave hindcasting (Cardone et al., 1995; Cardone and Cox, 2009), we also estimate the distribution of 100-year H_S, first using non-stationary extreme value analysis of storm peak *wind speed* (henceforth u), then propagating simulated extreme winds through a system model for H_S.

The article is structured as follows. In Section 2 we describe the application motivating this work, namely estimation of N-year maxima for storm peak significant wave height (H_S) and related variables at locations in the Danish sector of the North Sea offshore Jutland. In Section 3 we introduce the system model and describe in outline how the hindcast emulator and discrepancy model are estimated using Bayesian inference; mathematical details of Bayes linear inference are relegated to Appendices A and B. Section 4 then provides details of system model estimation for two related sources of data. The first data set (referred to as “Case A” for clarity) is comprised of wind field, wave hindcast and wave measurements corresponding to a period of 37 years in the recent past. The second data set (referred to as “Case B”) consists of extreme wind fields from a global high-resolution climate model (Shaffrey et al., 2009), for a period of approximately 1200 years, and corresponding wave hindcast and all relevant recent wave measurements. In Section 5 we outline how non-stationary extreme value analysis is used to estimate the tail of the distribution of H_S and u, given appropriate covariates. In Section 6, we combine outputs of system models for Case A or B with extreme value models for H_S or u to make different estimates for the distribution of 100-year maximum H_S. Section 7 provides discussion and conclusions.

2. Motivating application

This work is motivated by the need to estimate distributions of N-year maxima for the ocean environment at a neighbourhood of locations approximately 220 km offshore the west coast of Jutland, Denmark in the North Sea, at a water depth of approximately 40 m. There is particular interest in estimating the effect of different sources of uncertainty (from wave hindcast models and offshore measurements) on estimates of distributions for N-year maxima.

We consider two samples of data. The first (Case A) is based on wind fields and corresponding wave hindcast simulator outputs, and wave measurements for a period of 37 years from 10th January 1979 to 30th December 2015 for the neighbourhood. CFSR wind fields (Saha et al., 2014) are input to a MIKE21 spectral wave simulator model (Sorensen et al., 2005) for a number of different combinations of hindcast tuning and set-up parameters, specified using a Latin hypercube design, to generate multiple sets of wave hindcast outputs. These are then filtered to isolate storm peak wind and wave characteristics for storms in the period using the procedure similar to that outlined in Ewans and Jonathan (2008). Storm events are identified as exceedances of a threshold which is non-stationary with respect to season and direction, and therefore may not necessarily correspond to “storms” as defined from a meteorological perspective. A total of 2187 storm events is isolated. Partial measurements of the wave environment for these storm periods at 7 offshore locations in the neighbourhood are also available.

For illustration, Fig. 1 shows the Case A sample for a specific (and reasonable) choice of hindcast tuning and set-up parameters at a specific central location (termed “C” for convenience). Storms are characterised in terms of hindcast storm peak significant wave height (H_S) and wave direction (ϕ), spectral peak period (T_P) at the storm peak, storm peak wind speed (u) and wind direction (θ) and the number of individual waves σ in a storm. Some measured storm peak significant wave heights (H_S measured) are also available. Note that wind and wave direction are defined as the direction *from which* events emanate, measured clockwise from north in degrees. Variation of hindcast H_S and u with direction typical for the region is evident, as is the strong relationship between hindcast H_S and u. Hindcast H_S and measured H_S are generally in good agreement, and there is no appreciable bias between them in particular. The typical H_S-T_P relationship for wind waves is also observed in the hindcast. Hindcast storm length ($\log_{10}(\sigma)$) reduces with increasing H_S, but this relationship is relatively weak. Note that, in addition to the sample illustrated in Fig. 1, Case A hindcast simulator outputs are available for a subset of storms and locations for each of 128 different combinations of hindcast tuning and set-up parameters obtained from a designed computer experiment. As described in Section 4, these data are critical for estimation of the hindcast emulator for Case A.

The second data source (Case B) is based on wind fields (from a climate model) corresponding to the most extreme waves over a period of 1200 years (as described in Section 1 and Shaffrey et al. (2009)) for the same spatial neighbourhood. The corresponding sample of wind speed, direction and wave hindcast variables (obtained using the same hindcast tuning and set-up parameters as for Case A in Fig. 1) are illustrated in Fig. 2. The figure is generally similar to Fig. 1. However, note that the hindcast $H_S - T_P$ relationship, for example, is influenced by the selection process of severe events, and is different to that typically observed in data corresponding to a continuous period of observation or hindcast. It appears that isolation of the most extreme wind fields, identifies some storm events with very long peak periods, and other (typically less intense) storms exhibiting a linear relationship between H_S and T_P. The largest value of hincast H_S for Case B is approximately 11.5 m, corresponding to a storm from the north.

Samples for Cases A and B are used in Section 4 to estimate system models, in Section 5 for extreme value analysis, and in Section 6 to estimate distributions for the 100-year maximum H_S.

3. Bayesian uncertainty analysis

Here we describe in general terms how an emulator for a hindcast simulator and a discrepancy model for the difference between emulator outputs and measurements are specified (Section 3.1) and estimated (Section 3.2). The full procedure is referred to as Bayesian uncertainty analysis. A detailed description of the inference for Cases A (in detail)

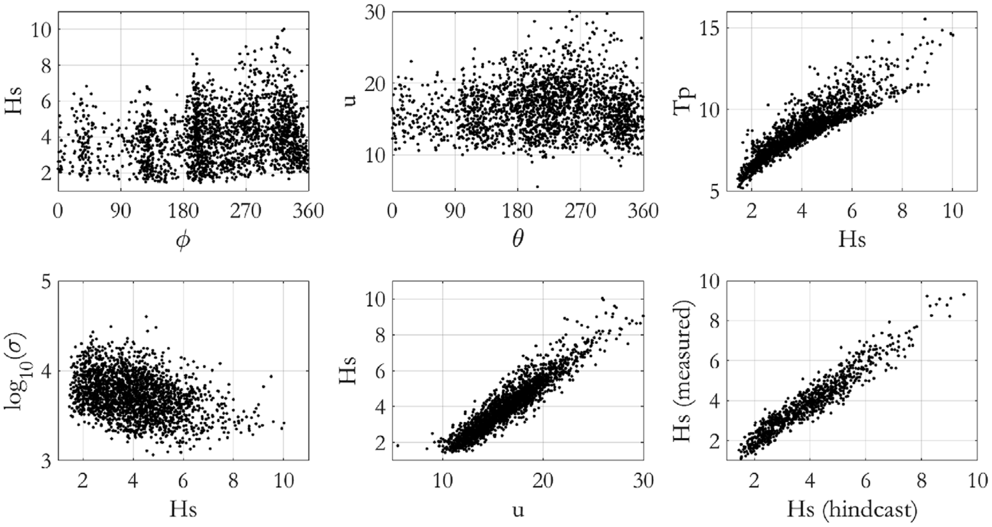

Fig. 1. Illustration of the Case A data at central location C. Hindcast outputs are: storm peak significant wave height (H_S), wave direction (ϕ), spectral peak period (T_P), wind speed (u) and wind direction (θ) and the number of waves σ in a storm. Measured variables are: storm peak significant wave height (H_S measured).

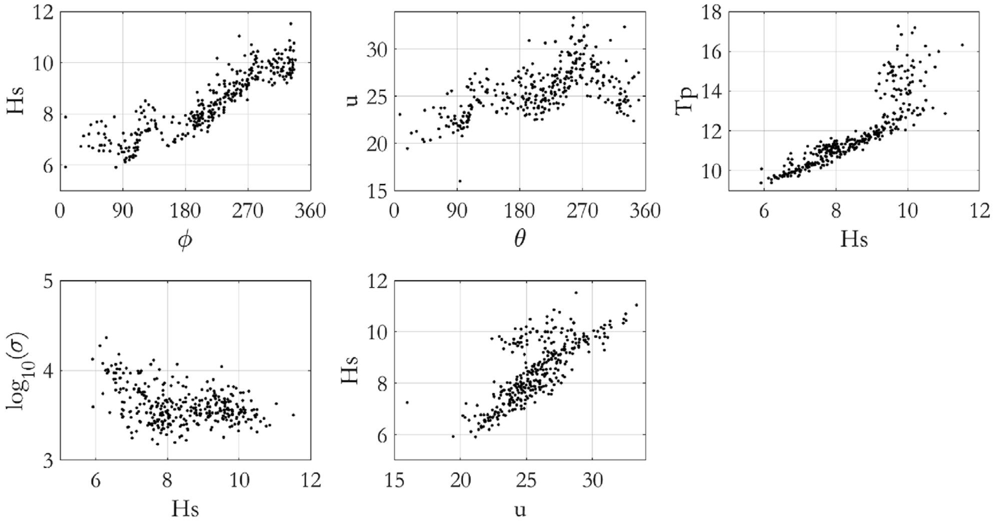

Fig. 2. Illustration of the Case B data at central location C. Hindcast outputs are: storm peak significant wave height (H_S), wave direction (ϕ), spectral peak period (T_P), wind speed (u) and wind direction (θ) and the number of sea states σ in a storm. There are no measurements available.

and B (in outline) is then given in Section 4.

In this work, models are estimated using Bayes Linear inference rather than fully probabilistic Bayesian inference. For completeness, a brief introduction to Bayes Linear analysis is provided in Appendix A; further mathematical details for the Bayes Linear analysis are also provided there. Since some readers may be more familiar with the corresponding fully probabilistic Bayesian inference, we also outline this approach in Appendix B for comparison.

We specify the system model using the following symbols: f (and F) refers to the output of a hindcast simulator, a deterministic function taking physical covariates x and hindcast tuning parameters ω as inputs. δ (and D) refer to the output of a discrepancy model, a deterministic function taking physical covariates x as inputs. System output y is then the sum of simulator and discrepancy model outputs, which we seek to relate to measurements z (and Z) of the system. For the current work, y refers to storm peak H_S, storm peak T_P and storm length σ, and x to physical covariates including location, storm peak wind speed u and storm peak wind direction θ. Further details are given in Section 4.

3.1. Model specification

3.1.1. Relating hindcast simulator, discrepancy and system

The set of system outputs under study is denoted by $y(x) = \{y_1(x), ..., y_{n_y}(x)\}$, where x is a vector of known physical covariates affecting the behaviour of y. Suppose that a hindcast simulator $f(x, \omega) = \{f_1(x, \omega), ..., f_{n_y}(x, \omega)\}$ has been constructed to model the behaviour of y, where $\omega = \{\omega_1, ..., \omega_{n_\omega}\}$ is a set of hindcast simulator tuning parameters that must be selected to run the hindcast simulator (but which do not correspond to physical covariates). The hindcast simulator f is linked to the system by means of a “best input” assumption: we assume that there is a setting ω^* of tuning inputs ω, such that if the simulator was to be run at this setting, we would obtain all of the information available from the simulator about the system. Under this assumption, the hindcast simulator is related to the system as

$$y_i(x) = f_i^*(x) + \delta_i(x) \quad , \tag{1}$$

where $f_i^*(x) = f_i(x, \omega^*)$ is the simulator evaluated at the best input, and $\delta_i(x)$ is the discrepancy between the simulator and the system at this setting of the tuning parameters. δ is assumed to be uncorrelated with $\{f, \omega^*\}$.

3.1.2. Modelling discrepancy

We choose to specify the discrepancy as the sum of a regression component representing global behaviour with respect to covariates, and a residual component representing local deviations from this behaviour

$$\delta_i(x) = \sum_p \alpha_{ip} h_p(x) + w_i(x) + \xi_i \quad . \tag{2}$$

Here $h(x) = \{h_1(x), ..., h_{n_h}(x)\}$ is a set of pre-specified basis functions, and α is a corresponding set of unknown regression weights to be estimated. $w(x) = \{w_1(x), ..., w_{n_y}(x)\}$ is a set of unknown residual functions, assumed to be correlated across the covariate domain, to be estimated. $\xi = \{\xi_1, ..., \xi_{n_y}\}$ is a set of unknown unstructured (“nugget”) residuals to be estimated, assumed uncorrelated across input space, used to capture unstructured variation in the discrepancy.

