

# Data driven analysis on the extreme wave statistics over an area

Tianning Tang<sup>a,1</sup>, Thomas A. A. Adcock<sup>a</sup>

<sup>a</sup>*Department of Engineering Science, University of Oxford, Oxford, UK*

---

## Abstract

In this paper we analyse ocean wave crest statistics over different sized areas using data driven methods. We use second order numerical simulations to generate extreme crest data. We consider a simplistic Gumbel distribution fit as well as using a Random Forest Model to map the sea-state parameters to extreme crest values. Our simulations are compared with the existing distributions in the literature. We find that existing distributions perform well for more straightforward cases but that as more parameters are introduced the data science approach can capture features other methods cannot. Our approach also highlights the importance of different parameters such as steepness or length in the mean wave direction. We conclude that machine learning model is promising approach to predicting wave crest distributions in complex scenarios.

*Keywords:* Wave statistics, Data driven method, Rogue wave

---

## 1. Introduction

Short-term wave crest distributions are essential for the design of offshore infrastructure. For example, an important criterion in the offshore platform design is to leave enough air gap between the mean sea level and the deck, which requires an accurate estimation of the crest distributions based on the sea state parameters. Linear theory predicts the short-term crest distribution closely follows the Rayleigh distribution. However, real water waves are nonlinear and this physics modifies the crest statistics.

---

<sup>1</sup>Email: tianning.tang@eng.ox.ac.uk

9 The water waves which occur naturally in the open ocean are direction-  
10 ally spread. This makes a fundamental difference in the nonlinear wave  
11 interactions and largely inhibits the build up of correlations between freely  
12 propagating components [1, 2, 3]. This indicates that extra amplification is  
13 usually small when compared to statistics predicted by second order models  
14 provided the spectrum is in equilibrium. This has been shown in experimen-  
15 tal and numerical simulations in [4, 5, 6, 7] as well as by field data [8]. Thus,  
16 in the present paper we model the waves as second order which implies that  
17 the underlying dynamics are linear.

18 Second order harmonics make the crests higher than predicted by lin-  
19 ear theory, which is particularly important in steep sea states. Forristall [9]  
20 proposed a Weibull distribution with empirical fitted parameters to capture  
21 the additional amplitude from second order contribution. His distribution  
22 is widely used in engineering practice. Fedele and Tayfun [10] replaced the  
23 empirical fitting with a rigorous theoretical framework. The existing crest  
24 distribution models predicts wave statistics well when compared to both ex-  
25 perimental and field measurements [10, 11, 12, 13]. Additionally, higher order  
26 wave-wave interactions can further modify the crests statistics [14]. Tayfun  
27 and Fedele [15] have established short term crest distributions accounting  
28 for the additional crests amplitude from these higher order nonlinear inter-  
29 actions (see also some further development of statistics of oceanic crests in  
30 [1, 10, 16]).

31 These models are developed and verified against the waves measured at a  
32 single Eulerian point. For large engineering structures using point statistics  
33 can underestimate the magnitude of the wave crests [17]. Hence, the wave  
34 statistics over an area becomes more relevant for engineering designs [17, 18,  
35 19]. The extra amplitude arises from the fact that a point measurement will  
36 tend to miss the true crest of the wave (which dominates for small areas) as  
37 well as there simply being more waves sampled (which dominates at larger  
38 areas) [19, 20, 21, 22].

39 Instead of the short term crest distributions, we study the maximum  
40 crest distributions, which only considers one maximum crest in a space-time  
41 ensemble. The maximum crest distributions are more stable, and can be pre-  
42 dicted with several existing theoretical models. Adler [23] and Piterbarg's  
43 theorem [24] provide estimations of the impact of the area to the maximum  
44 crest distribution for Gaussian multidimensional random wave fields. Piter-  
45 barg's theorem was extended to second order by Socquet-Juglard et. al.  
46 [25] with a scaling parameter for second order contributions and Forristall

47 [17] with empirical fitting accounting for second order effects. Fedele [20]  
 48 presented a model based on the Adler and Taylor’s Euler characteristics ap-  
 49 proach [23, 26, 27], which is further extended in [22] to account for second  
 50 order physics. These theoretical model have been found to be reasonably  
 51 accurate when compared to field measurements and numerical simulations  
 52 [28, 29].

53 In the present paper we have taken an alternative approach. Rather than  
 54 fitting data with a parameterised distribution we take a data science ap-  
 55 proach to the problem. There has been a rapid expansion of the application  
 56 of these data driven methods in recent years as a results of both the signifi-  
 57 cant improvement in the methodology of data driven methods and the boom  
 58 in computational hardware. The training and validating of these data driven  
 59 methods are more accessible than ever before. As a result, data driven meth-  
 60 ods are starting to be used in ocean environment studies [30, 31, 32, 33, 34],  
 61 and predicting extreme statistics with active sequential sampling methods  
 62 [35, 36, 37, 34, 38] and probabilistic decomposition method [39, 40, 41, 42].  
 63 The object of this paper is primarily to demonstrate the capability of tak-  
 64 ing a machine learning approach to the problem of predicting short-term  
 65 crest statistics from a given input spectrum. We also compare the existing  
 66 distributions against our numerical data.

67 In this paper, we first briefly introduce our simulation method in section  
 68 2.1, which forms the database for model training and comparison. The de-  
 69 tailed dataset information is given in section 2.3. We also present a step-wise  
 70 prediction routine in section 2.4 to clarify the methodology we used to com-  
 71 pare different wave statistical models. We describe the theoretical models  
 72 in section 3.1 and data driven models in section 3.3 and 3.4. We start with  
 73 second order maximum crest distributions at a single point as a limiting case  
 74 in section 4.1. We first aim to examine the feasibility of data driven methods  
 75 under a simplified condition. We further extended the simulation into waves  
 76 over an area, and compared the performance of different models in section 4.2  
 77 for linear waves and in section 4.3 for second order waves. In these sections,  
 78 we increase the complexity of the simulated ocean environment by including  
 79 second order effects to further explore the potential of data driven methods  
 80 in predicting real wave statistics in the open ocean. We summarise the per-  
 81 formance of different models in section 4.4, and present an analysis of the  
 82 relative importance of input parameters in section 4.5.

## 2. Methodology

### 2.1. Numerical simulations

To investigate the extreme wave statistics over an area, we have developed an envelope based code to simulate directional spread waves over an area at finite water depth with optimal speed. For the linear part of the surface elevation  $\eta_{linear}$ , the wave surface elevation can be obtained as:

$$\eta_{linear}(\mathbf{x}, t) = \sum_{i=1}^{\infty} [a_i \cos(\mathbf{k}_i \cdot \mathbf{x} - \omega_i t) + b_i \sin(\mathbf{k}_i \cdot \mathbf{x} - \omega_i t)], \quad (1)$$

where  $\mathbf{x}$  is the position vector,  $\mathbf{k}_i$  is the wavenumber vector for  $i^{th}$  component,  $\omega_i$  is the angular frequency for  $i^{th}$  component, which is connect with wavenumber through dispersion relationship.  $a_i$  and  $b_i$  are the independent random variables drawn from an normal distribution  $N(0, C_i^2)$ . The variance of each wave component  $C_i^2$  is defined by the spectrum of the sea-state as  $C_i^2 = S_o(\omega_i)\Delta\omega$ .  $S_o$  is the omnidirectional wave spectrum and  $\Delta\omega$  is the width of the angular frequency bins. The two dimensional inverse Fast Fourier Transform is applied to improve the speed of the simulations on the linear part.

To further optimise this simulation code for the speed, we also applied a multi-scale time-space frame to obtain the local maximum crest height over a given area and duration. The first simulation is run on a coarse time step at 0.75 seconds (2 steps per period) and a relatively coarse spatial step (0.3 meters) (13 steps per wavelength for deep water and 10 steps for shallowest water case). The time when the local maximum of the envelope occurs is recorded. The second simulation then starts from two periods before that maximum event to two periods after. This simulation is run on a fine time step of 0.05 seconds (30 steps per period) and a fine space step of 0.1 meters (38 steps per wavelength for deep water cases and 28 steps for shallowest water cases) to obtain the accurate value of this maximum crest height. Rigorous tests have been performed to confirm this multi-scale time frame method can capture the maximum events accurately.

All the second order corrections were computed with a hybrid envelope method. The leading order coefficients of the sum terms and difference terms are corrected according to [43]. The coefficients for the broad banded corrections are interpolated from the exact second order simulations [44] (see Appendix A for details). Despite the narrowbanded approximation used the

Table 1: Input and test conditions for all the cases in this study.  $H_s k_p$  is one of the wave steepness parameters we applied in this study,  $H_s$  is the significant wave height and  $k_p$  is the peak wavenumber.  $\lambda_p$  is the peak wave length, which is related with  $k_p$  as  $\lambda_p = 2\pi/k_p$ .  $x$  and  $y$  are the side length of sampling area along mean wave direction and transverse direction respectively, and  $k_p d$  is the normalised water depth parameter and  $d$  is the physical water depth. For all the simulated cases, the peak wave period is 1.5 seconds, and we simulate 150 periods in time domain.

Section	Sampling	Simulation	Input range	Test condition
4.1.1	Point	$2^{nd}$ order	$H_s k_p = 0.0358$ to $0.25$	$H_s k_p = 0.1788$
4.1.2	Point	$2^{nd}$ order	$k_p d = 1.4$ to $7$	$k_p d = 2.5$
4.1.3	Point	$2^{nd}$ order	$H_s k_p = 0.0358$ to $0.25$ $k_p d = 1.4$ to $7$	$H_s k_p = 0.178$ , $k_p d = 2.5$
4.2.1	Square	Linear	$x = y = 0.5\lambda_p$ to $20\lambda_p$	$x = y = 3.5\lambda_p$
4.2.2	Rectangular	Linear	$x, y = 2.5\lambda_p$ to $20\lambda_p$	$x, y = 4\lambda_p$ to $18\lambda_p$
4.3.1	Rectangular	$2^{nd}$ order	$x, y = 2.5\lambda_p$ to $20\lambda_p$	$x, y = 4\lambda_p$ to $18\lambda_p$
4.3.2	Rectangular	$2^{nd}$ order	$x, y = 2.5\lambda_p$ to $20\lambda_p$ , $H_s k_p = 0.075$ to $0.25$	$x, y = 4\lambda_p$ to $20\lambda_p$ , $H_s k_p = 0.1$ to $0.25$

112 results agree closely with simulations using Sharma & Dean [45] type double  
113 summation.

## 114 2.2. Input conditions

In this study, we simulate waves using a directional JONSWAP spectrum  $S(f, \theta) = S_o(f)D(\theta)$  with  $\gamma = 3.3$ , where  $\gamma$  is the peak enhancement factor [46],  $S_o(f)$  is omnidirectional wave spectrum and  $D(\theta)$  is the spreading function. We used a normal directional spreading function  $D(\theta)$  with a spreading parameter of  $G_\theta = 22^\circ$  to specify the energy distribution, which is given by:

$$D(\theta) = \frac{1}{\sqrt{2\pi}G_\theta} \exp\left(-\frac{\theta^2}{2G_\theta^2}\right), \quad (2)$$

115 where  $\theta$  is the angle deviated from the mean wave direction and  $G_\theta$  is the  
116 directional spreading parameter. For all the simulated cases, the peak wave  
117 period is 1.5 seconds, and we simulate 150 periods for all the realisations in  
118 this study.

119 In this study, we have covered seven input conditions, which are sum-  
120 marised in Table 1.

121 We choose the test condition for sections from 4.1.1 to sections 4.2.1 to be  
122 a single sample within the input range. This provides details in the difference  
123 between the shape of maximum crest distributions from different models.

Table 2: Summary of different types of dataset used in this paper

Name	Usage	Realizations per Case	Domain Coverage	Note
<b>Fitting dataset</b>	Empirical Fitting	2000	Grid-based	To obtain stable statistical results
<b>Training dataset</b>	Random Forest	500/2000	Random	RF favours more simulation cases for better coverage
<b>Test dataset</b>	Performance evaluation	5000/10000	Grid-based	To provide the most accurate results

124 However, we choose the relative error in the expected value of maximum  
125 crest to be the test matrix because of its significance in ocean engineering  
126 applications.

### 127 2.3. Datasets types

128 In this study, for each input condition, we simulated three different types  
129 of datasets for model predictions and performance assessment, which are  
130 summarised in Table 2. For all the test cases, a rigorous check on the statis-  
131 tical variance of distribution parameters is performed, additional realisations  
132 are simulated to reduce the confidence interval.