3.1.3. Relating system and measurements

Sets of system measurements are denoted $z = \{z_1, ..., z_{n_z}\}$, where $z_k = \{z_{1k}, ..., z_{n_y k}\}$. We assume that z_{ik} is a noise-corrupted measurement of the underlying system value $y_i(x_k)$ made at known system input setting x_k

$$z_{ik} = y_i(x_k) + \varepsilon_{ik} \quad , \tag{3}$$

where ε is a corresponding set of noise terms, assumed uncorrelated with each other, and with system value.

3.1.4. Emulating the simulator

Simulators for physical systems often consist of numerical solvers for sets of coupled differential equations on spatio-temporal grids. Such numerical solvers are generally computationally demanding. A computationally-slick approximation for the simulator is therefore practically very appealing. A common strategy is to approximate f using a statistical model known as an emulator. An emulator can be used to generate fast predictions for f at input settings where it has not yet been evaluated, allowing more detailed investigation of simulator behaviour. An emulator is typically specified as the sum of a regression surface, a structured residual component, and an unstructured residual component

$$f_i(x, \omega) = \sum_p \beta_{ip} g_p(x, \omega) + r_i(x, \omega) + \eta_i \quad . \tag{4}$$

Here, $g(x, \omega) = \{g_1(x, \omega), ..., g_{n_g}(x, \omega)\}$ is again a set of pre-specified basis functions, and β is a corresponding set of regression weights to be estimated. The regression component of the model captures the global structure of the simulator. $r(x, \omega) = \{r_1(x, \omega), ..., r_{n_y}(x, \omega)\}$ is a set of unknown, zero-mean residual processes, assumed correlated across input space, to be estimated. These capture local deviations from the global structure. $\eta = \{\eta_1, ..., \eta_{n_y}\}$ is a set of nugget residuals to be estimated, assumed uncorrelated across input space, used to capture unstructured variation in simulator output.

3.2. Bayes linear analysis

Using the model specified above, Bayesian uncertainty analysis is performed within a Bayes Linear (second-order) framework by the procedure described here. A Bayes Linear analysis proceeds in two steps. First, a second-order prior belief specification is made for all uncertain quantities, typically using data available initially. Then, beliefs are updated using further data.

3.2.1. Estimating the hindcast emulator

Referring to Equation (4), we first specify basis functions $\{g_p(x, \omega)\}$. Typically in the absence of other information, these correspond to linear, squared and interaction terms in the elements of x and ω; where possible, they should be chosen to reflect the underlying physical relationship between hindcast simulator inputs and outputs if known.

Next we estimate prior moments for all emulator components from Equation (4). Specifically, we estimate expectations $\{\mathrm{E}[\beta_{ip}]\}$ for regression weights, and covariances $\{\mathrm{Cov}[\beta_{ip}, \beta_{jq}]\}$ and $\{\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')]\}$ between pairs of regression weights and residual processes. We note that the set $\{\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')]\}$ is specified in terms of covariances $\{\mathrm{Cov}[r_i(x, \omega), r_j(x, \omega)]\}$ between residuals for outputs evaluated at the same input setting (x, ω), and correlations $\{\mathrm{Corr}[r_i(x, \omega), r_j(x', \omega')]\}$ between different input settings (x, ω) and (x', ω') (with $x' \neq x, \omega' \neq \omega$). We also typically set all covariances between regression weight, residual process and nugget $\{\eta_i\}$ components to zero. The procedure for prior specification is described in detail in Appendix A.1.

Then, we adjust emulator moments using Bayes Linear adjustment for a further set of hindcast simulator evaluations. The adjusted moments are then used to compute adjusted predictions $\{\mathrm{E}_F[f_i(x, \omega)]\}$ and $\{\mathrm{Cov}_F[f_i(x, \omega), f_j(x', \omega')]\}$ at input settings where the simulator value has not been observed. The procedure for computing adjusted simulator predictions is described in detail in Appendix A.1. Finally, as also explained in Appendix A.1, we set nugget covariances $\{\mathrm{Cov}[\eta_i, \eta_j]\}$ manually.

Typically in a Bayesian uncertainty analysis, we assume that there is an unknown setting ω^* (called the "best input setting") of tuning inputs ω, such that if the simulator was to be run at this setting, we would obtain all of the information available from the simulator about the system. As will be discussed in Section 4, for the current work, we find that hindcast emulator performance (an approximation for hindcast simulator performance) varies within the domain of ω specified prior to analysis, but that this variability is not large. That is, the whole domain of ω corresponds effectively to good simulator performance. For this reason, we assume that ω^* is uniformly distributed on the full domain of ω. Incorporating uncertainty about the choice of ω^* in the best input emulator is discussed in the next section.

In passing we note that the procedure described here and in Appendix A is also used to make a prior specification for the discrepancy model, required for Bayes Linear adjustment of the system model.

3.2.2. Propagating uncertainty about tuning inputs

The fitted emulator is used to compute beliefs about the best input emulator f^*. Specifically, we compute $\{\mathrm{E}[f_i^*(x)]\}$ and $\{\mathrm{Cov}[f_i^*(x), f_j^*(x')]\}$ by propagating uncertainty about ω^* through the fitted emulator. Details of this calculation are provided in Appendix A.2.

3.2.3. Estimating the system model

Referring to Equation (1) and Equation (2), beliefs about the system $y(x)$ are computed by updating $f^*(x)$ and $\delta(x)$ jointly. The procedure is similar to that for fitting the emulator. First, basis functions $\{h_p(x)\}$ are selected, then prior beliefs $\{\mathrm{E}[\alpha_{ip}]\}$, $\{\mathrm{Cov}[\alpha_{ip}, \alpha_{jq}]\}$ and $\{\mathrm{Cov}[w_i(x), w_j(x')]\}$ specified. Additionally, a second-order prior uncertainty specification is made for the measurement error components, consisting of expectations $\{\mathrm{E}[\varepsilon_{ik}]\}$ and covariances $\{\mathrm{Cov}[\varepsilon_{ik}, \varepsilon_{jl}]\}$. The system measurements $\{z_{ik}\}$ are then used to jointly adjust beliefs about $f^*(x)$ and $\delta(x)$; these updated beliefs can then be used to compute adjusted predictions $\{\mathrm{E}_z[y_i(x)]\}$ and $\{\mathrm{Cov}_z[y_i(x), y_j(x')]\}$ for the system at unobserved input settings. Finally, we use an independent test set to tune the nugget covariances $\{\mathrm{Cov}[\xi_i, \xi_j]\}$ manually. The procedure for computing adjusted system predictions is described in detail in Appendix A.3.

4. Model estimation

In this section we discuss the estimation of system models for Case A and Case B. Since model estimation for Case B follows the same steps as for Case A, we choose to describe Case A in detail (in Section 4.1–4.3) and overview Case B briefly (in Section 4.4).

Data for Case A are introduced in Section 2. This consists of hindcast simulation inputs and outputs, and offshore measurements, at multiple locations in the Danish Sector of the North Sea offshore Jutland for a recent period of 37 years. Our objective is to estimate a useful system model for the present day wave environment there. We note in particular that we have access to multiple sets of hindcast simulator outputs, corresponding to different choices of hindcast tuning inputs ω.

Modelling proceeds as follows. Step 1.1 of the analysis is described in Section 4.1.1. We fit an emulator to a set of hindcast simulator outputs corresponding to a limited number of offshore locations but a large number of combinations of tuning inputs ω. The fitted emulator is used to explore the effect of tuning input setting on simulator outputs, to be used subsequently to evaluate the hindcast simulator for a much larger number of locations. In Step 1.2 of the analysis (Section 4.1.2), we fit a second emulator to a larger set of data for hindcast simulator inputs and outputs (for all geographic locations, with a specific choice of tuning inputs $\omega^\dagger$). We use this emulator to explore physical covariate effects. We next combine the emulators, adding terms describing tuning input effects taken from the Step 1.1 emulator to the Step 1.2 emulator, so that the final emulator describes the effects of both physical covariate and simulator tuning inputs. In Step 2 of the analysis (Section 4.2), we discuss how uncertainty about tuning inputs is transferred into uncertainty about system outputs. In Step 3 (Section 4.3), we jointly update beliefs about the emulator and a discrepancy model to create a model for system behaviour, using data for hindcast emulator inputs, outputs and corresponding offshore measurements at a limited number of locations.

4.1. Step 1: emulation

4.1.1. Step 1.1: first emulator for limited number of locations but multiple ω

The first emulator is based on hindcast simulator inputs and outputs for 6 representative offshore locations. Terms in the emulator model are listed in Table 1 for completeness.

The hindcast simulator was executed at each of 128 different combinations of tuning input settings ω generated according to a Latin hypercube design on a domain of plausible tuning inputs, specified following consultation with hindcast experts at the Danish Hydraulics Institute. After the removal of a small number of suspect evaluations, a sample of simulator output for 20717 storm peak events was retained for subsequent analysis. Analysis is performed in three stages, using a random partition of the sample into three subsets: (a) 15000 storm peak events to fix the prior uncertainty specification for the components; (b) a further 5217 storm peak events to estimate the correlation structure of the residual processes, and to perform Bayes Linear adjustment as described in Appendix A.1; and (c) the remaining 500 storm peak

Table 1
Emulator terms for hindcast simulator physical covariate (x) and tuning (ω) inputs, and (y) outputs.

Emulator term	Symbol	Description
$g_1^{(x)}(x)$	u	Storm peak wind speed
$g_2^{(x)}(x)$	$\cos\theta$	Cosine of storm peak wind direction, θ
$g_3^{(x)}(x)$	$\sin\theta$	Sine of storm peak wind direction
$g_4^{(x)}(x)$	$\cos s$	Cosine of storm season, s
$g_5^{(x)}(x)$	$\sin s$	Sine of storm season
$g_6^{(x)}(x)$, $g_7^{(x)}(x)$, ..., $g_{11}^{(x)}(x)$	–	Indicators for 6 locations
$g_1^{(\omega)}(\omega)$	–	C dissipation
$g_2^{(\omega)}(\omega)$	–	D dissipation
$g_3^{(\omega)}(\omega)$	–	Percentage current
$g_4^{(\omega)}(\omega)$	–	Kn
$g_5^{(\omega)}(\omega)$	–	Indicator for triad interaction (on/off)
y_1	H_S	Storm peak significant wave height
y_2	T_P	Storm peak wave period
y_3	$\log_{10}(\sigma)$	$\log_{10}$(storm duration)

events to validate model performance.

We estimate the hindcast emulator using the model structure from Equation (4) and the procedure described in Section 3.1, separating regression components corresponding to physical covariate and tuning inputs

$$f_i(x,\omega) = \sum_p \beta_{ip}^{(x)} g_p^{(x)}(x) + \sum_q \beta_{iq}^{(\omega)} g_q^{(\omega)}(\omega) + r_i(x,\omega) + \eta_i \quad . \tag{5}$$

Mathematical details of the estimation procedure are relegated to Appendix A.1. Here we describe in words how the emulator is estimated.

We specify basis functions $\{g_p^{(x)}(x)\}$ and $\{g_q^{(\omega)}(\omega)\}$ as reported in Table 1. Then we make a prior belief specification for the uncertain components of the model, by performing an initial Bayesian linear regression using sub-sample (a). The prior moments for the regression coefficients ($\{\mathrm{E}[\beta_{ip}^{(x)}]\}$, $\{\mathrm{Cov}[\beta_{ip}^{(x)}, \beta_{jr}^{(x)}]\}$, $\{\mathrm{E}[\beta_{iq}^{(\omega)}]\}$, $\{\mathrm{Cov}[\beta_{iq}^{(\omega)}, \beta_{jr}^{(\omega)}]\}$ and $\{\mathrm{Cov}[\beta_{ip}^{(x)}, \beta_{jq}^{(\omega)}]\}$) are fixed using the posterior parameters from this initial regression, and the marginal output covariances $\{\mathrm{Cov}[r_i(x,\omega), r_j(x,\omega)]\}$ are fixed to the covariances of the regression residuals.

We then estimate prior covariances $\{\mathrm{Cov}[r_i(x,\omega), r_j(x',\omega')]\}$ between residual processes at different input settings (x,ω), (x',ω'). This is done by means of a Gaussian process regression, setting correlation lengths to values which provide good predictive performance under leave-one-out cross-validation for sub-sample (b). Once prior specification is complete, we jointly update beliefs about the regression and residual components again using sub-sample (b). We then generate adjusted emulator output for sub-sample (c), and compare emulator with simulator outputs. Fig. 3 shows predictions for sub-sample (c) generated under the prior linear regression fit and the fully updated emulator. Whereas the initial linear regression does relatively well for storm peak H_S, the fully-adjusted emulator predicts very well for all outputs.

We use the second regression component of the fully-adjusted emulator (see Equation (5)) to represent uncertainty about the global effect of the tuning inputs in the emulator developed at Step 1.2 (for which simulator outputs corresponding to only a single combination of tuning inputs was available).

4.1.2. Step 1.2: emulator for all locations and single $\omega^\dagger$

We find that hindcast simulator performance (as approximated by the emulator for Step 1.1) varies within the domain of ω specified prior to analysis, as illustrated in Fig. 4. It can be seen that the effect of varying tuning inputs (over the complete Latin hypercube design) on expectations and 95% credible intervals of predictions is small with respect to the effect of covariate u. For this reason, to reduce computational effort, we decided to gather further hindcast simulator output corresponding to all 23 locations of offshore platforms, with a single central choice $\omega^\dagger$ of tuning inputs chosen in consultation with hindcasting experts. Again, we used CFSR wind inputs corresponding to time periods of 2187 historical storms observed the 37-year period. This resulted in a sample of 50301 storm peak simulator outputs available to explore the effects of physical covariates.

The resulting sample was partitioned at random into three different sub-samples ((a), size 40000), ((b), size 9301) and ((c), size 1000), used as before for prior specification, Bayes Linear adjustment and validation. The model for fitting is

$$f_i(x,\omega) = \sum_p \beta_{ip}^{(x)} g_p^{(x)}(x) + \sum_q \beta_{iq}^{(\omega)} g_q^{(\omega)}(\omega) + r_i(x) + \eta_i \quad , \tag{6}$$

where beliefs about $\{\beta_{iq}^{(\omega)}\}$, and $\{g_q^{(\omega)}(\omega)\}$ are borrowed from Step 1.1, and the residual process does not depend on the tuning parameters, since all simulator evaluations are obtained at the same setting $\omega^\dagger$ of ω. Regression basis functions are chosen to be the same as for Step 1.1. Subsequent model fitting follows the description in Section 4.1.1. The predictive performance of the fully-updated emulator is evaluated by making predictions on sub-sample (c). Predictions generated using both the initial regression fit and the fully-updated emulator are shown in Fig. 5. Describing the effects of physical covariates is clearly more challenging over 23 locations. The final emulator adopted following Step 1.1 and Step 1.2 is then Equation (6), where the second term on the right-hand side is borrowed from the emulator at Step 1.1, and all other terms correspond to the emulator at Step 1.2.

4.2. Step 2: propagating best input uncertainty

Following Section 3.2, we make a best input assumption (see Section 3.1) to relate beliefs about $f(x,\omega)$ to beliefs about $y(x)$ at the corresponding system input x. Then we propagate uncertainty about the best input ω^* through the simulator. Uncertainty about ω^* is summarised by a probability distribution $p(\omega^*)$; in this analysis, we consider that we have no reason to favour any setting of ω^* over any other, and so we assume that $p(\omega^*)$ is a product of independent uniform distributions for the components of ω^* over the original input domain for ω specified. The procedure for computing expectations $\{\mathrm{E}[f_i^*(x)]\}$ and covariances $\{\mathrm{Cov}[f_i^*(x), f_j^*(x')]\}$ is summarised in Appendix A.2.

4.3. Step 3: estimating the system model

Following Section 3.2, here we combine the emulator developed in Section 4.1 and Section 4.2 with a prior specification for discrepancy components between simulator and offshore measurement, and update these components jointly to estimate a final system model for real ocean wave environments offshore Denmark. The procedure used is again similar to that for emulator fitting.

Measurements from 7 of 23 platforms were available for at least some interval during the 37 years for which storm characteristics were simulated. Time intervals for which storm measurements were available differ between platforms, with occasional missing and dubious measurements. All three storm peak characteristics corresponding to simulator outputs of interest were observed for a sample of 4615 storm peak events in total. This sample is used to estimate the system model.

We partition the sample at random into 3 sub-samples and use them as follows: sub-sample (a) consists of 1000 measurements, to fix prior specification for discrepancy components; sub-sample (b) consists of 3315 measurements, to update all components of the system model jointly; and sub-sample (c) consists of 300 measurements, to estimate predictive performance of the fitted system model.

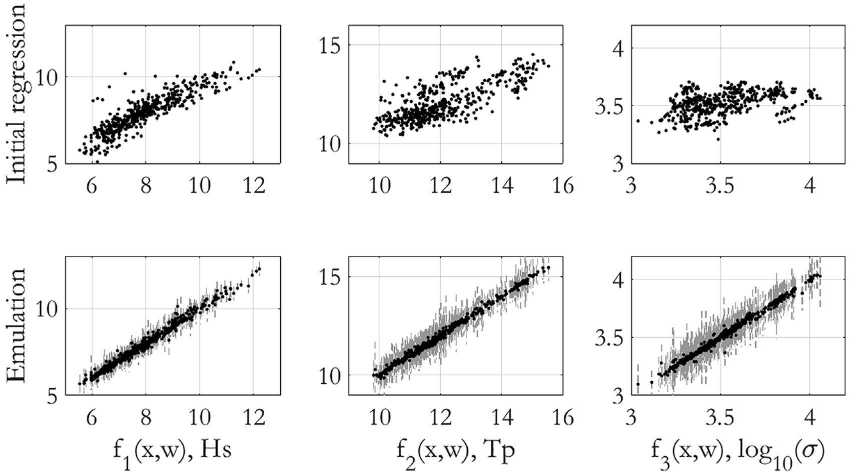

Fig. 3. Emulation for Case A: The top windows show predictions for sub-sample (c) generated from the prior linear regression surface, and the bottom windows show corresponding predictions from the fully-adjusted emulator. Vertical grey lines are 95% credible intervals for emulator predictions.

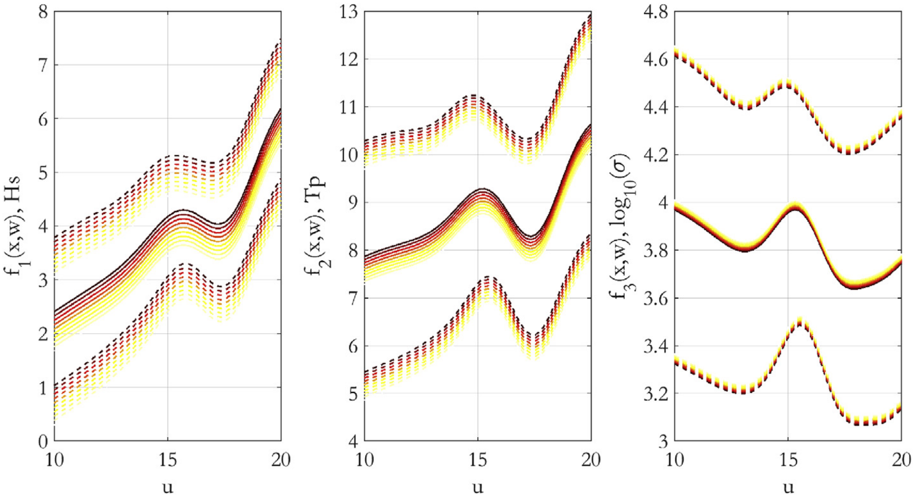

Fig. 4. Emulation (Case A, Step 1.1) of storm peak significant wave height (H_S), storm peak period (T_P) and (the $\log_{10}$ of) the number of waves in a storm, as a function of storm peak wind speed (u). Lines indicate medians and 95% credible intervals. Colour scale indicates varying settings of the tuning parameters ω. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

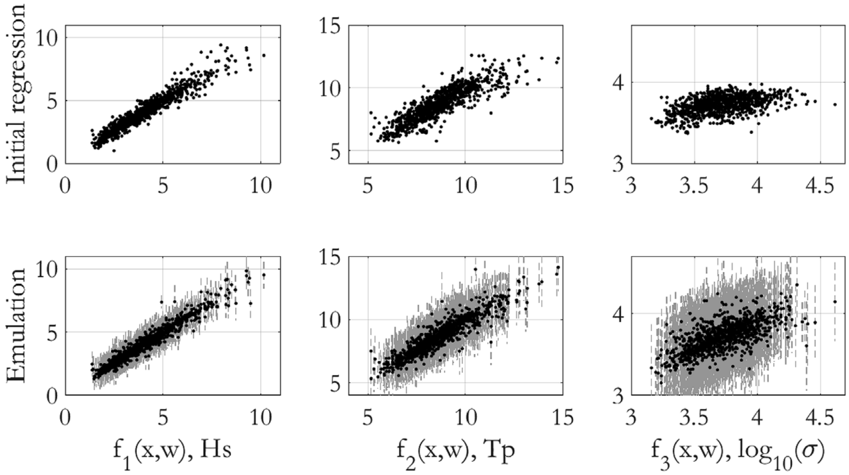

Fig. 5. Emulation of storm peak significant wave height (H_S), peak period (T_P) and storm length (σ), Case A, Step 1.2: The top windows show predictions generated from the prior linear regression, and the bottom windows show corresponding predictions from the fully-adjusted emulator. Vertical grey lines are 95% credible intervals for emulator predictions.

Referring to Equation (1) and Appendix A, linear regression on sample discrepancies $D = \{D_{ik}\}$, where $D_{ik} = z_{ik} - \mathrm{E}[f_i^*(x_k)]$ for each storm measurement z_{ik} from sub-sample (a) is used to obtain a prior specification for discrepancy components. Linear regression mean and covariance parameters $\{\mathrm{E}[\alpha_{ip}]\}$ and $\{\mathrm{Cov}[\alpha_{ip}, \alpha_{jq}]\}$ are adopted as prior regression parameter moments, and residual covariances $\{\mathrm{Cov}[w_i(x), w_j(x)]\}$ are fixed empirically using covariances of regression residuals.

We then estimate prior covariances $\{\mathrm{Cov}[w_i(x), w_j(x')]\}$ between residual processes at different input settings x, x'. This is done by means of a Gaussian process regression, setting correlation lengths to values which provide good predictive performance under leave-one-out cross-validation for sub-sample (b). Having made a prior specification for moments, we then jointly update both simulator and discrepancy moments using sub-sample (b).

Finally, sub-sample (c) is used to assess predictive performance. Fig. 6 shows predictions for sub-sample (c) generated under the prior linear regression fit and under the fully-adjusted system model.

4.4. Modelling case B

Inference for Case B, introduced in Section 2, follows the same procedure as that described above for Case A. For Case B, hindcast simulator output was only made available for one setting $\omega^\dagger$ of tuning inputs ω. Hence, the effect of tuning input variation is again approximated by including the linear regression component estimated for Case A in Section 4.1.1. Hindcast simulator evaluations are available for 345 storms with high storm peak wind speeds over a 1200 year period at 18 locations. A sample of 6210 simulator evaluations was therefore available for modelling. Emulator estimation was based on a random partitioning of this sample into 1500 individuals for initial regression (sub-sample (a)), 4510 to estimate residual correlation structure and jointly update regression and residual components (sub-sample (b)), and 200 to assess predictive performance (sub-sample (c)). The associated system model was similarly estimated, using the same measured data as for Case A.