### 133 2.4. Prediction routine

134 In this study, we applied following prediction routine to examine the  
135 performance of different wave statistical models as:

- 136 1. Based on the input parameters listed in Table 1, numerical simulations  
137 are performed to obtain maximum crest distributions.
- 138 2. Gumbel distribution is fitted to parameterise the maximum crest dis-  
139 tributions for fitting, training and test dataset (see Table 2 for dataset  
140 details). Figure 1 shows a typical parameter space of coefficient  $A$  when  
141 only one parameter ( $H_s$ ) is varied. Figure 7 shows a typical parameter  
142 space of coefficient  $A$  (where  $A$  is the parameter for Gumbel distri-  
143 bution shown in Equation 6) when two parameters ( $H_s$  and  $k_p d$ ) are  
144 varied simultaneously.
- 145 3. Wave statistical models are applied based on the input sea state pa-  
146 rameters to predict the maximum crest distributions.

- 147 4. Models predictions are compared with the test dataset to determine  
 148 the error.  
 149 5. Step 1-4 are repeated for all the different input conditions listed in  
 150 Table 1.

### 151 3. Wave statistical models

152 In this study, we reviewed existing models for space-time wave statistics.  
 153 Each model uses slightly different definition of some common wave field pa-  
 154 rameters (e.g. wave steepness). We clarify this in Appendix B with also a  
 155 nomenclature table for reference.

#### 156 3.1. Theoretical models

##### 157 3.1.1. Forristall2006 model

158 For short-term wave distribution, linear waves crests measured at a given  
 159 point closely follow the Rayleigh distribution. For second order waves, For-  
 160 ristall [9] proposed a two parameter Weibull distribution:

$$P(\eta) = \exp \left[ - \left( \frac{\eta}{\alpha_F H_s} \right)^{\beta_F} \right]. \quad (3)$$

161 For deep water waves,  $\alpha_F$  and  $\beta_F$  are coefficient for Weibull distribution,  
 162 which found by Forristall to be:

$$\alpha_F = \sqrt{1/8} + 0.2568S_1 \quad \text{and} \quad \beta_F = 2 - 1.7912S_1, \quad (4)$$

163 where  $S_1$  is one measure of steepness of the sea state, which is defined as:

$$S_1 = \frac{2\pi}{g} \frac{H_s}{T_z^2}, \quad (5)$$

164 where  $T_z$  is the zero-crossing period and  $g$  is the gravitational acceleration.  
 165 However, for most offshore structures, the design of these platforms should be  
 166 able to survive a wave crest reaching any part of the platform area. Hence, it  
 167 is vital for the designer to be able to predict the maximum crests distribution  
 168 over a given area during a certain period.

169 As for maximum crests distribution over an area, the linear statistics  
 170 closely follow the Gumbel distribution according to the Piterbarg theorem  
 171 [24]. For second order waves, Forristall [17] proposed a two-parameter Gum-  
 172 bel distribution:

$$P(\eta_{max} > s\sigma) = \exp\{-\exp[-(-B + s)/A]\}, \quad (6)$$

173 where  $\eta_{max}$  is the space-time maximum surface elevation,  $\sigma$  is the standard  
 174 deviation of the surface elevation,  $s$  is the quantity of interest, parameter  $A$   
 175 and  $B$  are linked with  $\alpha_F$  and  $\beta_F$ :

$$A = (\beta_F/4\alpha_F)(\text{Log}_e N)^{1-1/\beta_F} \quad \text{and} \quad B = 4\alpha_F(\text{Log}_e N)^{1/\beta_F}, \quad (7)$$

176 and  $N$  is the equivalent number of waves, which can be estimated as follows:  
 177 For a point measurement,

$$N = T/T_z, \quad (8)$$

178 where  $T$  is the duration of the time series. For relatively large areas,

$$N = \frac{2\pi xyT}{1.25\lambda_x\lambda_yT_z}, \quad (9)$$

179 where  $\lambda_x$  is the averaged wave length in the mean wave direction and  $\lambda_y$  is  
 180 the averaged wave length in the lateral direction. Both wave parameters and  
 181 irregularity parameters can be calculated from the moments of the directional  
 182 wave spectrum  $S(f, \theta)$ :

$$\lambda_x = 2\pi\sqrt{\frac{m_{000}}{m_{200}}}, \lambda_y = 2\pi\sqrt{\frac{m_{000}}{m_{020}}}, \quad (10)$$

183 where  $m_{i,j,l}$  is the moments of the directional spectrum, which can be com-  
 184 puted as:

$$m_{i,j,l} = \int \int k_x^i k_y^j f^l S(f, \theta) df d\theta, \quad (11)$$

185 where the  $k_x$  and  $k_y$  are the wave number vector component in  $x$  and  $y$   
 186 directions respectively,  $f$  is the frequency bins of the wave spectrum.

187 For relatively small areas,

$$N = \frac{2xT}{\lambda_x T_z}. \quad (12)$$

188 For this theoretical model, the value of  $\alpha_F$  and  $\beta_F$  are obtained by follow-  
 189 ing equation 4. The value of  $A$  and  $B$  are then computed following equation  
 190 7 and 8. The calculated coefficients are used to generate the curve referred  
 191 as ‘Forristall2006’.



### 192 3.2. Forristall2015 model

193 Forristall provides an updated fit for linear waves over an squared area and  
 194 recommends a simple second order correction with a simple multiplication  
 195 in [29]. The updated fit suggests that the effects from the duration and side  
 196 length of the sampling area is additive, the expected value of maximum crest  
 197 height for linear waves can be estimated as:

$$E(\eta_{\max} \mid \text{area}) / \sigma = E(\eta_{\max} \mid \text{point}) / \sigma + F_1(L/\lambda_p), \quad (13)$$

198 where  $E$  is the expected value of probability density, and  $L$  is the side length  
 199 of the squared area, the expected value of for a given Gumbel distribution is

$$E(\eta_{\max}) = (B + 0.5772/A)\sigma, \quad (14)$$

200 and the fitting function  $F_1(L/\lambda_p)$  of side length is given as:

$$F_1(L/\lambda_p) = 0.9829 + 0.4170 \ln(L/\lambda_p) + 0.0427 \ln^2(L/\lambda_p). \quad (15)$$

201 Unfortunately, although this updated fit can provide accurate estimation on  
 202 the expect value of maximum crest height, the performance of this model  
 203 is not fully examined in this study due to the lack of distribution curve  
 204 information.

#### 205 3.2.1. Tromans and Vanderschuren fitting equation

206 This is a fit which follows the report OTC 7683 [47] and the extreme  
 207 crests distribution can be expressed as:

$$P(\eta_{\max} > s\sigma) = \exp\{-\exp[-\text{Log}_e(N_p)((-b_p + s)^2 - 1)]\}. \quad (16)$$

208 We note that the equation 16 is originally intended for maximum wave  
 209 heights distributions with varying wave spectra during a storm at North Sea,  
 210 which is different from the cases studied in this study. As this modified  
 211 Gumbel distribution shows excellent performance in the original report, we  
 212 used a modified Gumbel distribution in similar format for maximum crest  
 213 distributions. There are no theoretical expressions for the value of  $N_p$  and  $b_p$   
 214 in the original report. In this paper, the value of  $N_p$  and  $b_p$  are obtained by  
 215 fitting the results from the simulations.

216 *3.2.2. Fedele2012 model*

217 Fedele [20] presented a model based on the Adler and Taylor's Euler  
 218 characteristics approach [23, 26, 27] for predicting the probabilities of the  
 219 highest crest within the space-time ensemble, which gives:

$$P_{F2012,max}\{\eta_{max}/\sigma > z\} \approx \exp \left\{ - \exp \left[ - (z - h) \left( h - \frac{2N_V h + N_S}{N_V h^2 + N_S h + N_B} \right) \right] \right\}, \quad (17)$$

220 where  $z$  is a threshold parameter for Fedele2012 model and the modal value  
 221  $h$  satisfies:

$$(N_V h^2 + N_S h + N_B) \exp(-h^2/2) = 1, \quad (18)$$

222 where  $N_V$  is the average number of waves within the volume,  $N_S$  for the waves  
 223 on the boundary surfaces and  $N_B$  accounts the waves along the perimeter  
 224 (see Fig.1 in [20] for details). The formula for  $N_V, N_S, N_B$  are given as:

$$\begin{aligned} N_V &= 2\pi \frac{xyT}{\lambda_x \lambda_y T_m} \sqrt{1 - \alpha_{xt}^2 - \alpha_{yt}^2 - \alpha_{xy}^2 + 2\alpha_{xt}\alpha_{yt}\alpha_{xy}}, \\ N_S &= \sqrt{2\pi} \left( \frac{xT}{\lambda_x T_m} \sqrt{1 - \alpha_{xt}^2} + \frac{xy}{\lambda_x \lambda_y} \sqrt{1 - \alpha_{xy}^2} + \frac{yT}{\lambda_y T_m} \sqrt{1 - \alpha_{yt}^2} \right), \\ N_B &= \frac{x}{\lambda_x} + \frac{y}{\lambda_y} + \frac{T}{T_m}, \end{aligned} \quad (19)$$

225 where  $T_m$  is the mean wave period, which can be calculated as:

$$T_m = \sqrt{\frac{m_{000}}{m_{002}}}. \quad (20)$$

226  $\alpha_{xy}, \alpha_{xt}, \alpha_{yt}$  are the irregularity parameters of the sea state, which can be  
 227 calculated as:

$$\alpha_{xt} = \frac{m_{101}}{\sqrt{m_{200}m_{002}}}, \alpha_{yt} = \frac{m_{011}}{\sqrt{m_{020}m_{002}}}, \alpha_{xy} = \frac{m_{110}}{\sqrt{m_{200}m_{020}}}. \quad (21)$$

228 The expected value of the maximum crest height for linear waves is given as:

$$E_{max}/\sigma \approx h + \frac{0.5772}{h - \frac{2N_V h + N_S}{N_V h^2 + N_S h + N_B}}. \quad (22)$$

229 Based on the results in [10], Fedele further extend extended the theory  
 230 with second order corrections in [19] and compare the model prediction with

field data. The extended version of space-time maximum crest distributions is further compared with field data in Benetazzo et. al. [22] and found to be relatively accurate. In the second order version of Fedele2012 model, the probabilities of the highest crest within the space-time ensemble corrected to second order within the narrow banded limit is given as:

$$P_{F2012-2,max}\{\eta_{max}/\sigma > z\} \approx \exp \left\{ - \exp \left[ - \frac{1}{1 + \mu h} \left( z - h - \frac{\mu^2}{2} h^2 \right) \left( h - \frac{2N_V h + N_S}{N_V h^2 + N_S h + N_B} \right) \right] \right\}, \quad (23)$$

where  $\mu$  is an integral measure of wave steepness corrected with spectral bandwidth  $\nu = \sqrt{m_{000}m_{002}/m_{001}^2 - 1}$ , which can be computed as:

$$\mu = \mu_m(1 - \nu + \nu^2), \quad (24)$$

where  $\mu_m = \sigma(2\pi m_{001}/m_{000})^2/g$  is one of a measure of wave steepness.

In the Fedele2012 model, scale dimension parameter  $\beta$  is introduced as a measure of the relative scale of the wave when compared to the volume size. The  $\beta$  can be computed as:

$$\beta = \frac{4N_S\xi + 2N_B}{16N_V\xi^2 + 4N_S\xi + N_B}, \quad (25)$$

where  $\xi$  is the intermediate variable for calculating  $\beta$ , which can be obtained through:

$$(16N_V\xi^2 + 4N_S\xi + N_B) \exp(-8\xi^2) = 1. \quad (26)$$

For linear waves with  $\beta = 1$ , the crest distribution closely follows Rayleigh distribution, which can be interpreted as waves at a point with varying time. The limiting case of  $\beta = 2$  describes the waves with either length in  $x$  or length in  $y$  is null or time duration is instant. For example,  $\beta = 2$  can be attained for waves over an area at an instantaneous time. The maximum value of  $\beta = 3$  suggests a sufficient number of waves can be observed on all three dimensions.

This Fedele2012 model is further extended to include third order nonlinearities by Fedele in [48]. The extended model has been used in the analysis of the sinking of the El Faro and has been compared with Higher Order Spectral simulations [16]. In the present work we do not examine the performance of this updated version as our numerical simulation is only to second order accurate. In this study, we include the data driven model, which requires a large amount of training data. Hence, we optimised our simulation for the speed of simulation to explore a wider range of parameter space.