5. Extreme value analysis

Emulator and system models developed in Section 4 allow the simulation of wave environments (in terms of storm peak H_S, T_P and storm length σ) offshore Jutland. Our aim now is to use these models to estimate the distribution of the 100-year maximum storm peak significant wave height H_S.

A number of different estimators are feasible, based directly on extreme value analysis of measured or simulated H_S, or indirectly on extreme value analysis of storm peak wind speed u propagated through a system model for H_S in terms of u and other physical covariates. In all cases, a method of extreme value analysis accommodating covariate variation is required. Here we describe a simple approach to non-stationary extreme value analysis following Ross et al. (2018) applicable to peaks over threshold of a variable y conditional on covariate x, such as H_S (measured or hindcast) with wave direction ϕ, or storm peak wind speed u with wind direction θ.

We adopt a piecewise stationary extreme value model as a particularly simple description of non-stationarity of y with respect to x. For each observation (x_i, y_i) in the sample $\{x_i, y_i\}_{i=1}^n$, the value of covariate x_i is used to allocate the observation to one and only one of m covariate intervals $\{C_k\}_{k=1}^m$ by means of an allocation vector A such that $k = A(i)$. For each k, all observations in the set $\{y_{i'}\}_{A(i')=k}$ with the same covariate interval C_k are assumed to have common extreme value characteristics.

Threshold exceedances of y are assumed to follow the generalised Pareto distribution with shape $\xi \in \mathbb{R}$ and scale $\zeta_k > 0$, with cumulative distribution function

$$F_{GP}(y; \xi, \zeta_k, \psi_k) = 1 - (1 + (\xi/\zeta_k)(y - \psi_k))^{-1/\xi}$$

for $y \in (\psi_k, y_k^+)$ where $y_k^+ = \psi_k - \zeta_k/\xi$ when $\xi < 0$ and ∞ otherwise. Since estimation of shape parameter is particularly problematic, ξ is assumed constant (but unknown) across covariate intervals, and the reasonableness of the assumption assessed by inspection of diagnostic plots. Threshold parameters $\{\psi_k\}$ are estimated empirically per covariate bin as the bin quantile with given non-exceedance probability $\tau \in [0,1]$. The joint posterior distribution of parameters ξ, $\{\zeta_k\}$ is estimated using Markov chain Monte Carlo with a Metropolis-Hastings-within-Gibbs algorithm. Uniform prior distributions on intervals $\mathcal{I}_\xi$, $\{\mathcal{I}_{\zeta_k}\}$ are assumed for ξ and $\{\zeta_k\}$, and the value of τ is sampled using a Gaussian random walk restricted to interval $\mathcal{I}_\tau$ judged reasonable from inspection of diagnostic plots. For the analysis reported here, we set $\mathcal{I}_\xi = [-0.5, 0.1]$, $\mathcal{I}_{\zeta_k} = [0.01, 100]$ and $\mathcal{I}_\tau = [0.9, 0.95]$. Further details of the piecewise stationary model are given in Ross et al. (2018). We choose to outline a typical extreme value analysis using the storm peak wind speed u and direction θ data from Case B, noting that the equivalent analysis was undertaken for all sources of storm peak wind and significant wave height for Cases A and B.

Inspection of diagnostic plots for u on θ in Case B suggested that partitioning the sample by θ into four directional quadrants (with cardinal directions as boundaries) is reasonable. Fig. 7 shows parameter estimates for extreme value threshold $\{\psi_k\}$, generalised Pareto scale $\{\zeta_k\}$

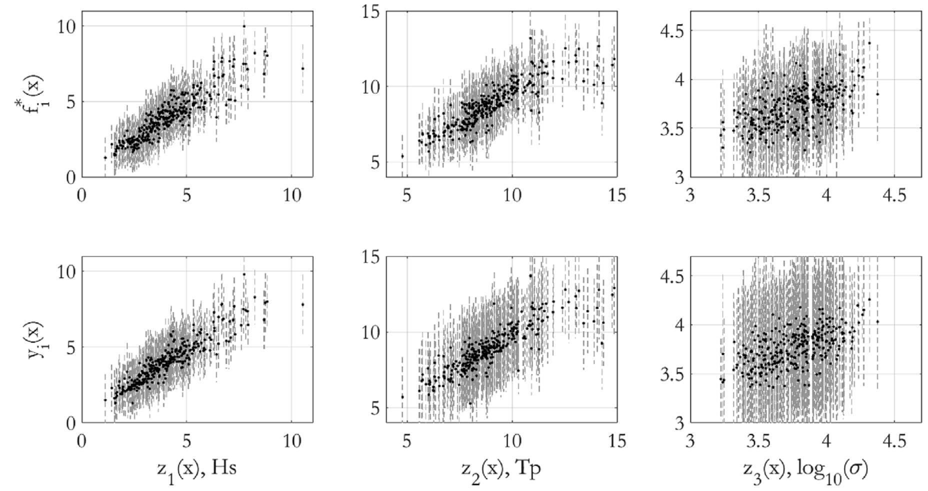

Fig. 6. System model of storm peak significant wave height (H_S), peak period (T_P) and storm length (σ), Case A, Step 3: The top windows compare predictions from the emulator $f^*(x)$ with measurements for sub-sample (c), and bottom windows show predictions y for the same data using the fully-adjusted system model.

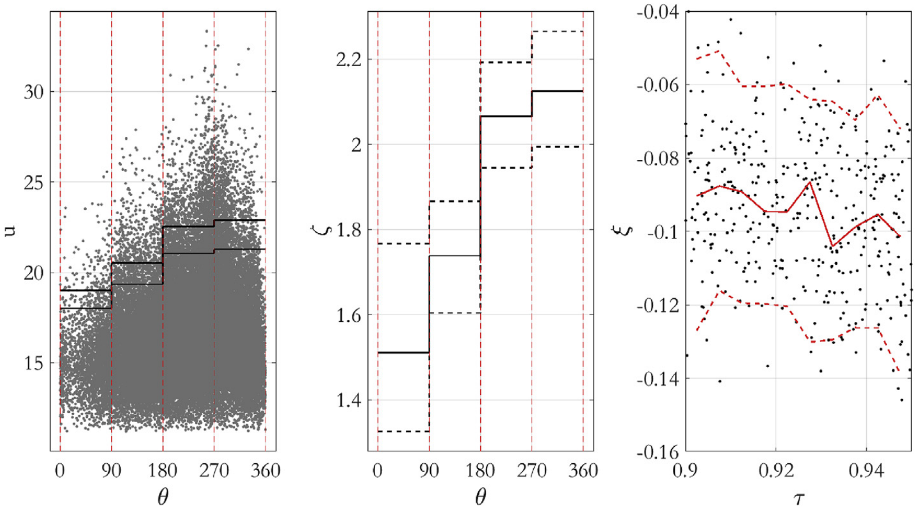

Fig. 7. Piecewise stationary modelling for storm peak wind speed (u) at central location C. The left hand panel shows the sample of u on storm peak wind direction (θ); it also shows the values of extreme value threshold (ψ) per covariate bin corresponding to the limits of interval $\mathcal{I}_\tau$ for the threshold non-exceedance probability (τ). Boundaries of covariate intervals are shown in red. The centre panel gives estimates for the generalised Pareto scale parameter (ζ) in terms of its posterior median and 95% credible interval per covariate bin. The right hand panel shows the estimated stationary generalised Pareto shape parameter (ξ) in terms of it posterior median and 95% credible interval as a function of τ. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and shape ξ.

Fig. 8 compares the tails of u per covariate interval with estimates under the fitted model. There is good agreement between empirical and model-based tails. From the figure we also see that tails corresponding to $\theta \in (180{,}270]$ and $(270{,}360]$ are considerably longer than elsewhere. We would expect these features to be reflected in Fig. 9, illustrating posterior predictive estimates for the distribution of the 100-year maximum storm peak H_S per covariate interval of storm peak wind speed θ, and "omni-directionally" over all covariate intervals. The omni-directional 100-year maximum is dominated by the most severe quadrants as expected.

6. Estimation of distributions of *N*-year maxima

The purpose of this section is to compare different estimates for the distribution of the 100-year maximum storm peak H_S at central location C. A total of six estimates are made. Estimates correspond to different combinations of data source (Cases A and B), and analysis type (e.g. extreme value analysis of measured or hindcast H_S, or extreme value analysis of u propagated through a system model for H_S).

The first estimate ("Measured" in Fig. 10) is obtained by extreme value analysis of the measured data available at C. With $F_H(h|\phi, \mathscr{G})$ representing generalised Pareto cumulative distribution function for storm peak H_S given wave direction ϕ for generalised Pareto parameters $\mathscr{G} = (\{\psi_k\}, \{\sigma_k\}, \xi)$, the cumulative distribution function $F_{H_{100}}$ of the 100-year maximum H_S is a generalised extreme value distribution (e.g. Jonathan and Ewans, 2013)

$$F_{H_{100}}(h|\phi, \mathscr{G}) = \exp[-\rho_{100}(\phi)(1 - F_H(h|\phi, \mathscr{G}))] \quad ,$$

where $\rho_{100}(\phi)$ is the expected number of occurrences of storm peak events in 100 years for covariate value ϕ, and the corresponding posterior predictive distribution for covariate interval $\mathcal{I}_\phi$ is

$$F_{H_{100}}(h|\phi \in \mathcal{I}_\phi) = \int_{\mathcal{I}_\phi} F_{H_{100}}(h|\phi, \mathscr{G}) f(\mathscr{G}|\phi) f(\phi)\, d\phi d\mathscr{G} \quad .$$

From Fig. 10, we see that the median value of the 100-year maximum is approximately 11.3 m, but that the 95% credible interval is very wide, and certainly includes [9.7, 14] m. The second estimate ("Hindcast A" in Fig. 10) is obtained by extreme value analysis of the

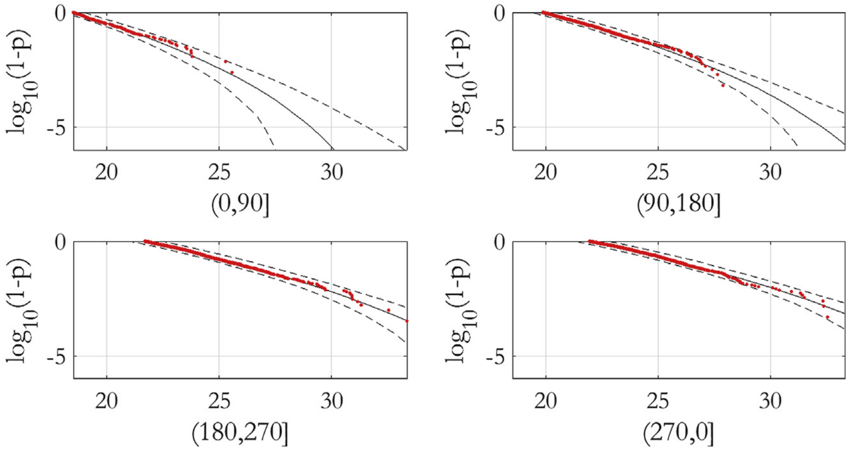

Fig. 8. Plots of tail probability (on $\log_{10}$ scale) for storm peak wind speed (u) by covariate interval of storm peak wind direction (θ), comparing an empirical estimate from the original sample (red) with the posterior median and 95% credible intervals from the fitted model. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

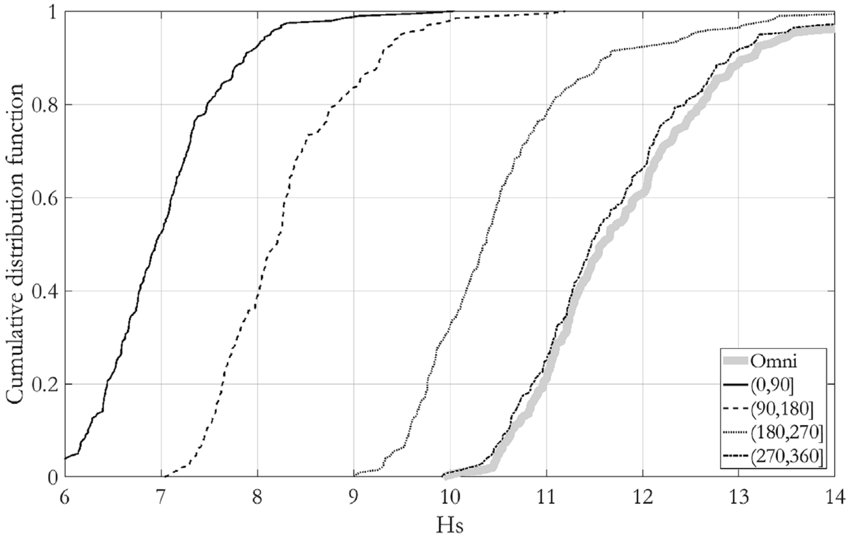

Fig. 9. Posterior predictive distributions for 100-year maximum storm peak significant wave height (H_S) per covariate interval of storm peak *wind* direction (θ) and omni-directionally.

hindcast data at C for Case A, in the same way. The median value is approximately 11.0 m, with a much narrower 95% credible interval.

The third estimate (“Emulator A”) is obtained by performing extreme value analysis *on winds* for Case A (as illustrated in Section 5), estimating the distribution of 100-year maximum wind (as for the “Measured” estimate), then using the emulator developed in Case A to estimate the corresponding distribution of 100-year wave. Writing the cumulative distribution function of storm peak H_S from the emulator, given storm peak wind speed u, wind direction ϕ and season s, as $F_H(h|u, \phi, s)$, the posterior predictive distribution for the 100-year storm peak H_S becomes

$$F_{H100}(h|\phi \in \mathcal{I}_\phi) = \int_{I_\phi} F_H(h|u, \phi, s) f_{U_{100}}(u|\phi, \mathscr{G}) f(\mathscr{G}|\phi) f(\phi|s) f(s)\, ds d\phi d\mathscr{G} du \quad ,$$

where $f_{U_{100}}(u|\phi, \mathscr{G})$ is the probability density function for the 100-year maximum wind speed. The probability density function for $\mathscr{G}|\phi$ is estimated in the extreme value analysis. Probability density functions for $\phi|s$ and s are estimated empirically from the original sample. The median 100-year maximum storm peak H_S using this estimator is approximately 12.1 m, with a distributional width similar to that of the “Hindcast A” estimator. The fourth estimate (“System A”) is similar to the third, but that the full system model for Case A (consisting of emulator and discrepancy model) is used, rather than the emulator model alone. The median 100-year maximum value estimate is now approximately 11.8 m.

The remaining estimates “Emulator B” and “System B” are the analogues of “Emulator A” and “System A” estimates using data from Case B rather than Case A. Median estimates are 11.0 m, 11.4 m and 11.6 m respectively.

In assessing the results in Fig. 10, one key comparison is between estimates for the distribution of the 100-year maximum event from measurements only (solid black) and from system models (red). We also include estimates from Emulators A and B (which are components of the corresponding system models) and from Hindcast A for interest only. The distribution of the 100-year maximum event from measured data is relatively wide due to the small sample size of measurements available, and consequent relatively large epistemic uncertainty.

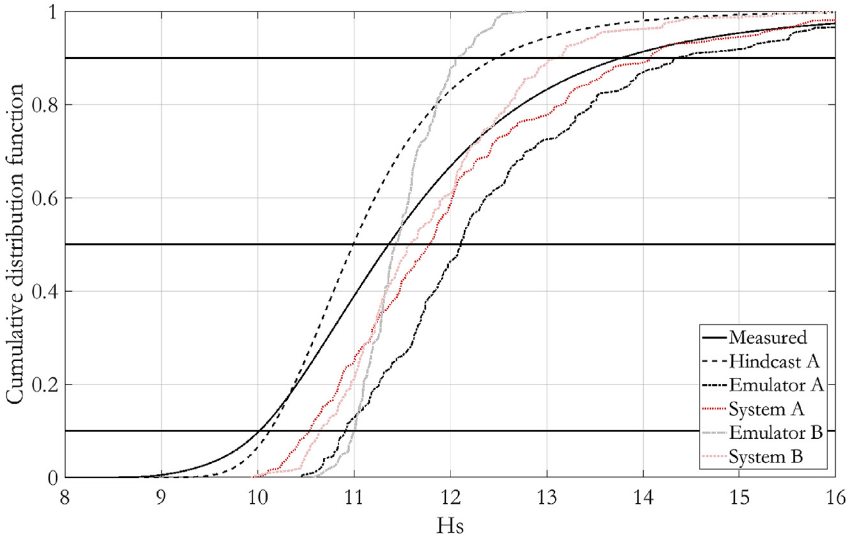

Fig. 10. Comparison of posterior predictive distributions for the 100-year maximum storm peak significant wave height (H_S) using different estimators. Measured: from extreme value analysis of measured H_S; Hindcast A: from extreme value analysis of hindcast H_S for Case A; Emulator A: from extreme value analysis of storm peak wind speed (u), propagated through H_S emulator for Case A; System A: from extreme value analysis of u, propagated through H_S system model for Case A. Labels Emulator B and System B refer to corresponding estimates for Case B.

Specifically, a total of 496 measurements was available, corresponding to a sampling period of approximately 16 years. To incorporate the effects of threshold uncertainty in extreme value inferences, a random extreme value threshold corresponding to a quantile with non-exceedance probability drawn at random from a uniform distribution on $\mathcal{I}_\tau = [0.6{,}0.9]$ was necessary. Hence, on average, a total of only approximately 120 observations was used in the extreme value analysis of measurements. We judge this to be a small sample size. As a result, in this sense, the width of the measured distribution might be expected to be biased high due to sampling variability. If a larger sample was available, we expect that the width of the distribution would reduce.

The corresponding system model estimates (in red) agree well with each other, and are located to the right of the distribution from measurements (at least up to the quantile with non-exceedance probability 0.7). The width of distributions from system models in cases A and B is slightly narrower than that estimated directly using the measured data. Recall that, in a system model, extreme value analysis is actually performed on **wind speed** (rather than on H_S directly), and that an emulator and discrepancy model are subsequently used together to convert extreme wind realisations to those of H_S, facilitating estimation of the cumulative distribution of the 100-year maximum H_S event.

For Emulator A, a total of 1220 observations with threshold $\mathcal{I}_\tau = [0.6{,}0.9]$, corresponding to an average number of observations for modelling of approximately 300 and a period of 37 years, is used for extreme value analysis of wind speed. For Emulator B, a total of 46034 observations with threshold $\mathcal{I}_\tau = [0.9{,}0.95]$, corresponding to a average number of observations for modelling of approximately 3450 and a period of 1200 years, is used for extreme value analysis of wind speed. We judge the relatively narrow width for the distribution of Emulator B to be in part due to the large number of wind observations available for extreme value analysis, and in part to the fact that Case B represents a period of observation of 1200 years (much larger than the 100-year period over which maxima are estimated). Therefore, Emulator B requires a lesser degree of extrapolation of the wind extreme value model and the wind-wave emulator. Since each system model is composed of the sum of emulator and discrepancy model components, we might expect that the width of the distribution from an emulator alone would be no larger than that from a full system model; Fig. 10 supports this finding.

It is interesting further to note that the estimated distribution from Emulator A is located to the left of the corresponding system model, by some 0.4 m at the median quantile level. In contrast, the estimated distribution from Emulator B shows quite a different shape to that based on the corresponding system model. These observations illustrate the importance of the discrepancy model, estimated over multiple locations, in adjusting the emulator towards observations.

We note that the distribution width of hindcast-based estimate Hindcast A in Fig. 10 is smaller than that based directly on measurements. The number of observations used to estimate the extreme value model for Hindcast A is 1220 with threshold specification $\mathcal{I}_\tau = [0.6{,}0.9]$, and hence on average approximately 300 observations are used for extreme value analysis. Therefore, all other things being equal, we expect a lower sampling uncertainty for the distribution of the 100-year maximum from Hindcast A than for the distribution estimated using measurements only; this is observed in the figure.

7. Discussion

In this work, we use Bayesian uncertainty analysis to propagate uncertainties due to approximate physical simulators of extreme wave environments, and uncertainties due to imperfect measurements of that environment, into estimates for the distribution of 100-year maximum storm peak significant wave height H_S. Statistical emulators, discrepancy and system models are estimated using Bayes Linear inference, and coupled to Bayesian non-stationary extreme value analysis. Using hindcast simulator data and offshore measurements for two scenarios (Case A and Case B), six different estimates for the median 100-year maximum storm peak H_S (see Fig. 10) are found, all of which lie in the interval [11.0,12.1]m. Estimates for the median 100-year maximum storm peak H_S based directly on extreme value analysis of hindcast waves (i.e. “Hindcast A” and “Hindcast B” estimates in Fig. 10) are lower than the corresponding estimates based on extreme value analysis of wind speed. There is good agreement between System A and System B estimates, despite the fact that Emulator A and Emulator B estimates are rather different. Overall, we conclude that, for this application, there is little difference in estimates from measurements, from H_S hindcasts or from extreme value analysis of u propagated through emulator or system models for H_S, for Case A and Case B.

We only report extreme value analysis of storm peak H_S in Section 5 and Section 6. Similar marginal analysis has been conducted for storm peak period T_P and storm length σ. We also note that outputs of system models estimated here provide the basis for joint characterisation of extreme wave environments, for example using the conditional extremes model of Heffernan and Tawn (2004) as reported further in Hansen et al. (2018).

We recommend the adoption of a system model approach to estimate the distribution of ***N***-year maxima (and of ***N***-year return values) in metocean design. This approach provides a rigorous, rational, scalable statistical framework for quantifying the relationships between physical models for wind fields, resulting wave fields and measurement characteristics. Uncertainties associated with different model and measurement components are represented explicitly, and can be inferred from the available data in an unambiguous, systematic fashion. The system model combines information from different locations in a coherent manner to improve estimation of the ***N***-year maximum at any of the locations. The system model can be used to estimate the distribution of ***N***-year maxima at locations in an ocean basin, other than those at which measurement or hindcast data are available. The system model estimate should therefore be preferred in general over that obtained from a limited sample of measurements.

In the current application, two different system models (corresponding to different data sources, referred to as Cases A and B) yield estimates of the distribution of the 100-year maximum event showing good agreement. They suggest that the corresponding distribution estimated directly from extreme value analysis of a limited set of measurements (corresponding to approximately 16 years of observation) is underestimated by approximately 0.5 m at quantile levels with non-exceedance probabilities of $\exp(-1)$ and 0.5.