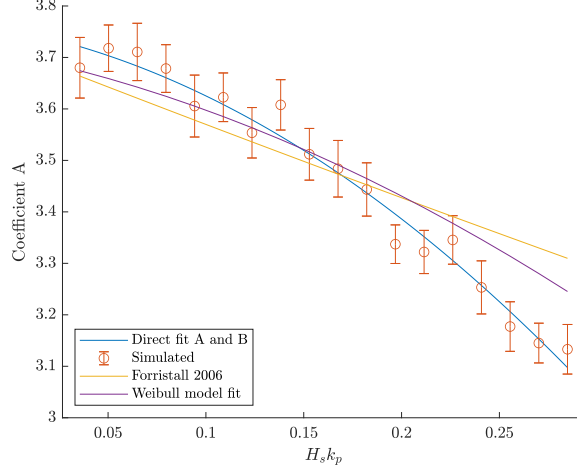


Figure 1: Comparison of the prediction curve from different models for  $A$  (equation 7) for different  $H_s$  with a heavy fitting weight towards the tail. The error bar gives 95% of confidence interval based on bootstrapping method. The duration of each simulation lasts for 150 periods. All the cases used in this figure is simulated in deep water. Each simulated point is obtained by fitting the maximum crest of 1000 realisations.

### 259 3.3. Empirical Fitting model

260 Followed by Piterbarg's theorem [24] and other theoretical models [20,  
 261 17, 49], the maximum crest distributions over an area can be described as  
 262 a Gumbel distribution with two parameters  $A$  and  $B$  shown in Equation 6.  
 263 Polynomial fitting  $f_{\text{fit}}(x)$  can be applied to predict the correlation between  
 264 these two parameters and input parameters (i.e. length in  $x$ , length in  $y$  etc.)  
 265 as:

$$\begin{aligned} A &= F_{\text{fit},A}(In_1, In_2 \dots) \\ B &= F_{\text{fit},B}(In_1, In_2 \dots), \end{aligned} \quad (27)$$

266 where  $In_1, In_2 \dots$  are input parameters. Herein we present a simple example  
 267 for illustration purposes. A polynomial fit is used to interpolate the  $A$  and  
 268  $B$  at desired input values (see Figure 1). The interpolated coefficients are  
 269 used to generate the curve labelled as "direct fit  $A$  and  $B$ ".

270 Alternatively, the value of  $\alpha_F$  and  $\beta_F$  can be obtained by fitting the crest  
 271 exceedance distributions, following Equation 3:

$$\begin{aligned}\alpha &= F_{\text{fit},\alpha}(In_1, In_2\dots) \\ \beta &= F_{\text{fit},\beta}(In_1, In_2\dots).\end{aligned}\tag{28}$$

272 A polynomial fit is used to interpolate the  $\alpha_F$  and  $\beta_F$  at specific input values  
 273 and the value of  $A$  and  $B$  are then computed following Equation 7 and 8.  
 274 The calculated coefficients are used to generate the curve labelled as “Weibull  
 275 model fit”.

276 However, the difficulty of obtaining a proper fit can be significantly in-  
 277 creased when the number of inputs is large. Hence, for second order problems,  
 278 we proposed a two step fitting method, which greatly reduces the difficulty  
 279 of fitting:

$$\begin{aligned}A &= F_{\text{fit,linear},A}(In_1, In_2\dots) + F_{\text{fit,2nd},A}(\mu), \\ B &= F_{\text{fit,linear},B}(In_1, In_2\dots) + F_{\text{fit,2nd},B}(\mu),\end{aligned}\tag{29}$$

280 where  $\mu$  is the wave steepness. In this simplified model, the extra crest  
 281 heights from the second order contributions are only fitted with the wave  
 282 steepness and hence reduced the number of data points required.

### 283 3.4. Random Forest model

284 The Random Forest model is an ensemble learning method that predicts  
 285 the output by constructing multiple decision trees (see Figure 2 as an ex-  
 286 ample). Each tree is built based on a different bootstrapped sample of the  
 287 original training dataset, which is referred to as bootstrap aggregating or  
 288 bagging. Additionally, at each leaf node, only a specific number of randomly  
 289 sampled features are selected as the candidates of the split. The best split  
 290 point can only be selected within these nominated features. This charac-  
 291 teristic is usually referred to as the random selection of features or feature  
 292 sampling. Both bagging and feature sampling are the core features of the  
 293 Random Forest models, which greatly reduce the correlations between dif-  
 294 ferent individual trees. Hence, the output value of each individual decision  
 295 tree can be treated as an independent prediction, and the average of these  
 296 independent predictions is the output as the predicted value of the Random  
 297 Forest model (see [50] for details).

298 To minimise the bias introduced during the random split of the training  
 299 and validation data, we applied the k-fold cross validation process [51] to

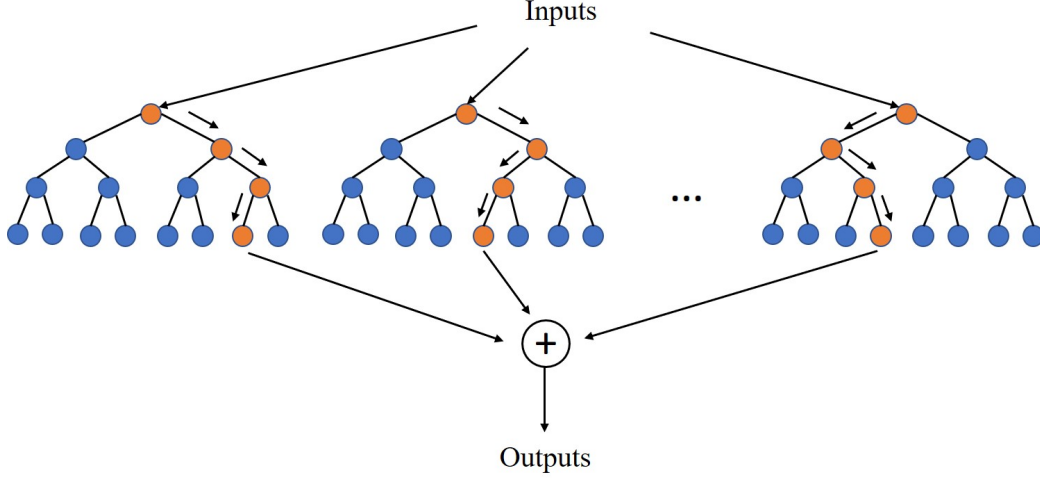


Figure 2: General structure of a Random Forest.

300 obtain more accurate predictions based on the Random Forest model. The  
 301 key idea of the k-fold cross validation process is to improve the validity and  
 302 accuracy of the model by splitting the dataset into subsets and evaluating  
 303 the performance of the model on each subset. For example, during a k-fold  
 304 cross validation process, the dataset is uniformly divided into  $q$  subsets. The  
 305 model is then trained, based on  $q - 1$  uniform-sized subsets. The trained  
 306 model is validated against the left-out subset as the out-of-bag samples. The  
 307 training and the validating process are repeated  $q$  times with a different  
 308 subset as the out-of-bag samples. The performance measurements for each  
 309 subset is calculated, and the average of them is treated as the final prediction  
 310 accuracy.

311 Apart from the splitting between training and validation data, hyper-  
 312 parameters also have a significant impact on the training quality of the Ran-  
 313 dom Forest model and hence will influence the accuracy of the predictions.  
 314 For the Random Forest model, there are two main hyper-parameters: the  
 315 number of trees in the forest and minimum required samples at a leaf node.  
 316 In this study, we tune all the hyper-parameters within the k-fold cross vali-  
 317 dation process. When building a Random Forest for each fold, we also search  
 318 for the optimal combinations of hyper-parameters. We use either grid search  
 319 as an uninformed method for simple models or Bayesian optimisation algo-  
 320 rithm [52] as informed tuning method for complicated methods. Although

Table 3: Summary of prediction models

Model Names	Model Type	Primary Equations	Input Type
Weibull model fit	Empirical Fitting	Equation 3 & 28	$P(\eta)$
TV model fit	Empirical Fitting	Equation 16 & 27	$P(\eta_{max})$
Direct fit $A$ and $B$	Empirical Fitting	Equation 6 & 27	$P(\eta_{max})$
Forristall2006	Theoretical	Equation 6	$P(\eta_{max})$
Random Forest	Data Driven	Equation 6 & 30	$P(\eta_{max})$

an informed method generally takes fewer iterations and may achieve better final tuning results [53], a grid search method has an advantage in parallel computing for simple models.

In this study, different Random Forest models are trained for different test conditions. Although the inputs for these models vary, the output from these models are the two coefficients shown in Equation 6. Hence, the Random Forest model can be simplified as:

$$[A, B] = F_{RF}(In_1, In_2...). \quad (30)$$

Based on these two predicted coefficients, the Equation 6 gives a probability of exceedance of the maximum crest height and the expected value of maximum crest height under given input conditions.

## 4. Results

### 4.1. Extreme second order wave statistics at a point

We start our investigation on the statistics from a point measurement as a limiting case to examine the performance of different models in a simplified test. We summarised the prediction models used in this section in Table 3.

#### 4.1.1. Point measurement models including steepness effects

In this section, we examine the performance of different models, when steepness varies from  $H_s k_p = 0.0358$  to  $H_s k_p = 0.25$ . When the wave steepness is the only parameter, the performance of different models at  $H_s k_p = 0.1788$  is shown in Figure 3, which is representative of other test cases with different wave steepness values. The model with the Tromans and Vanderschuren fit to the data and Random Forest model seems to predict the

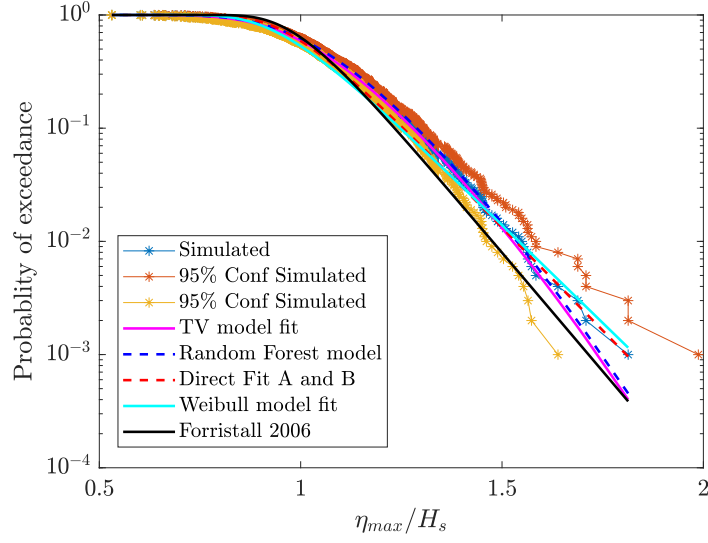


Figure 3: Comparison between simulated maximum crest distribution for test condition with steepness of  $H_s k_p = 0.1788$  against the predictions from different point measurement statistical models, which only consider wave steepness effects.  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

343 extreme crests distribution accurately except for a few extreme cases. For  
 344 all the models following Equation 6, the accuracy for the intermediate values  
 345 of maximum crest height are relatively poor. We also found that the direct  
 346 fit *A* and *B* model outperforms other models based on the Equation 6. The  
 347 general performance trend is similar at other wave steepness.

348 Apart from the exceedance plot, which focuses on the tail of the distri-  
 349 bution, the probability density function of the extreme crests distributions is  
 350 shown in Figure 4. This illustrates the performance around the probability  
 351 density peak. All Forristall models tends to under predict the probability of  
 352 relative small extreme values and over predict the probability of intermediate  
 353 and extreme values. Tromans and Vanderschuren fit to the data seems to  
 354 agree better with the simulated probability density function.