There are numerous extensions and improvements which should be considered. It would be interesting to perform direct emulation of the wave hindcast simulator, rather than of its storm peak characteristics. This might allow us to investigate the possibility of particular (potentially not the most extreme) wind field occurrences generating extreme wave events. This would then allow for a more detailed characterisation of the relatively well-understood wind-wave physics involved, and possibly increase the importance of characterising the wind field and its uncertainty. The resulting emulator and system models might be more complex, but would focus attention on what the wave hindcasting community already knows: that the wind field description is key to accurate hindcasting. Further, there is scope to perform a full history-matching to more fully characterise the domain of plausible ω, and hence the probability distribution of best tuning input ω^*. In the work reported here, the full domain of tuning parameters ω considered corresponds to relatively good emulator performance. More generally, had this domain been defined more loosely, history matching would have been necessary to identify sub-domains corresponding to good emulator performance. Finally, there is considerable scope for improving emulator quality for the worst storms. In the current work, the emulator was estimated based on a sample of physical covariates (e.g. wind speed and direction) corresponding to actual historical occurrences (or estimates thereof). We could have estimated the hindcast emulator using hindcast simulator output for a space-filling design of physical covariates on an

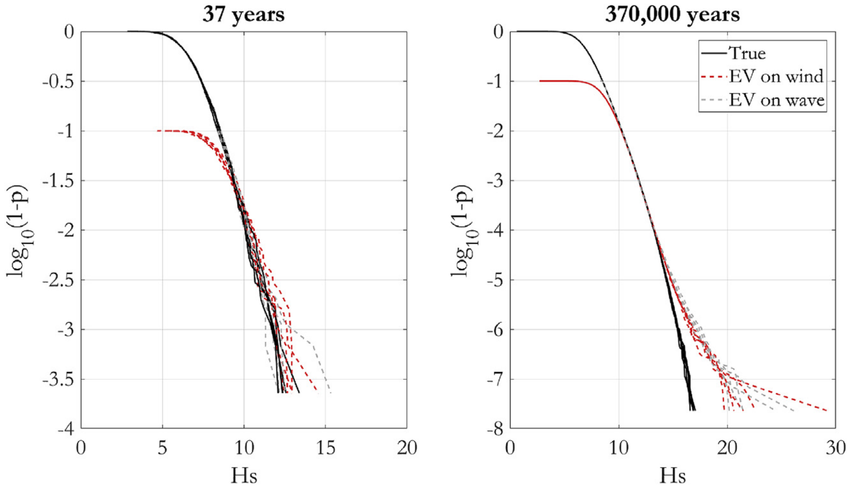

Fig. 11. Illustrations of tail distributions for storm peak significant wave height (H_S) from a simulation study. The assumed environment has generalised Pareto distributed storm peak wind speed (u), and a non-linear wave regression with Gaussian error for H_S as a function of u. Five realisations of true tails for samples corresponding to 37 years of observations are shown in solid black in the left hand panel. As explained in Section 7, tails based on extreme value analysis of samples of u for the same time period, then propagated through the wave regression model to obtain H_S, are shown in dashed red. Tails based on extreme value analysis of H_S obtained by propagating samples of u for the same time period through the wave regression are shown in dashed grey. The right hand panel shows the corresponding tails for a period of 370,000 years (10^4 times the length of the original sample). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

extended domain, exploring unusual combinations of covariates, and in particular including large (even potentially unphysically large) winds to improve our emulator approximation for the most extreme storms.

In light of the comparisons made in Section 6, it is interesting to consider, in general terms, whether it is better (a) to perform extreme value analysis on storm peak wind speed u (non-stationary with respect to storm peak wind direction θ and possibly other covariates) and propagate predicted extreme winds through a system wave model to estimate a distribution for the 100-year maximum of storm peak H_S, or (b) to propagate a sample of storm peak winds (and directions) through the system wave model, and perform extreme value analysis on the resulting sample of storm peak H_S. We investigate this further here using a simple "toy" example motivated by the models estimated in Section 6, assuming for simplicity that storm peak u and H_S are stationary with respect to covariates. Specifically, we assume from inspection of Fig. 7 that in reality $u \sim \mathrm{GP}(\xi, \zeta, \psi)$ where shape $\xi = -0.09$, scale $\zeta = 1.8$ and threshold $\psi = 20$. We further assume that in reality $h = au^b + \nu$ where, from inspection of Fig. 4, $a = 0.05$, $b = 1.6$ and $\nu \sim \mathrm{N}(0,1)$. We then simulate 5 realisations of samples of size $n = 2200$ (corresponding to 37 years of recent extreme storms in Case A, shown in solid black in the left hand panel of the tail plot in Fig. 11), and 5 realisations of size $n' = 2.2 \times 10^7$ (corresponding to 370,000 years of extreme storms in Case A, shown in solid black in the right hand panel of Fig. 11).

The effect of sampling uncertainty on tail location is clear in the left hand panel, but is negligible to a tail probability of 10^{-7} in the right hand panel. The dashed red lines in the left hand panel correspond to approach (a), using tail estimates obtained by performing extreme value analysis on u for threshold exceedances of the 90% sample quantile for a sample of size n. The fitted generalised Pareto model is then used to simulate a further sample of size $n/10$ which is propagated through the wave model. The analogous dashed red lines in the right hand panel are obtained by using n' instead of n. The dashed grey lines in the left hand panel correspond to tail estimates obtained using approach (b), by simulating a sample of u of size n, then propagating the sample through the wave model to obtain a sample of H_S. Extreme value analysis is then performed on threshold exceedances of the 90% sample quantile for this sample, and the fitted generalised Pareto model used to simulate a further sample of size $n/10$ for plotting. The dashed grey lines in the right hand panel are again obtained using n' not n. It can be seen that the variability between realisations of the dashed red (approach (a)) and grey (approach (b)) tails is comparable in the left and right hand panels. Further, tail estimates show the same degree of bias with respect to true (solid black) tails. This suggests, for the model and parameters considered, there is little difference between approaches (a) and (b).

This work constitutes an illustration of Bayesian uncertainty analysis applied to met-ocean design. Whereas the current research should be considered preliminary in many respects, we hope it provides a reasonable demonstration that uncertainty analysis is possible and useful in providing rational estimation and comparison of estimates for extreme events such as the N-year maximum.

Acknowledgement

We thank Michael Goldstein of Durham University for discussions throughout the course of the work reported here. We further acknowledge the support of colleagues at the Danish Hydraulics Institute, TOTAL E&P Danmark (formerly Maersk Oil) and Shell.

Appendix A. Estimation using Bayes linear analysis

This appendix provides a mathematical outline of the estimation of the hindcast emulator, the discrepancy model and hence the system model using Bayes Linear analysis. We start by describing the fundamentals of a Bayes Linear analysis. Following Goldstein and Wooff (2007), suppose we want to learn about quantities $A = \{A_1, ..., A_{n_a}\}$ and $B = \{B_1, ..., B_{n_b}\}$. Our knowledge of these quantities is summarised through expectations $\mathrm{E}[A_i]$, $\mathrm{E}[B_i]$, and (co-) variances $\mathrm{Var}[A_i]$, $\mathrm{Var}[B_i]$ and $\mathrm{Cov}[A_i, B_j]$ of components, referred to as moments. A Bayes Linear analysis proceeds in two steps. First we make a prior specification for these moments. This can be achieved by analysis of an initial sample of data, or from other prior information about the characteristics of the quantities. Secondly, when new data $D = \{D_1, ..., D_{n_D}\}$ become available, we update or adjust our beliefs about A and B, given D. Specifically, we adjust all the moments above, using expectations $\mathrm{E}[D_i]$ and (co-) variances $\mathrm{Var}[D_i]$, $\mathrm{Cov}[A_i, D_j]$ and $\mathrm{Cov}[B_i, D_j]$ as follows. The adjusted expectation of A given D is

$$\mathrm{E}_D[A] = \mathrm{E}[A] + \mathrm{Cov}[A, D]\mathrm{Var}[D]^{-1}[D - \mathrm{E}[D]] \quad ,$$

where $\mathrm{E}[A]_i = \mathrm{E}[A_i]$, $\mathrm{Var}[D]_{ij} = \mathrm{Cov}[D_i, D_j]$ etc. The adjusted covariance between A and B given D is

$$\mathrm{Cov}_D[A, B] = \mathrm{Cov}[A, B] - \mathrm{Cov}[A, D]\mathrm{Var}[D]^{-1}\mathrm{Cov}[D, B] \quad .$$

These equations are used repeatedly in the current work to estimate the hindcast emulator, the discrepancy model and the system model, as explained below. In Appendix A.1, we illustrate Bayes linear analysis to estimate the hindcast emulator and the best input emulator. In Appendix A.2, we describe how uncertainty about the best input setting is captured in the emulator. Finally, in Appendix A.3, we illustrate joint adjustment of moments related to both emulator and discrepancy to obtain adjusted system moments.

Appendix A.1. Estimating the hindcast emulator

The hindcast emulator is defined by Equation (4), and its characteristics quantified by moments $\{\mathrm{E}[\beta_{ip}]\}$, $\{\mathrm{Cov}[\beta_{ip}, \beta_{jq}]\}$, $\{\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')]\}$ and $\mathrm{Cov}[\eta_i, \eta_j]$. We fit the emulator in two steps: first we find prior estimates for these moments; then we adjust emulator moments using Bayes Linear analysis.

Estimating prior moments

We estimate prior moments for regression coefficients and residual process using linear regression on a set of hindcast simulator inputs $\{(x_1, \omega_1),\ldots,(x_{n_F}, \omega_{n_F})\}$ and outputs $\{F_{ij}\}$, where $F_{ij} = f_i(x_j, \omega_j)$ is simulator output i at input setting $\{x_j, \omega_j\}$.

First we estimate prior moments $\{\mathrm{E}[\beta_{ip}]\}$, $\{\mathrm{Cov}[\beta_{ip}, \beta_{jq}]\}$ for regression coefficients. For computational convenience, we re-arrange the data, stacking simulator outputs so that $\widetilde{F}_{i+(j-1)n_y} = F_{ij}$ and define $\widetilde{G} = G^{\mathrm{T}} \otimes I_{n_y}$. Then we make the standard linear regression assumption that $\widetilde{F} = \widetilde{G}b + e$, and make Gaussian assumptions for the prior distribution of b ($\sim N(0, v_r)$) and error vector e ($\sim N(0, v_e)$). Then a-posteriori $b \sim N(\mu, W)$ where

$$W = \left[\frac{1}{v_e}\widetilde{G}^{\mathrm{T}}\widetilde{G} + \frac{1}{v_r}I\right]^{-1} \quad \text{and} \quad \mu = W\left[\frac{1}{v_e}\widetilde{G}^{\mathrm{T}}\widetilde{F}\right] \quad .$$

Prior moments for regression coefficients are thus set at $\mathrm{E}[\beta_{ip}] = \mu_{i+(p-1)n_g}$ and $\mathrm{Cov}[\beta_{ip}, \beta_{kq}] = W_{(i+(p-1)n_g)(k+(q-1)n_g)}$.

Next we estimate prior marginal covariances $\{\mathrm{Cov}[r_i(x, \omega), r_j(x, \omega)]\}$ of the residual process at a single input setting (x, ω) using residuals from the regression fit above. We compute residuals R as $R_{ij} = F_{ij} - \sum_p \mathrm{E}[\beta_{ip}]G_{pj}$, and set marginal residual covariances to

$$\mathrm{Cov}[r_i(x, \omega), r_j(x, \omega)] = \frac{1}{n_F}\sum_p (R_{ip} - \overline{R}_i)(R_{jp} - \overline{R}_j) = V_{ij} \quad ,$$

where $\overline{R}_i = \frac{1}{n_F}\sum_p R_{ip}$.

To use Equation (4) for prediction (see e.g. Equation (A.2) below), we also need a prior specification for the covariances $\{\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')]\}$ between the residual processes at different input settings (x, ω) and (x', ω'). We assume that these covariances have the form

$$\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')] = V_{ij}\,\rho(x, \omega, x', \omega'|\lambda) \quad ,$$

where, with $v = \{x, \omega\}$ and $\lambda = \{\lambda_1, \lambda_2,\ldots,\lambda_{n_\lambda}\}$, the kernel ρ takes the squared exponential form

$$\rho(v, v'|\lambda) = \prod_j \exp\left[-\frac{\lambda_j}{2}(v_j - v'_j)^2\right] \quad .$$

Once the values of λ are fixed, the above equation can be used to estimate $\mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')]$ at any combination of input settings. We estimate λ using a Gaussian process regression model on residuals R from the regression above (but see discussion of sample partitioning in Section 4.1.1, Section 4.1.2 and Section 4.3), choosing λ to give good predictive performance assessed by leave-one-out cross-validation. The fitting and cross-validation procedure used is outlined in Chapter 2 and Chapter 5 of Rasmussen and Williams (2006). For the purposes of the cross-validation, we assume that R_{ij} is an observation of $r_i(x_j, \omega_j)$, and that r_i is a zero-mean Gaussian process. Again, we stack the differences R_{ij} between the data and corresponding predictions under the regression model to obtain the vector $\widetilde{R}$. We assess the quality of a correlation parameter λ by leaving out each data point $\widetilde{R}_i$ in turn and predicting its value using a Gaussian process fit to the remainder of the data. The criterion used to compare different λ is the sum of the predictive log likelihoods

$$L(\lambda) = \sum_{i=1}^{n_y \times n} l_i(\lambda) \quad ,$$

where

$$l_i(\lambda) = -\frac{1}{2}\log(v_i) - \frac{(\widetilde{R}_i - m_i)^2}{2v_i}$$

is the predictive Gaussian likelihood for the i^{th} residual. The predictive mean and variance for the i^{th} residual are

$$m_i = \widetilde{R}_i - \frac{[\widetilde{K}^{-1}\widetilde{R}]_i}{[\widetilde{K}^{-1}]_{ii}} \text{ and } v_i = \frac{1}{[\widetilde{K}^{-1}]_{ii}} \quad ,$$

where $\widetilde{K} = K \otimes V$, and $K_{ij} = \rho(x_i, \omega_i, x_j, \omega_j|\lambda)$. The availability of these expressions means that we must only invert the matrix $\widetilde{K}$ once for each setting of λ that we evaluate, making this cross-validation procedure computationally efficient. We generate a space-filling collection of λ settings (according to a Latin hypercube) and evaluate $L(\lambda)$ for each setting; we then select the setting which maximises this criterion. This setting is then used for the joint update of the regression and residual components of the model.

Bayes linear adjustment

Henceforth adopting the notation $\{(x_1, \omega_1), \ldots, (x_{n_F}, \omega_{n_F})\}$ and $\{F_{ij}\}$ (where $F_{ij} = f_i(x_j, \omega_j)$)) to refer to a further set of simulator inputs and outputs, we now update prior emulator moments described above using Bayes Linear adjustment. Referring again to Equation (4), the adjusted expectation for emulator output at new (unobserved) input setting $\{x, \omega\}$ is

$$\mathrm{E}_F[f_i(x, \omega)] = \mathrm{E}_F[\beta_{ip}]g_p(x, \omega) + \mathrm{E}_F[r_i(x, \omega)] \quad , \tag{A.1}$$

and the adjusted covariance between two new inputs is

$$\begin{aligned}\mathrm{Cov}_F[f_i(x, \omega), f_k(x', \omega')] = {} & g_p(x, \omega)\mathrm{Cov}_F[\beta_{ip}, \beta_{kq}]g_q(x', \omega') + g_p(x, \omega)\mathrm{Cov}_F[\beta_{ip}, r_k(x', \omega')] \\ & + \mathrm{Cov}_F[r_i(x, \omega), \beta_{kq}]g_q(x', \omega') + \mathrm{Cov}_F[r_i(x, \omega), r_k(x', \omega')] \\ & + \mathrm{Cov}[\eta_i, \eta_k].\end{aligned} \tag{A.2}$$

In these equations, and those following in this appendix, we assume the Einstein summation convention, such that repeated indices are understood to be summed over. We note from Equation (A.1) and Equation (A.2) that adjusting the emulator is equivalent to adjusting the moments $\mathrm{E}[\beta_{ip}]$, $\mathrm{E}[r_i(x, \omega)]$, $\mathrm{Cov}[\beta_{ip}, \beta_{kq}]$, $\mathrm{Cov}[\beta_{ip}, r_k(x', \omega')]$, $\mathrm{Cov}[r_i(x, \omega), \beta_{kq}]$ and $\mathrm{Cov}[r_i(x, \omega), r_k(x', \omega')]$. Computation of these moments is outlined below. First, note that prior moments of simulator outputs can be derived from the prior specification using

$$\begin{aligned}\mathrm{E}[F_{ij}] &= \mathrm{E}[\beta_{ip}]G_{pj} \quad , \\ \mathrm{Cov}[F_{ij}, F_{kl}] &= G_{pj}\mathrm{Cov}[\beta_{ip}, \beta_{kq}]G_{ql} + \mathrm{Cov}[r_i(x_j, \omega_j), r_k(x_l, \omega_l)] + \mathrm{Cov}[\eta_{ij}, \eta_{kl}] \quad ,\end{aligned}$$

where G is a regression design matrix with elements $G_{pj} = g_p(x_j, \omega_j)$.

Regression coefficients have adjusted expectations

$$\mathrm{E}_F[\beta_{ip}] = \mathrm{E}[\beta_{ip}] + \mathrm{Cov}[\beta_{ip}, F_{rs}]\mathrm{Var}[F]^{-1}_{rsvw}[F_{vw} - \mathrm{E}[F_{vw}]]$$

and adjusted covariances

$$\mathrm{Cov}_F[\beta_{ip}, \beta_{kq}] = \mathrm{Cov}[\beta_{ip}, \beta_{kq}] - \mathrm{Cov}[\beta_{ip}, F_{rs}]\mathrm{Var}[F]^{-1}_{rsvw}\mathrm{Cov}[F_{vw}, \beta_{kq}] \quad ,$$

where prior covariances of regression weights with the data are

$$\mathrm{Cov}_F[\beta_{ip}, F_{rs}] = \mathrm{Cov}[\beta_{ip}, \beta_{rt}]G_{ts} \quad .$$

Residual process moments have adjusted expectations

$$\mathrm{E}_F[r_i(x, \omega)] = \mathrm{Cov}[r_i(x, \omega), F_{rs}]\mathrm{Var}[F]^{-1}_{rsvw}[F_{vw} - \mathrm{E}[F_{vw}]]$$

and adjusted covariances

$$\mathrm{Cov}_F[r_i(x, \omega), r_j(x, \omega')] = \mathrm{Cov}[r_i(x, \omega), r_j(x', \omega')] - \mathrm{Cov}[r_i(x, \omega), F_{rs}]\mathrm{Var}[F]^{-1}_{rsvw}\mathrm{Cov}[F_{vw}, r_j(x', \omega')] \quad ,$$

where prior covariances of residual components with data are

$$\mathrm{Cov}[r_i(x, \omega), F_{rs}] = \mathrm{Cov}[r_i(x, \omega), r_r(x_s, \omega_s)] \quad .$$

Finally, cross-covariances between regression coefficients and residual process have adjusted covariances

$$\mathrm{Cov}_F[\beta_{ip}, r_k(x', \omega')] = -\mathrm{Cov}[\beta_{ip}, F_{rs}]\mathrm{Var}[F]^{-1}_{rstu}\mathrm{Cov}[F_{tu}, r_k(x', \omega')] \quad .$$

Tuning the nugget covariance

Finally, we tune the nugget covariances $\{\mathrm{Cov}[\eta_i, \eta_j]\}$ by assuming $\mathrm{Cov}[\eta_i, \eta_j] = aV_{ij}$, and adjusting a manually so that marginally three standard deviation predictive intervals include 95% of predictions for an independent test set.

Appendix A.2. Incorporating tuning input uncertainty in the emulator

Uncertainty about the setting ω^* is represented through a probability distribution $p(\omega^*)$. Our expectation for the simulator $f^*(x)$ evaluated at this best input is

$$\mathrm{E}[f_i^*(x)] = \mathrm{E}[\mathrm{E}_F[f_i(x, \omega^*)]] \quad ,$$

where the outer expectation is taken with respect to $p(\omega^*)$. Our specification for the covariance between $f_i^*(x)$ and $f_j^*(x')$ is

$$\begin{aligned}\mathrm{Cov}\left[f_i^*(x), f_j^*(x')\right] &= \mathrm{E}[\mathrm{Cov}_F[f_i(x, \omega^*), f_j(x', \omega^*)]] + \mathrm{Cov}[\mathrm{E}_F[f_i(x, \omega^*)], \mathrm{E}_F[f_j(x', \omega^*)]] \\ &= \mathrm{E}[\mathrm{Cov}_F[f_i(x, \omega^*), f_j(x', \omega^*)]] \\ &\quad + \mathrm{E}[\mathrm{E}_F[f_i(x, \omega^*)]\mathrm{E}_F[f_j(x', \omega^*)]] - \mathrm{E}[f_i^*(x)]\mathrm{E}\left[f_j^*(x')\right] \quad ,\end{aligned}$$

where, again, the outer expectations are take with respect to $p(\omega^*)$. We use $\{\mathrm{E}[f_i^*(x)]\}$ and $\left\{\mathrm{Cov}\left[f_i^*(x), f_j^*(x')\right]\right\}$ as moments of the best input emulator for subsequent inference.

Appendix A.3. Estimating the system model

Having observed data Z, referring to Equation (1) and Equation (2), the adjusted expectation for the system value at a new (unobserved) input setting x is

$$\mathrm{E}_z[y_i(x)] = \mathrm{E}_z[f_i^*(x)] + \mathrm{E}_z[\alpha_{ip}]h_p(x) + \mathrm{E}_z[w_i(x)] \quad , \tag{A.3}$$

and the adjusted covariance between system values any pair $\{x, x'\}$ of new input values is

$$\mathrm{Cov}_z[y_i(x), y_j(x')] = \mathrm{Cov}_z\left[f_i^*(x), f_j^*(x')\right] + \mathrm{Cov}_z[f_i^*(x), \delta_j(x')] + \mathrm{Cov}_z\left[\delta_i(x), f_j^*(x')\right] + \mathrm{Cov}_z[\delta_i(x), \delta_j(x')] \quad . \tag{A.4}$$

Individual adjusted moments appearing in these expressions are computed below. First we note that prior beliefs about the data Z can be derived from the moments computed in Section Appendix A.1 and the prior specification for the discrepancy and measurement error components using

$$\begin{aligned} \mathrm{E}[z_{ij}] &= \mathrm{E}[f_i^*(x_j)] + \mathrm{E}[\alpha_{ip}]H_{pj} \quad , \\ \mathrm{Cov}[z_{ij}, z_{kl}] &= \mathrm{Cov}[f_i^*(x_j), f_k^*(x_l)] + H_{pj}\mathrm{Cov}[\alpha_{ip}, \alpha_{kq}]H_{ql} \\ &\quad + \mathrm{Cov}[w_i(\omega_j), w_k(\omega_l)] + \mathrm{Cov}[\varepsilon_{ij}, \varepsilon_{kl}] \quad , \end{aligned}$$

where $H_{pj} = h_p(x_j)$ is the usual regression design matrix. Adjusted predictive moments of the simulator are

$$\begin{aligned} &\mathrm{E}_z[f_i^*(x)] = \mathrm{E}[f_i^*(x)] + \mathrm{Cov}[f_i^*(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}[z_{rs} - \mathrm{E}[z_{rs}]] \quad , \\ &\mathrm{Cov}_z\left[f_i^*(x), f_j^*(x')\right] = \mathrm{Cov}\left[f_i^*(x), f_j^*(x')\right] - \mathrm{Cov}[f_i^*(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}\mathrm{Cov}\left[z_{rs}, f_j^*(x')\right] \quad , \end{aligned}$$

where the covariances between new simuator values and system data points are

$$\mathrm{Cov}[f_i^*(x), z_{pq}] = \mathrm{Cov}\left[f_i^*(x), f_p^*(x_q)\right]$$

and the moments of f^* are computed as in Appendix A.2. Adjusted moments of the discrepancy regression coefficients are computed using

$$\begin{aligned} \mathrm{E}_z[\alpha_{ik}] &= \mathrm{E}[\alpha_{ik}] + \mathrm{Cov}[\alpha_{ik}, z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}[z_{rs} - \mathrm{E}[z_{rs}]] \quad , \\ \mathrm{Cov}_z[\alpha_{ik}, \alpha_{jl}] &= \mathrm{Cov}[\alpha_{ik}, \alpha_{jl}] - \mathrm{Cov}[\alpha_{ik}, z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}\mathrm{Cov}[z_{rs}, \alpha_{jl}] \quad , \end{aligned}$$

where the covariances between the coefficients and the system data values are

$$\mathrm{Cov}[\alpha_{ik}, z_{pq}] = \mathrm{Cov}[\alpha_{ik}, \alpha_{pr}]H_{rq} \, .$$