355 The weight function can also affect the fitting results and hence affects  
 356 the final performance of different models. Figure 5 shows the results with a  
 357 normal fitting weight. From the figure, the general performance of different  
 358 models around the peak of the probability density function is improved but  
 359 at the cost of the loss of accuracy towards the tail. As the expected value  
 360 of the maximum crest height has more significance in engineering practice,



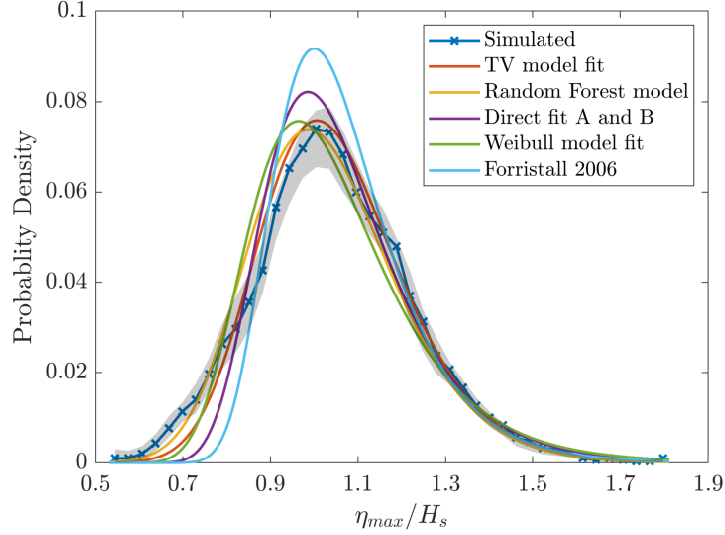


Figure 4: Comparison between simulated maximum crest distribution for test condition at steepness of  $H_s k_p = 0.1788$  against the predictions from different point measurement statistical models, which only consider wave steepness effects. The shaded area represents 95% confidence interval.  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

we used the normal fitting weight function to examine the performance of different space-time models.

#### 4.1.2. Point measurement models including water depth effects

In this subsection, we fix the wave steepness but vary the relative water depth from  $k_p d = 1.4$  to  $k_p d = 7$  to explore the impact of water depth on second order wave statistics.

Figure 6 compares the accuracy of different models at a fixed wave steepness of  $H_s k_p = 0.178$ , when the water depth varies. We choose the water depth of test dataset to be  $k_p d = 2.5$ , which represents well the model performance over the test domain. The Tromans and Vanderschuren fit to the data works best for the majority of the cases. The Random Forest model gives almost the same results as the polynomial fit. All of the models following equation 6 seem to over predict the probability at the probability peak.

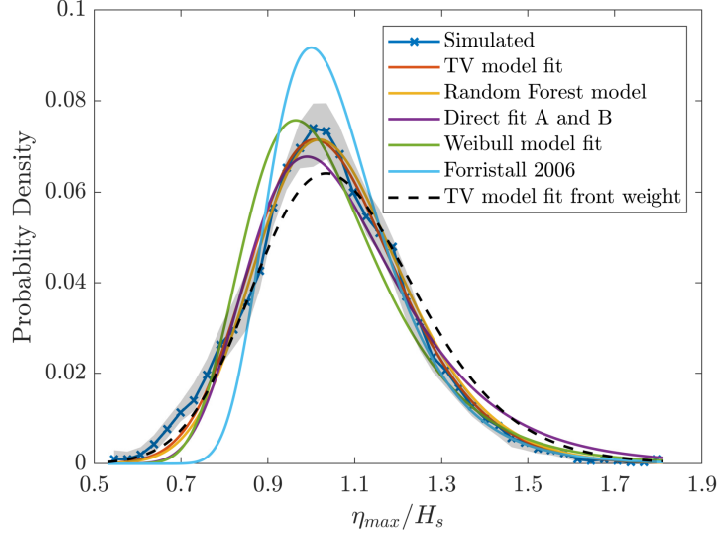


Figure 5: Comparison between simulated maximum crest distribution for test condition at steepness of  $H_s k_p = 0.1788$  against the predictions from different point measurement statistical models, which only consider wave steepness effects with a normal fitting weight during Gumbel fitting process (step 2 in section 2.4).  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

#### 375 4.1.3. Point measurement models including steepness and water depth co- 376 effects

377 We now analyse the performance of different models when water depth  
378 and wave steepness vary simultaneously. In this section, the relative water  
379 depth varies from  $k_p d = 1.4$  to  $k_p d = 7$  and the wave steepness varies from  
380  $H_s k_p = 0.040$  to  $H_s k_p = 0.25$  simultaneously.

381 For empirical fitting models, we used a total of 49 different combinations  
382 of input parameters as a grid. All the coefficients are averaged from 5 sim-  
383 ulations to reduce the statistical variability. This training dataset is then  
384 fitted with a third order polynomial to get the prediction curves (see Figure  
385 7 as an example).

386 A simple Random Forest model is used here to reduce the demands on the  
387 training datasets. We present a typical test case with relative water depth  
388 of  $k_p d = 2.5$  and wave steepness of  $H_s k_p = 0.178$ . From Figure 8, Random  
389 Forest seems to have an edge when comparing to traditional fitting methods  
390 when two inputs vary at the same time. All the theoretical models have  
391 worse performance when compared to the cases with only one variable, and

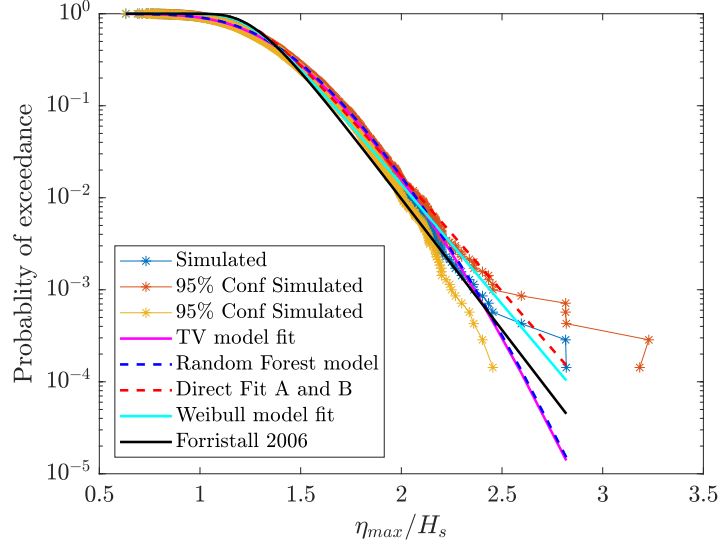


Figure 6: Comparison between simulated maximum crest distribution for test condition at water depth of  $k_p d = 2.5$  against the predictions from different point measurement statistical models, which only consider finite water depth effects.  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

also tend to underestimate the position of the probability peak.

We present the maximum crest statistics at a single point as a simplified example to investigate the feasibility of data driven methods. For a single input parameter from the sea state, all the models work well. However, when a more complicated problem is considered, the Random Forest model demonstrates advantages when compared to other methods. Although the Random Forest model can provide an accurate prediction in this particular example, we still consider this model as an early stage prototype instead of the state of the art.

#### 4.2. Linear waves over an area

Wave maximum crest distributions over an area differs from that measured at a single point particularly for relatively large areas. In this section, we examine maximum crests distributions of linear waves over an area. We start with the linear waves over a squared area, and the only variable is the side length of the area. We then further investigate the maximum crests height over a rectangular area, where the two side lengths vary simultaneously.

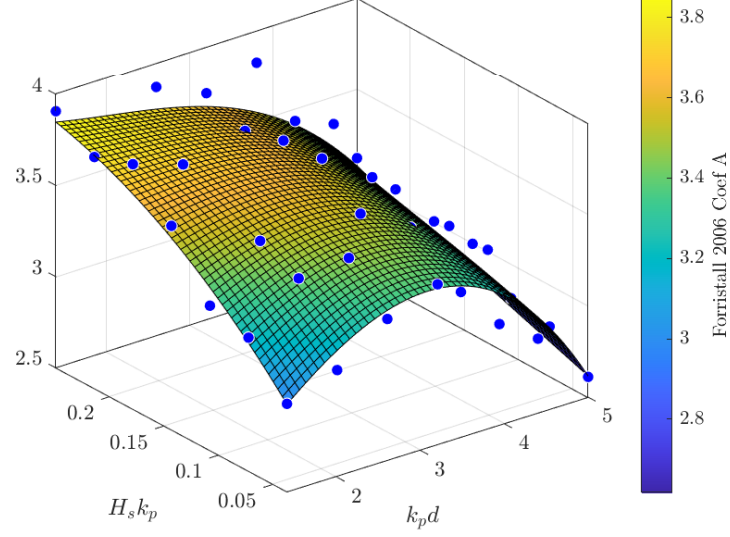


Figure 7: Prediction curve from Fit *A* and *B* model for coefficient *A* values for different significant wave heights and water depths.

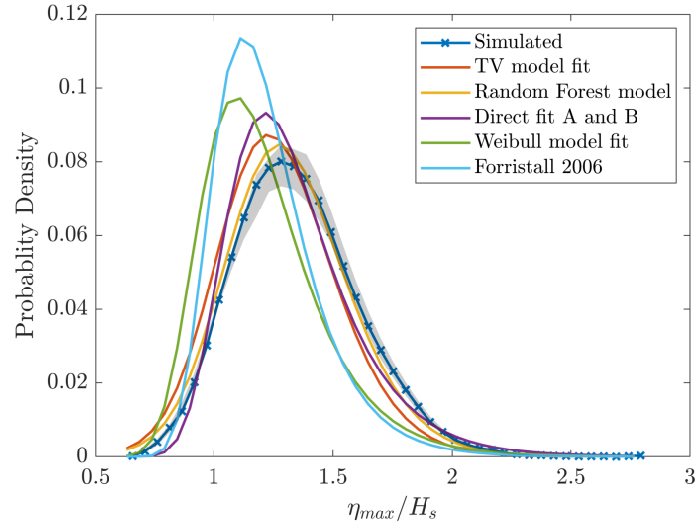


Figure 8: Comparison between simulated maximum crest distribution for test condition at water depth of  $k_p d = 2.5$  and wave steepness of  $H_s k_p = 0.178$  against the predictions from different point measurement statistical models, which consider both wave steepness and finite water depth effects.  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

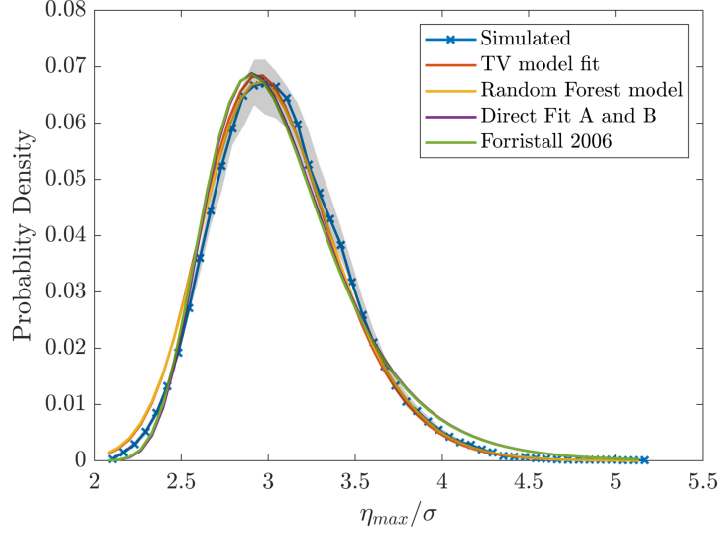


Figure 9: Comparison between linear simulated maximum crest distribution over a squared area with side length of  $3.5\lambda_p$  against the predictions from different linear space-time statistical models, which consider only the side length effects.  $\eta_{max}$  is the maximum crest elevation from a realisation of a simulation for 150 wave periods.

#### 4.2.1. Linear space-time models for a square area

In this section, we will first examine the performance of different models for a square area. We only change the length of the side from  $0.5\lambda_p$  to  $20\lambda_p$ . The prediction curve of the fitting method is based on a 25 point polynomial fit. Simulations with random side lengths are used for Random Forest training to obtain predictions of coefficients of the Gumbel distribution.

To examine the performance of different models, we have presented a test case with a side length of  $3.5\lambda_p$ . We have chosen this test case as it represents the general trends well for other test cases within the test domain. Figure 9 shows the predictions of different models when compared to the test case. The shade shows the 95% confidence interval. In general, all the theoretical, fitting and Random Forest models predict the probability density function well. Random Forest model tends to over predict at the lower end. Both Forristall2006 model and Fedele2012 model tend to slightly over predict at the higher end.

#### 4.2.2. Linear space-time models for a rectangular area with side lengths effects

We further examine the performance of different models when the area is rectangular, and both the length in  $x$  and the length in  $y$  varies. Figure 10 shows the percentage error of expected maximum crest height predicted from different models at different combinations of lengths in  $x$  and  $y$  directions.

From Figure 10, both theoretical models can predict the maximum crest height over a near square sized area accurately. However, when side length in  $x$  direction is small compare to the length in  $y$ , both models tend to over predict the extreme crest height, particularly for the Forristall2006 model. For small lengths in  $y$  direction, both theoretical models tend to under predict the maximum crest height. However, both fitting model and Random Forest model tend to provide accurate predictions for areas with extreme aspect ratios. Random Forest model slightly outperforms the fitting model with less random errors in the middle of the test domain.

#### 4.3. Second order waves over an area

Waves in open oceans are modified by nonlinear physics. This will make the crests higher than those predicted by linear theory, particularly in steep sea states. In this section, we simulate waves with second order corrections to further examine the accuracy of different models. We first look at rectangular areas with fixed wave steepness in 4.3.1, and we further extend this with varying wave steepness in 4.3.2. For both cases examined in this subsection only deep water waves are considered.

For consistency we match the order of theoretical model with the order of accuracy in the numerical simulations. In this section, as the second order waves being simulated, we applied the second order version for both Forristall2006 and Fedele2012 model.

##### 4.3.1. Second order space-time models for a rectangular area with side lengths effects

We examine the performance of different models for second order waves over a rectangular area in deep water. In this subsection, the wave steepness is fixed at  $H_s k_p = 0.178$ . The relative error in the expected value of maximum crests height for second order waves at different combination of length in  $x$  and length in  $y$  is shown in Figure 11. When compared to the performance in linear simulations in Figure 10, the relative error of two theoretical models is significantly increased in second order simulations. For Forristall2006 model,

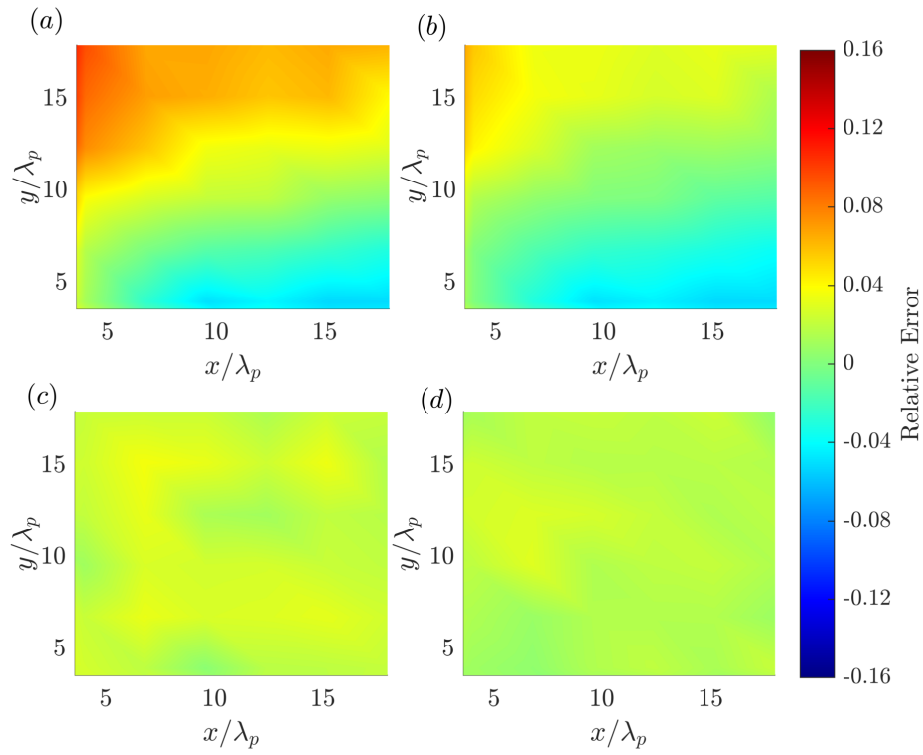


Figure 10: Comparison of the relative error in the expected value of maximum linear wave crest value for 150 wave periods from (a): Forristall 2006, (b) Fedele 2012, (c) Fit *A* and *B* and (d) Random Forest model at different area sizes for linear simulation.

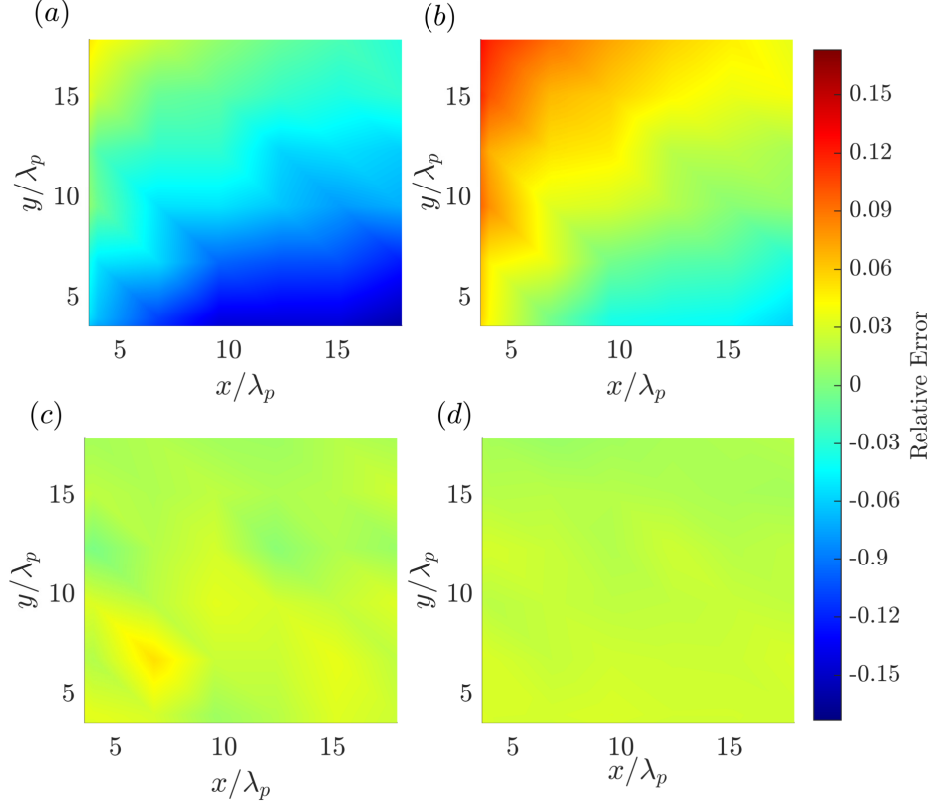


Figure 11: Comparison of the relative error in the expected value of maximum crest value for 150 wave periods from (a): Forristall 2006, (b) Fedele 2012, (c) Fit *A* and *B* and (d) Random Forest model at different area sizes for second order simulation at fixed wave steepness.

460 inaccurate predictions of the extra crest height due to second order harmonics  
 461 leads to a significant bias, which also leads to large errors (over 14%) at large  
 462 length in  $x$ . Fedele2012 model maintain its performance when the aspect  
 463 ratio is close to 1. However, the relative error is increased for extreme aspect  
 464 ratio cases. The increase in relative error is likely due to the magnification  
 465 effect from the extra crest height in second order wave field. The fitting  
 466 method and Random Forest method maintains their high performance as  
 467 these two models can easily include the extra crest height from second order  
 468 harmonics at a fixed wave steepness.



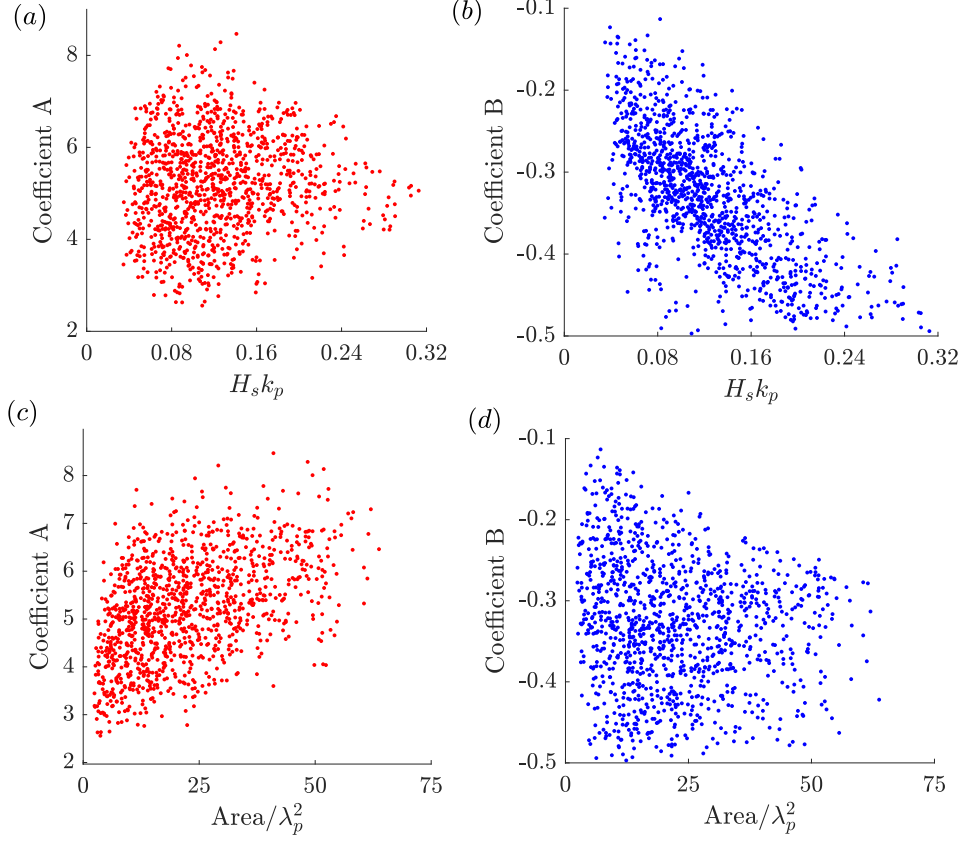


Figure 12: Scatter plot of coefficients  $A$  and  $B$  against (a), (b): wave steepness, (c), (d) area size.

#### 4.3.2. Second order space-time models for a rectangular area with steepness and side lengths co-effects

We first examine the impact of both area size and the wave steepness on the two coefficients of the Gumbel distribution. In Figure 12, we present the scattering of two coefficients  $A$  and  $B$ , when we change the area size and the wave steepness simultaneously. From Figure 12 (c), coefficient  $A$  seems to have a strong correlation with the area size, which indicates the area size will affect mostly on the shape of the curve. Coefficient  $B$  shows strong correlation with the wave steepness as the wave steepness mainly has an impact on the mean value of maximum crest distribution.

Finally, the performance of all the models will be examined by predicting

the maximum crest height over a rectangular area with varying wave steepness in deep water. In Figure 13, we present the relative error at different cross planes of the total test domain. For both theoretical models, the horizontal cross section at  $H_s k_p = 0.178$  is the same as the error plot shown in Figure 11. The results at the lowest wave steepness  $H_s k_p = 0.089$  is similar to the error plot shown in Figure 10 with linear simulation. This similarity is primarily because the second effect is not significant in low steepness sea states. However, for high steepness sea states, both theoretical models tend to give large errors, especially for Forristall2006 model, which underestimates the maximum crest height. This increasing relative error from low steepness to high steepness for the Forristall2006 model indicates there is a steepness correlated offset, which is most likely due to the underestimation of extra crest height from the second order corrections. However, the Fedele2012 model generally provides a reasonably accurate estimation of second order effects for a near squared area even at high steepness. The increased relative error for high steepness cases is likely due to the amplification effect from the extra crest height in second order effects.

The fitting method tends to overestimate the maximum crest height, especially for areas with a short length in  $x$  direction. This overestimation is probably due to the large change rate in both coefficients  $A$  and  $B$  in that region. The third order polynomial fitting method seems to struggle in this fast-changing region.

The Random Forest model gives the best overall performance when compared to the other three models. The relative error is comparatively small in the middle of the training domain, but the Random Forest model tends to give a large error at the boundary of the training domain, which shows at the lowest steepness with a short length in  $x$ . This is primarily because the nature of the Random Forest model as there is generally less training data available at the boundary of the training domain. The same trend applies to the relatively large error at high wave steepness with short lengths in  $y$ .

#### 4.4. Space-time model performance comparison

We summarise the performance of different models under different wave conditions as the averaged absolute relative error and relative error range. The former parameter provides a measure of the general performance of the model, and the latter parameter shows the worst performance of the model within the test range.