Adjusted moments of the discrepancy residual components are

$$\begin{aligned} \mathrm{E}_z[w_i(x)] &= \mathrm{E}[w_i(x)] + \mathrm{Cov}[w_i(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}[z_{rs} - \mathrm{E}[z_{rs}]] \quad , \\ \mathrm{Cov}_z[w_i(x), w_j(x')] &= \mathrm{Cov}[w_i(x), w_j(x')] - \mathrm{Cov}[w_i(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}\mathrm{Cov}[z_{rs}, w_j(x')] \quad , \end{aligned}$$

where covariances between residuals and system data values are

$$\mathrm{Cov}[w_i(x), z_{pq}] = \mathrm{Cov}[w_i(x), w_p(\omega_q)] \quad .$$

Adjusted covariances between the coefficients and the residual components are

$$\mathrm{Cov}[\alpha_{ik}, w_j(x)] = -\mathrm{Cov}[\alpha_{ik}, z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}\mathrm{Cov}[z_{rs}, w_j(x')] \quad .$$

Adjusted covariances between the simulator components and the discrepancy are

$$\begin{aligned} &\mathrm{Cov}[f_i^*(x), \delta_j(x')] = -\mathrm{Cov}[f_i^*(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}\mathrm{Cov}[z_{rs}, \delta_j(x')] \\ &= -\,\mathrm{Cov}[f_i^*(x), z_{pq}]\mathrm{Var}[z]^{-1}_{pqrs}[\mathrm{Cov}[z_{rs}, \alpha_{jl}]h_l(x') + \mathrm{Cov}[z_{rs}, w_j(x')]] \quad . \end{aligned}$$

Appendix B. Probabilistic Bayesian uncertainty analysis

The procedure for a fully probabilistic Bayesian uncertainty analysis is similar to that for a Bayes Linear analysis described in Section 3 and Appendix A. In both cases, the objective is to provide a statistical description of the system y given knowledge of hindcast simulation evaluations (represented by F) for physical covariate inputs x and tuning inputs ω, and measurements z. Fully probabilistic inference, described in detail in e.g. Kennedy and O'Hagan (2001), is more widespread in the literature, and arguably easier to outline and understand concisely than Bayes Linear inference. But fully probabilistic inference is also considerably more demanding to implement in terms of complexity of full model and prior specification, and computational burden.

In outline for the probabilistic approach, referring to Section 3, a Gaussian prior specification is typically made for hindcast emulator basis parameters β, and a Gaussian process prior specification is made for hindcast emulator residual $r(x, \omega)$. Similarly, Gaussian and Gaussian process prior specifications are made for discrepancy coefficients α and discrepancy residual $w(x)$ respectively. The joint distribution of hindcast simulator outputs F and measurements z is computed, and beliefs about the tuning parameters are obtained using Bayes theorem as

$$p(\omega|z, F) \propto p(z, F|\omega)p(\omega) \quad ,$$

where

$$p(z, F|\omega) = \int p(z, F|\beta, \alpha, \omega)p(\beta)p(\alpha)d\beta d\alpha$$

can be computed in closed-form. Predictions for the system at input settings are then obtained by evaluating

$$p(y(x)|x, z, F) = \int p(y(x)|x, \omega)p(\omega|z, F)d\omega$$

using numerical approximation. The last expression is the full probabilistic Bayesian equivalent to the system adjusted expectation and covariance for Bayes Linear inference given in Equation (A.3) and Equation (A.4).

References

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

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer.

Brooker, D.C., K Cole, G., McConochie, J.D., 2004. The influence of hindcast modelling uncertainty on the prediction of high return period wave conditions. In: In Proc. 23th International Conf. On Offshore Mechanics and Arctic Engineering, Vancouver, Canada.

Cardone, V.J., Graber, H.C., Jensen, R.E., Hasselmann, S., Caruso, M.J., 1995. In search of the true surface wind field in SWADE IOP-1: ocean wave modelling perspective. Glob. Atmos. Ocean Syst. 3, 107–150.

Cardone, V.J., Cox, A.T., 2009. Tropical cyclone wind field forcing for surge models: critical issues and sensitivities. Nat. Hazards 51, 29–47.

Chavez-Demoulin, V., Davison, A.C., 2012. Modelling time series extremes. Rev. Stat. 10, 109–133.

Coles, S.G., Simiu, E., 2003. Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. J. Eng. Mech. 129, 1288.

Cooley, D., Naveau, P., Jomelli, V., Rabatel, A., Grancher, D., 2006. A Bayesian hierarchical extreme value model for lichenometry. Environmetrics 17, 555–574.

Davison, A.C., Hinkley, D.A., 1997. Bootstrap methods and their application (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, Cambridge, UK.

Davison, A.C., Padoan, S.A., Ribatet, M., 2012. Statistical modelling of spatial extremes. Stat. Sci. 27, 161–186.

Ewans, K.C., Jonathan, P., 2008. The effect of directionality on northern North Sea extreme wave design criteria. J. Offshore. Arct. Eng. 130 (041604) 1–041604:8.

Forristall, G., Heideman, J.C., Leggett, I.M., Roskam, B., Vanderschuren, L., 1996. Effect of sampling variability on hindcast and measured wave heights. J. Waterw. Port, Coast. Ocean Eng. 122, 216–225.

Goldstein, Michael, Wooff, David, 2007. Bayes linear statistics: theory and methods, first ed. Wiley, Chichester.

Guedes-Soares, C., Scotto, M., 2001. Modelling uncertainty in long-term predictions of significant wave height. Ocean Eng. 28, 329–342.

Hansen, H.F., Zeeburg, A.R., Randell, D., Jonathan, P., 2018. Seasonal-directional extreme value analysis of North Sea storm conditions. In preparation for Coastal Eng.

Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values. J. Roy. Stat. Soc. B 66, 497–546.

Jonathan, P., Ewans, K.C., 2007. Uncertainties in extreme wave height estimates for hurricane dominated regions. J. Offshore. Arct. Eng. 129, 300–305.

Jonathan, P., Ewans, K.C., 2013. Statistical modelling of extreme ocean environments with implications for marine design: a review. Ocean Eng. 62, 91–109.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B 63, 425–464.

MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M., Russell, G., 2011. A flexible extreme value mixture model. Comput. Stat. Data Anal. 55, 2137–2157.

Mackay, E.B.L., Challenor, P.G., Bahaj, A.S., 2010. On the use of discrete seasonal and directional models for the estimation of extreme wave conditions. Ocean Eng. 37, 425–442.

Malde, N.P., Tozer, S., Oakley, J., Gouldby, B.P., Liu, Y., Wyncoll, D., 2018. Applying Gaussian process emulators for coastal wave modelling. Submitted to Coastal Eng.

Mendez, F.J., Menendez, M., Luceno, A., Medina, R., Graham, N.E., 2008. Seasonality and duration in extreme value distributions of significant wave height. Ocean Eng. 35, 131–138.

Mendez, F.J., Menendez, M., Luceno, A., Losada, I.J., 2006. Estimation of the long-term variability of extreme significant wave height using a time–dependent pot model. Journal Geophys. Res. 11, C07024.

Northrop, P., Attalides, N., Jonathan, P., 2017. Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. J. Roy. Stat. Soc. C 66, 93–120.

Orimolade, A.P., Haver, S., Gudmestad, O.T., 2016. Estimation of extreme significant wave heights and the associated uncertainties: a case study using NORA10 hindcast data for the barents sea. Mar. Struct. 49, 1–17.

Oyebamiji, O.K., Wilkinson, D.J., Jayathilake, P.G., Curtis, T.P., Rushton, S.P., Li, B., Gupta, P., 2017. Gaussian process emulation of an individual-based model simulation of microbial communities. J. Comput. Sci. 22, 69–84.

Papastathopoulos, I., Tawn, J.A., 2013. Extended generalised pareto models for tail estimation. J. Stat. Plann. Inference 143, 131–143.

Randell, D., Turnbull, K., Ewans, K., Jonathan, P., 2016. Bayesian inference for nonstationary marginal extremes. Environmetrics 27, 439–450.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, pp. 248. http://www.gaussianprocess.org/.

Reeve, D.T., Randell, D., Ewans, K.C., Jonathan, P., 2012. Accommodating measurement scale uncertainty in extreme value analysis of North Sea storm severity. Ocean Eng. 53, 164–176.

Reich, Brian J., Shaby, Benjamin A., 2012. A hierarchical max-stable spatial model for extreme precipitation. Ann. Appl. Stat. 6, 1430–1451.

Ross, E., Sam, S., Randell, D., Feld, G., Jonathan, P., 2018. Estimating surge in extreme North Sea storms. Ocean Eng. 154, 430–444.

Saha, Suranjana, Moorthi, Shrinivas, Wu, Xingren, Wang, Jiande, Nadiga, Sudhir, Tripp, Patrick, Behringer, David, Hou, Yu-Tai, ya Chuang, Hui, Iredell, Mark, Ek, Michael, Meng, Jesse, Yang, Rongqian, Pea Mendez, Malaquas, van den Dool, Huug, Zhang, Qin, Wang, Wanqiu, Chen, Mingyue, Becker, Emily, 2014. The NCEP climate forecast system Version 2. J. Clim. 27 (6), 2185–2208.

Sanchez-Archilla, A., G Aguar, J., J Egozcue, J., Prinos, P., 2008. Extremes from scarce data: the role of Bayesian and scaling techniques in reducing uncertainty. J. Hydraul. Res. 46, 224–234.

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. Rev. Stat. 10, 33–60.

Shaffrey, L.C., Stevens, I., Norton, W.A., Roberts, M.J., Vidale, P.L., Harle, J.D., Jrrar, A., Stevens, D.P., Woodage, M.J., Demory, M.E., Donners, J., Clark, D.B., Clayton, A., Cole, J.W., Wilson, S.S., Connolley, W.M., Davies, T.M., Iwi, A.M., Johns, T.C., King, J.C., New, A.L., Slingo, J.M., Slingo, A., Steenman-Clark, L., Martin, G.M., 2009. U.K. HiGEM: the new U.K. high-resolution global environment model. Model description and basic evaluation. J. Clim. 22 (8), 1861–1896.

Sorensen, O.R., Kofoed-Hansen, H., Rugbjerg, M., S Sorensen, L., 2005. A third-generation spectral wave model using an unstructured finite volume technique. In: Proceedings of the 29th International Conference on Coastal Engineering, Lisbon, Portugal, 2004.

Tancredi, A., Anderson, C.W., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9, 87–106.

Vanem, E., 2015. Non-stationary extreme value models to account for trends and shifts in the extreme wave climate due to climate change. Appl. Ocean Res. 52, 201–211.

Vernon, I., Goldstein, M., Bower, R.G., 2010. Galaxy formation: a Bayesian uncertainty analysis. Bayesian Anal. 5, 619–669.

Wada, R., Waseda, T., 2018. Benchmark for sources of uncertainty in extreme wave analysis. In: Proc. 37rd Conf. Offshore Mech. Arct. Eng. (Accepted 2018).

Wadsworth, J.L., Tawn, J.A., Jonathan, P., 2010. Accounting for choice of measurement scale in extreme value modelling. Ann. Appl. Stat. 4, 1558–1578.