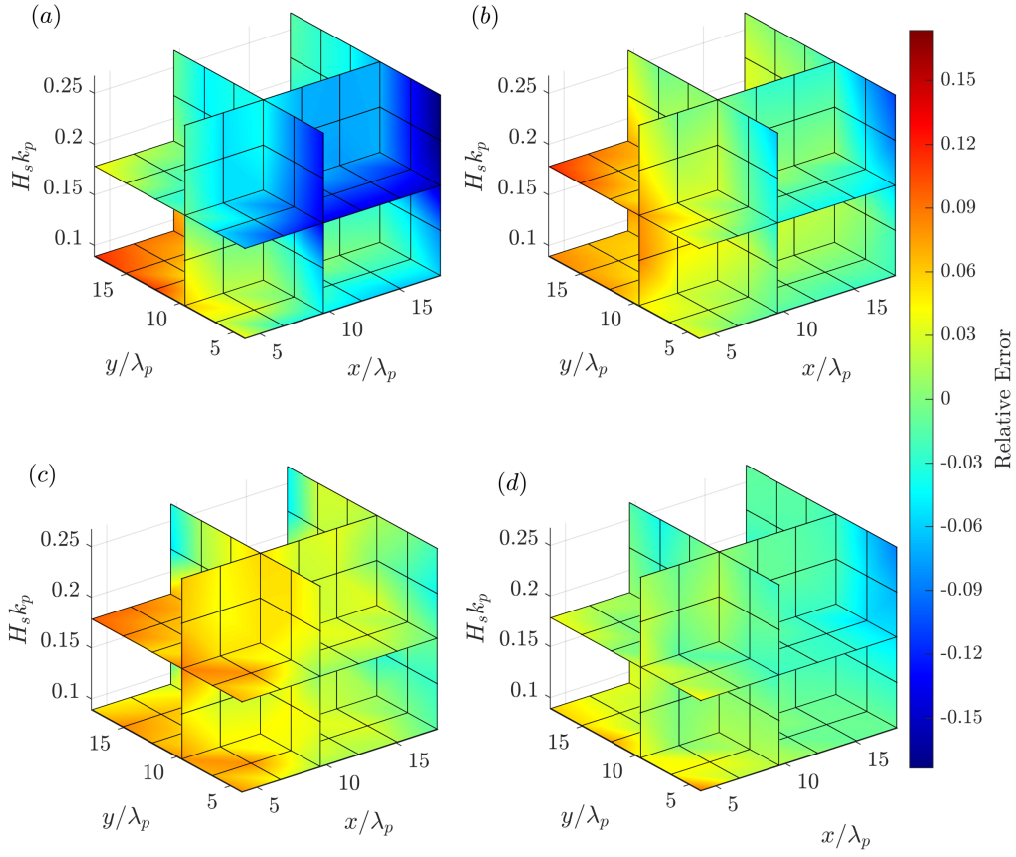


Figure 13: Comparison of the relative error in the expected value of maximum crest value for 150 wave periods from (a): Forristall 2006, (b) Fedele 2012, (c) Fit A and B and (d) Random Forest model at different area sizes for second order simulations with different wave steepnesses and area side lengths.

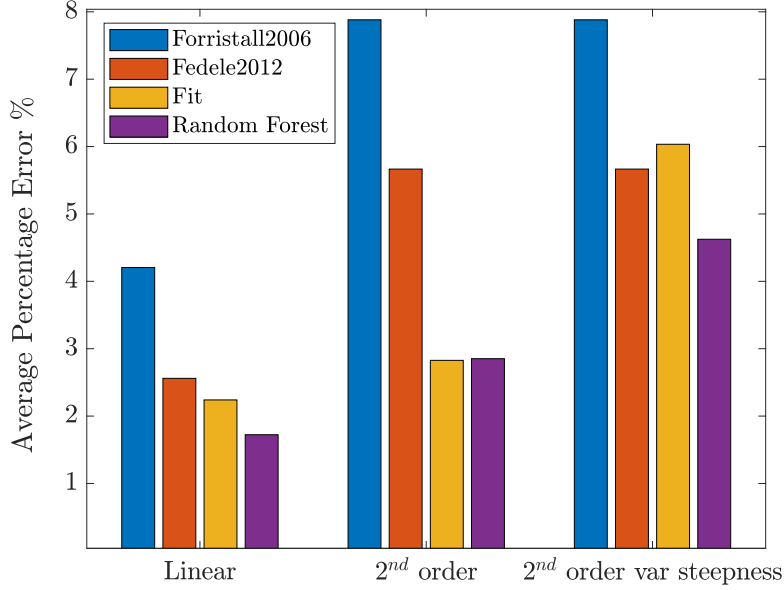


Figure 14: Comparison of the absolute average relative error in the expected value of maximum crest value from different models.

Figure 14 shows the averaged absolute relative error for all the models with linear waves, second order waves with fixed wave steepness and second order waves with varying wave steepness. In general, for linear waves, all the models give lower overall errors, and the Random Forest model has the best performance. For second order waves with fixed wave steepness, both fitting methods and the Random Forest model provide relatively accurate predictions on the maximum crest height. When wave steepness is introduced as an additional parameter, the fitting method requires more training data and hence leads to increased relative error. The Random Forest model still has the best performance among all the models tested herein.

Figure 15 shows the maximum error of different models in the test domain. For second order waves with varying wave steepness cases, we examine the error range at the wave steepness of  $H_s k_p = 0.178$ . In general, the fitting method and the Random Forest model tend to provide relatively small error ranges for both linear and second order waves, and all models generally produce less error for linear cases. For second order waves, the error ranges of both theoretical models increase significantly. Both fitting method and the Random Forest model tend to slightly over predict the results for the second

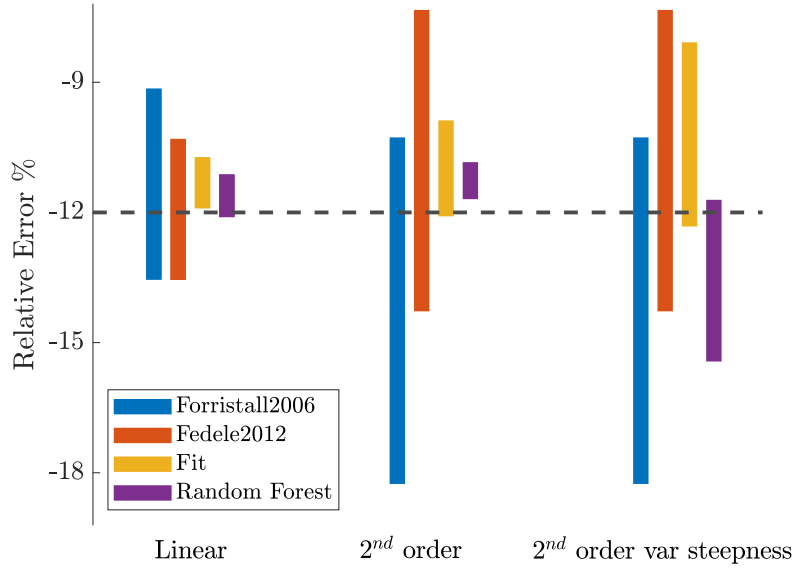


Figure 15: Comparison of the maximum error range in the expected value of maximum crest value from different models.

order waves with fixed wave steepness. When wave steepness is introduced as an additional parameter, the fitting method and Random Forest models tends to have larger error range, and the fitting method continues to over predict the maximum crest height.

We apply the data driven methods in space-time wave statistics as the second example in this paper. For relatively simple cases with few input parameters, such as linear waves over a square area, all the theoretical, fitting and the Random Forest model works well. However, as the situation becomes more complicated, additional input parameter increases the complexity of the prediction models. Prediction errors increase for all the models, but the Random Forest model starts to show its potential in handling complicated problems.

#### 4.5. Importance of the parameters

Apart from the providing predictions on the maximum crest distributions, the Random Forest model can also provide an advanced variable importance measure as ‘permutation accuracy importance’. This measurement indicates the relative importance of different input parameters based on the prediction

551 trees. This importance estimate parameter can help both in eliminating the  
 552 irrelevant inputs to increase the training speed and accuracy of the model,  
 553 and also can provide a brief insight into the nature of maximum crest statis-  
 554 tics.

555 Figure 16 shows the relative importance parameter when the test matrix  
 556 is divided into four subgroups based on the wave steepness and area size.  
 557 The general trend of the importance for all of the inputs is similar for both  
 558 high and low steepness models. Wave steepness become more critical with  
 559 high steepness cases as the second order effect tends to have more impact  
 560 in a steep sea state. This double confirmation further demonstrates that  
 561 these importance estimates from the Random Forest model can provide extra  
 562 insights into the nature of the maximum crest distributions.

563 The relative importance of different inputs changes dramatically when  
 564 the area size changes. For small areas, the scale dimension coefficient  $\beta$   
 565 seems to be much more important when comparing to large area cases. This  
 566 is primarily due to the scale dimension coefficient  $\beta$  which quantifies the  
 567 statistical change when a 3D wave reduces to the 2D or even 1D wave at  
 568 small areas. The importance of wave steepness increases for large areas,  
 569 which could be because for relatively large areas, increase area size is not as  
 570 effective as increase wave steepness. The importance of length in  $y$  direction is  
 571 also increased for large areas, which could be because length in the  $y$  direction  
 572 is more effective for 3D waves. When an area is small, the difference in the  
 573 peak wavelength in  $x$  and  $y$  direction will lead to differences in the number  
 574 of waves in 2D waves, which agrees well with Fedele’s theory [20].

## 575 5. Discussions and conclusions

576 This paper reviewed existing theoretical models for maximum crest dis-  
 577 tributions over an area and also proposed two data driven models to estimate  
 578 the crest distributions based on the sea state parameters without any simu-  
 579 lations required during the prediction phase. We use second order numerical  
 580 simulations to study the accuracy of both the theoretical models and data  
 581 driven models under various conditions.

582 In this study, we present two examples of applying data driven methods  
 583 to predict wave statistics. In the first example, we explore the second order  
 584 maximum crest distributions at a single point, where the Random Forest  
 585 model shows its advantages in accuracy for the most complicated scenario.

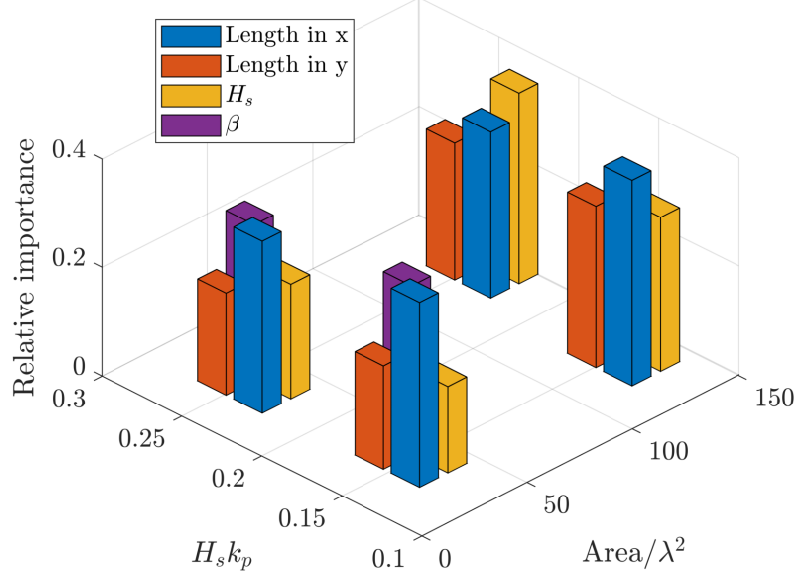


Figure 16: Relative importance estimated from the Random Forest model based on the input range of area size and wave steepness for waves with second order correction.

586 In the second example, we extend our research into the maximum crest dis-  
587 tributions over an area. Of two theoretical models, the Fedele2012 model  
588 [20] tends to provide more accurate predictions, particularly for second order  
589 waves. The polynomial fitting method works equally well with the Random  
590 Forest model for the linear cases. The error in the empirical fitting method  
591 increased significantly for the second order case, which is primarily because  
592 of the fitting difficulty in a complex problem with many input parameters.  
593 However, the Random Forest model continues to perform well in the most  
594 complicated case and thus shows significant potential for modelling complex  
595 situations.

596 The two examples studied in this paper are intended as practicability tests  
597 of applying data driven method to predict maximum crest distributions under  
598 various conditions. The complexity of the Random Forest model presented  
599 herein is restricted by the number of inputs, which ultimately depends on the  
600 time constraint and accessible computational power. The underlying physics  
601 for these two examples examined herein are relatively straightforward with  
602 proper assumptions. Hence, two theoretical models can provide relatively

603 accurate predictions particularly for the areas close to a square.

604 However, real water wave statistics in the open ocean are modified by  
605 weak and strong nonlinear physics. For complicated situations such as wave-  
606 current interactions, wind wave interactions, wave shoaling effects, non-  
607 equilibrium spectrum or wave breaking effects, it is very difficult to derive  
608 satisfactory analytical solutions to account for all these phenomena. This  
609 makes design difficult.

610 For relatively complicated problems, the traditional polynomial fitting  
611 methods requires large amount of fitting data yet with large errors for mul-  
612 tiple inputs cases. However, the Random Forest model requires less training  
613 data and provides more accurate predictions even for two simplified situ-  
614 ations examined herein. For engineering applications, the Random Forest  
615 model can be used just as the traditional fitting equations. Given the input  
616 from the sea state parameters (i.e. steepness, sampling area dimensions), the  
617 Random Forest model can make predictions without requiring much compu-  
618 tational power once it is properly trained and validated. However, care needs  
619 to be taken when applying these data driven methods directly as random er-  
620 ror peak could occur especially at the boundary of the training domain.

621 Based on the performance of two examples examined, data driven ap-  
622 proaches look promising in terms of predicting space-time wave statistics.  
623 This type of statistical model has substantial potential in predicting maxi-  
624 mum crest height distributions under intricate situations, which can also be  
625 used for engineering purposes with proper validation.

## 626 **Acknowledgement**

627 This work was funded by UK/China ORE funding (EPSRC/NERC/NSFC  
628 EP/R007632/1).

## 629 **Appendix A. Envelope based second order corrections**

630 For the second order corrections, an envelope method is used to compute  
631 the sum and difference terms:

$$\eta = \Re(\eta_{linear} + \eta_{2-} + \eta_{2+}), \quad (\text{A.1})$$

where  $\Re$  is the real part of the signal,  $\eta_{2-}$  is the second order difference terms  
and  $\eta_{2+}$  is the second order sum terms For the linear part  $\eta_{linear}$ , we follow



the standard envelope equation:

$$\eta_{linear} = U \exp(i(k_p x - \omega_p t)), \quad (\text{A.2})$$

where  $k_p$  is the dominant wave number,  $U$  is the linear complex envelope, and  $\omega_p$  is the peak wave frequency.

For second order sum terms with bandwidth correction can be expressed as:

$$\eta_{2+} = \left[ \text{Coef0} \frac{kU^2}{2} - \text{Coef1} iU \frac{\partial U}{\partial x} + \text{Coef2} U \frac{\partial^2 U}{\partial y^2} + \text{Coef3} \frac{1}{k_p} \left( \frac{\partial U}{\partial y} \right)^2 \right] \exp(2i(k_p x - \omega t)), \quad (\text{A.3})$$

where Coef0 is the leading order coefficient, which is given in [43] for finite water depth. For the bandwidth correction coefficients Coef1, Coef1, Coef3, Trulsen and Dysthe [54] and Toffoli *et al.* [5] give the values in the deep water:

$$\text{Coef1} = -\frac{1}{2}, \quad \text{Coef2} = \frac{1}{2}, \quad \text{and} \quad \text{Coef3} = -\frac{3}{4}. \quad (\text{A.4})$$

To obtain the bandwidth correction coefficients at finite water depth, we used a global optimisation code to find the best combination of all three bandwidth correction coefficients. The target of the optimisation code is the exact solution presented in [44]. This process is repeated at different water depths to obtain the bandwidth correction coefficient values at different the water depths.

The obtained bandwidth correction coefficients at different relative water depths is shown in Figure A.17. All the bandwidth correction coefficients converges to the deep water value shown in Equations A.4 when the relative water depth tends to infinity. This suggests that the global optimisation method seems to successfully find the optimal combinations to provide the best bandwidth corrections for the second order sum term.

For the second order difference term, we followed Trulsen and Dysthe [54] and Toffoli *et al.* [5] for deep water cases:

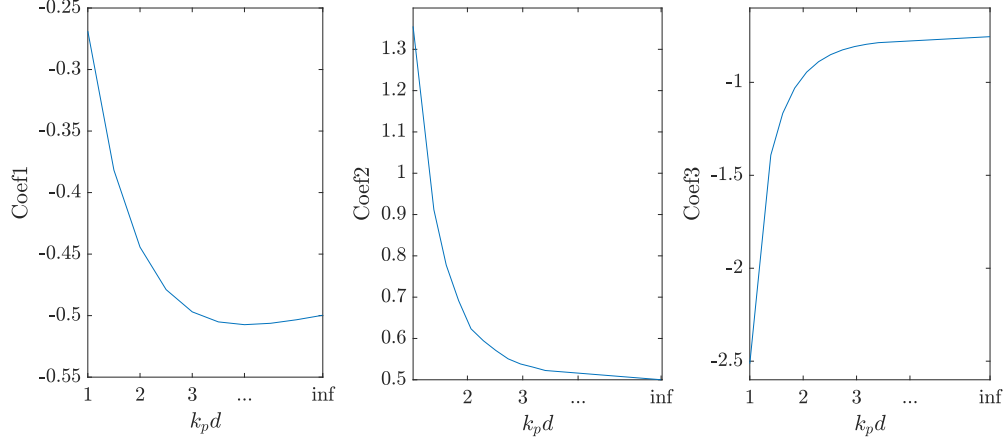


Figure A.17: Bandwidth correction coefficients at different relative water depths.

$$\eta_{2-} = \frac{1}{2\omega_p} \frac{\partial \phi}{\partial x} - \frac{1}{16k_p} \frac{\partial^2 |U|^2}{\partial x^2} - \frac{1}{8k_p} \frac{\partial^2 |U|^2}{\partial y^2}, \quad (\text{A.5})$$

where  $\phi$  is a velocity potential of the return flow at mean water level, which can be obtained as:

$$\left. \frac{\partial \phi}{\partial v} \right|_{v=0} = \frac{\omega_p}{2} \frac{\partial |U|^2}{\partial x}, \quad (\text{A.6})$$

and the potential satisfies Laplace's equation in the fluid, thus

$$\nabla^2 \phi = 0, \quad (\text{A.7})$$

where  $v$  is the position in vertical direction. For the second order difference term in finite water depth, since we are only interested at the peak of the largest crest, we followed the an simplified equation used by McAllister *et. al.* [55] as:

$$\eta_{2-,c} = -\frac{|U|^2}{4d} \frac{1}{1+R}, \quad (\text{A.8})$$

where  $\eta_{2-,c}$  is the difference term at the envelope peak,  $d$  is the water depth, and  $R$  is the spatial aspect ratio of the wave group (see [55] for more details of this formula and its derivation). We examine the accuracy of the

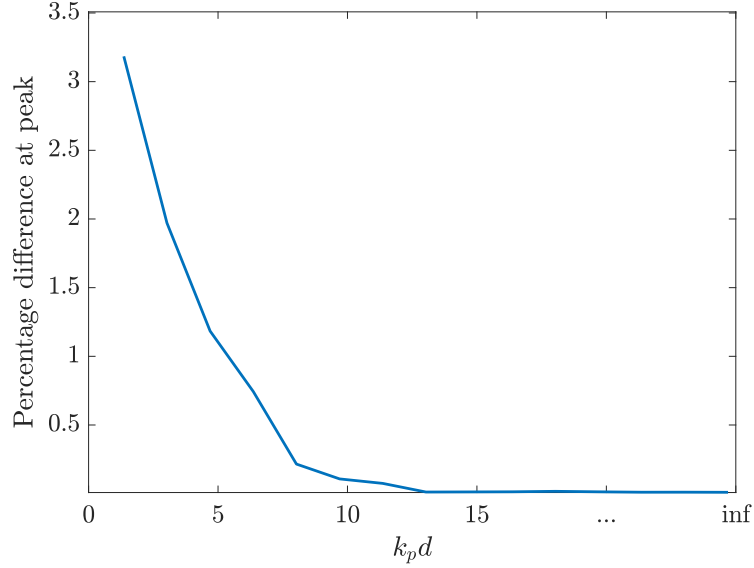


Figure A.18: Percentage difference of second order difference term at different water depths.

664 difference term by comparing the results against the exact solution presented  
665 in [44] at different water depths. Figure A.18 presents the percentage differ-  
666 ence at different water depth. There is about 3% of difference in terms of  
667 the trough of the difference term at the shallow water end and everything is  
668 accurate on the deep water side.

## 669 Appendix B. Nomenclature and symbol clarification

670 In this study, we have used three different parameters for measuring wave  
671 steepness, so as to be consistent with the studies these equations are taken  
672 from, we clarify the difference between these parameters as:

- 673 1.  $\mu$  (Equation 24) is a measure of the wave steepness, which is calculated  
674 based on the directional wave spectrum.
- 675 2.  $S_1$  (Equation 5) can be calculated directly from the time series as  $T_z$   
676 can be obtained with zero-crossing points and  $H_s$  can be approximated  
677 as  $4\sigma$ .
- 678 3.  $H_s k_p$  (Table 1) in this study is computed from the input linear wave  
679 spectrum.

680 We also used three parameters to characterising the wave length:

- 681 1.  $\lambda_p$  (Table 1) is computed from the directly with  $T_p$ , which is a fixed  
682 value of 1.5 seconds in this study through dispersion relationship.
- 683 2.  $\lambda_x$  and  $\lambda_y$  (Equation 10) can be calculated based on the directional  
684 wave spectrum.

685 We also used three parameters to characterising the wave period:

- 686 1.  $T_z$  can be obtained with zero-crossing points from surface elevation  
687 records.
- 688 2.  $T_m$  (Equation 20) can be calculated based on the simulated directional  
689 wave spectrum.
- 690 3.  $T_p$  is a fixed value of 1.5 seconds in this study.

691 Symbol  $T$  without subscript is for the duration of the time series which is  
692 150 periods in this study.

Table B.4: Nomenclature

Symbol	Definition
$\eta_{linear}$	Linear part of surface elevation
$\mathbf{x}$	Position Vector
$\mathbf{k}_i$	Wavenumber vector for $i^{th}$ component
$\omega_i$	Angular frequency for $i^{th}$ component
$t$	Time of the surface elevation
$a_i$	Independent random variables drawn from an normal distribution
$b_i$	Independent random variables drawn from an normal distribution
$C_i^2$	Variance of $i^{th}$ wave component
$\mathbb{N}$	A normal distribution
$S(f, \theta)$	Directional wave spectrum
$S_o(f)$	Omnidirectional wave spectrum
$H_s$	Significant wave height
$k_p$	Peak wavenumber
$H_s k_p$	One of the wave steepness parameters
$x$	Side length of sampling area along mean wave direction
$y$	Side length of sampling area in transverse direction
$k_p d$	Normalised water depth
$d$	Water depth
$\lambda_p$	Peak wave length

Table B.4: Nomenclature

Symbol	Definition
$\gamma$	Peak enhancement factor
$D$	Spreading function
$G_\theta$	Directional spreading parameter
$\theta$	Angle deviated from the mean wave direction
$P$	Probability distribution
$\eta$	Surface elevation
$\alpha_F$	Coefficient for Forristall distribution
$\beta_F$	Coefficient for Forristall distribution
$S_1$	Steepness of the sea-state
$g$	Gravitational acceleration
$T_z$	Zero-crossing period
$T_m$	Mean wave period
$T_p$	Peak wave period
$\eta_{max}$	Space-time maximum surface elevation
$\sigma$	Standard deviation of the surface elevation
$s$	Quantity of interest
$N$	Equivalent number of waves
$A$	Parameter of Gumbel distribution
$B$	Parameter of Gumbel distribution
$T$	Duration of the time series
$\lambda_x$	Averaged wave length in the mean wave direction for simulated wave field
$\lambda_y$	Averaged wave length in the lateral direction for simulated wave field
$E$	Expected value of probability density
$L$	Side length of the squared area
$N_p$	Parameter fitted for Tromans and Vanderschuren model
$b_p$	Parameter fitted for Tromans and Vanderschuren model
$z$	Threshold parameter for Fedele2012 model
$N_V$	Average number of waves within the volume
$N_S$	Average number of waves on the boundary surfaces
$N_B$	Average number of waves along the perimeter
$h$	Modal value for Fedele2012 model
$\mu$	A measure of wave steepness corrected with bandwidth
$\mu_m$	A measure of wave steepness
$\nu$	Wave bandwidth
$q$	Number of subsets

Table B.4: Nomenclature

Symbol	Definition
$\Re$	Real part of complex number
$\eta_{2-}$	Second order difference terms
$\eta_{2+}$	Second order sum terms
$\omega_p$	Peak wave frequency
$U$	Linear complex envelope
Coef1-4	Coefficients for second order sum terms
$\phi$	Velocity potential of the return flow at mean water level
$R$	Spatial aspect ratio of the wave group
$\eta_{2-,c}$	Difference term at the envelope peak
$v$	Position in vertical direction
$f$	Frequency bins of the wave spectrum
$F_1(L/\lambda_p)$	Fitting function for Forristall2015
$\xi$	A intermediate variable for calculating $\beta$
$m_{i,j,l}$	Moments of the directional wave spectrum
$\alpha_{xy}, \alpha_{xt}, \alpha_{yt}$	Irregularity parameters of the sea state

## References

- [1] F. Fedele, J. Brennan, S. P. De León, J. Dudley, F. Dias, Real world ocean rogue waves explained without the modulational instability, Scientific Reports 6 (2016) 27715.
- [2] F. Fedele, A. Benetazzo, G. Gallego, P. C. Shih, A. Yezzi, F. Barbariol, F. Ardhuin, Space-time measurements of oceanic sea states., Ocean Modelling 70 (2013) 103–115.
- [3] F. Fedele, On the kurtosis of deep-water gravity waves, Journal of Fluid Mechanics 782 (2015) 25–36.
- [4] W. Xiao, Y. Liu, G. Wu, D. K. P. Yue, Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution, Journal of Fluid Mechanics 720 (2013) 357–392.
- [5] A. Toffoli, O. Gramstad, K. Trulsen, J. Monbaliu, E. Bitner-Gregersen, M. Onorato, Evolution of weakly nonlinear random directional waves: laboratory experiments and numerical simulations, Journal of Fluid Mechanics 664 (2010) 313.

- 709 [6] M. Onorato, A. R. Osborne, M. Serio, L. Cavaleri, C. Brandini, C. T.  
710 Stansberg, Extreme waves, modulational instability and second order  
711 theory: wave flume experiments on irregular waves, *European Journal*  
712 *of Mechanics - B/Fluids* 25 (5) (2006) 586–601.
- 713 [7] S. Y. Annenkov, V. I. Shrira, Evolution of kurtosis for wind waves,  
714 *Geophysical Research Letters* 36 (13) (2009).
- 715 [8] M. Christou, K. Ewans, Field measurements of rogue water waves, *Jour-*  
716 *nal of Physical Oceanography* 44 (9) (2014) 2317–2335.
- 717 [9] G. Z. Forristall, Wave Crest Distributions: Observations and Second-  
718 Order Theory, *Journal of Physical Oceanography* 30 (8) (2000) 1931–  
719 1943.
- 720 [10] F. Fedele, M. A. Tayfun, On nonlinear wave groups and crest statistics,  
721 *Journal of Fluid Mechanics* 620 (2009) 221.
- 722 [11] M. Onorato, A. R. Osborne, M. Serio, L. Cavaleri, C. Brandini, C. T.  
723 Stansberg, Observation of strongly non-Gaussian statistics for random  
724 sea surface gravity waves in wave flume experiments, *Physical Review*  
725 *E* 70 (6) (2004) 067302.
- 726 [12] Z. Cherneva, M. A. Tayfun, C. Guedes Soares, Statistics of nonlinear  
727 waves generated in an offshore wave basin, *Journal of Geophysical Re-*  
728 *search: Oceans* 114 (C8) (2009).
- 729 [13] G. Z. Forristall, Comparing hindcasts with wave measurements from  
730 hurricanes lili, ivan, katrina and rita, in: *Proc. 10th International Work-*  
731 *shop on Wave Hindcasting and Forecasting and Coastal Hazards Sym-*  
732 *posium*, North Shore, Oahu, HI, Nov, 2007, pp. 11–16.
- 733 [14] P. A. E. M. Janssen, Nonlinear four-wave interactions and freak waves,  
734 *J. Phys. Ocean.* 33 (4) (2003) 863–884.
- 735 [15] M. A. Tayfun, F. Fedele, Wave-height distributions and nonlinear effects,  
736 *Ocean Engineering* 34 (11-12) (2007) 1631–1649.
- 737 [16] F. Fedele, C. Lugni, A. Chawla, The sinking of the el faro: predicting  
738 real world rogue waves during hurricane joaquin, *Scientific reports* 7 (1)  
739 (2017) 1–15.

- [17] G. Z. Forristall, Maximum crest heights over an area and the air gap problem, International Conference on Offshore Mechanics and Arctic Engineering 47489 (2006) 11–15.
- [18] K. Dysthe, H. E. Krogstad, P. Müller, Oceanic rogue waves, Annual Review of Fluid Mechanics 40 (2008) 287–310.
- [19] F. Fedele, A. Benetazzo, G. Gallego, P.-C. Shih, A. Yezzi, F. Barbariol, F. Ardhuin, Space–time measurements of oceanic sea states, Ocean Modelling 70 (2013) 103–115.
- [20] F. Fedele, Space–time extremes in short-crested storm seas, Journal of Physical Oceanography 42 (9) (2012) 1601–1615.
- [21] G. Z. Forristall, Maximum crest heights under a model tlp deck, in: International Conference on Offshore Mechanics and Arctic Engineering, Vol. 44342, 2011, pp. 571–577.
- [22] A. Benetazzo, F. Barbariol, F. Bergamasco, A. Torsello, S. Carniel, M. Sclavo, Observation of Extreme Sea Waves in a Space-time Ensemble, Journal of Physical Oceanography 45 (9) (2015) 2261–2275.
- [23] R. J. Adler, The Geometry of Random Fields, Vol. 62, SIAM, 1981.
- [24] V. I. Piterbarg, Asymptotic methods in the theory of Gaussian processes and fields, Vol. 148, American Mathematical Society, 1996.
- [25] H. Socquet-Juglard, K. Dysthe, K. Trulsen, H. E. Krogstad, J. Liu, Probability distributions of surface gravity waves during spectral changes (2005).
- [26] J. E. Adler, R. J. Taylor, Random fields and geometry (2007).
- [27] F. Fedele, G. Gallego, A. Yezzi, A. Benetazzo, L. Cavaleri, M. Sclavo, M. Bastianini, Euler characteristics of oceanic sea states, Mathematics and Computers in Simulation 82 (6) (2012) 1102–1111.
- [28] A. Benetazzo, F. Fedele, G. Gallego, P.-C. Shih, A. Yezzi, Offshore stereo measurements of gravity waves, Coastal Engineering 64 (2012) 127–138.



- 768 [29] G. Z. Forristall, Maximum crest heights over an area: laboratory mea-  
769 surements compared to theory, in: International Conference on Offshore  
770 Mechanics and Arctic Engineering, Vol. 56499, American Society of Me-  
771 chanical Engineers, 2015, p. V003T02A044.
- 772 [30] M. Ali, R. Prasad, Significant wave height forecasting via an extreme  
773 learning machine model integrated with improved complete ensemble  
774 empirical mode decomposition, *Renewable and Sustainable Energy Re-  
775 views* 104 (2019) 281–295.
- 776 [31] A. Callens, D. Morichon, S. Abadie, M. Delpey, B. Liquey, Using random  
777 forest and gradient boosting trees to improve wave forecast at a specific  
778 location, *Applied Ocean Research* 104 (2020) 102339.
- 779 [32] S. L. Brunton, B. R. Noack, P. Koumoutsakos, Machine learning for  
780 fluid mechanics, *Annual Review of Fluid Mechanics* 52 (2020) 477–508.
- 781 [33] T. P. Sapsis, Statistics of extreme events in fluid flows and waves, *Annual  
782 Review of Fluid Mechanics* 53 (2020).
- 783 [34] F. Ragone, F. Bouchet, Computation of extreme values of time averaged  
784 observables in climate models with large deviation techniques, *Journal  
785 of Statistical Physics* (2019) 1–29.
- 786 [35] T. P. Sapsis, Output-weighted optimal sampling for bayesian regression  
787 and rare event statistics using few samples, *Proceedings of the Royal  
788 Society A: Mathematical, Physical and Engineering Sciences* 476 (2234)  
789 (2020) 20190834.
- 790 [36] O. Gramstad, C. Agrell, E. Bitner-Gregersen, B. Guo, E. Ruth,  
791 E. Vanem, Sequential sampling method using gaussian process regres-  
792 sion for estimating extreme structural response, *Marine Structures* 72  
793 (2020) 102780.
- 794 [37] M. A. Mohamad, T. P. Sapsis, Sequential sampling strategy for extreme  
795 event statistics in nonlinear dynamical systems, *Proceedings of the Na-  
796 tional Academy of Sciences* 115 (44) (2018) 11138–11143.
- 797 [38] D. Mj, D. Dutykh, Learning extreme wave run-up conditions, *Applied  
798 Ocean Research* 105 (2020) 102400.

- 799 [39] M. A. Mohamad, T. P. Sapsis, Probabilistic response and rare events  
800 in mathieu's equation under correlated parametric excitation, *Ocean*  
801 *Engineering* 120 (2016) 289–297.
- 802 [40] W. Cousins, T. P. Sapsis, Quantification and prediction of extreme  
803 events in a one-dimensional nonlinear dispersive wave model, *Physica*  
804 *D: Nonlinear Phenomena* 280 (2014) 48–58.
- 805 [41] M. Farazmand, T. P. Sapsis, Reduced-order prediction of rogue waves  
806 in two-dimensional deep-water waves, *Journal of Computational Physics*  
807 340 (2017) 418–434.
- 808 [42] M. A. Mohamad, W. Cousins, T. P. Sapsis, A probabilistic decomposi-  
809 tion synthesis method for the quantification of rare events due to internal  
810 instabilities, *Journal of Computational Physics* 322 (2016) 288–308.
- 811 [43] A. V. Slunyaev, A high-order nonlinear envelope equation for gravity  
812 waves in finite-depth water, *Journal of Experimental and Theoretical*  
813 *Physics* 101 (5) (2005) 926–941.
- 814 [44] J. Dalzell, A note on finite depth second-order wave–wave interactions,  
815 *Applied Ocean Research* 21 (3) (1999) 105–111.
- 816 [45] J. N. Sharma, R. G. Dean, Second-order directional seas and associated  
817 wave forces, *Soc. Pet. Eng. J.* 21 (01) (1981) 129–140.
- 818 [46] K. Hasselmann, T. P. Barnett, E. Bouws, H. Carlson, D. E. Cartwright,  
819 K. Enke, J. A. Ewing, H. Gienapp, D. E. Hasselmann, P. Kruseman,  
820 Measurements of wind-wave growth and swell decay during the Joint  
821 North Sea Wave Project (JONSWAP), *Ergänzungsh.* 8-12 (1973).
- 822 [47] P. S. Tromans, L. Vanderschuren, Response based design conditions in  
823 the north sea: application of a new method, in: *27th Offshore technology*  
824 *conference*, Vol. 1, Offshore Technology Conference, 1995, p. 387.
- 825 [48] F. Fedele, On oceanic rogue waves, arXiv preprint arXiv:1501.03370  
826 (2015).
- 827 [49] H. E. Krogstad, J. Liu, H. Socquet-Juglard, K. B. Dysthe, K. Trulsen,  
828 Spatial extreme value analysis of nonlinear simulations of random sur-  
829 face waves, in: *International Conference on Offshore Mechanics and*  
830 *Arctic Engineering*, Vol. 37440, 2004, pp. 285–295.

- 831 [50] L. Breiman, Random forests, *Machine learning* 45 (1) (2001) 5–32.
- 832 [51] J. D. Rodriguez, A. Perez, J. A. Lozano, Sensitivity analysis of k-fold  
833 cross validation in prediction error estimation, *IEEE transactions on*  
834 *pattern analysis and machine intelligence* 32 (3) (2009) 569–575.
- 835 [52] J. Snoek, H. Larochelle, R. Adams, Practical bayesian optimization of  
836 machine learning algorithms advances in neural information processing  
837 systems 25 (2012).
- 838 [53] J. S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-  
839 parameter optimization, in: *Advances in neural information processing*  
840 systems, 2011, pp. 2546–2554.
- 841 [54] K. Trulsen, K. B. Dysthe, A modified nonlinear schrödinger equation for  
842 broader bandwidth gravity waves on deep water, *Wave motion* 24 (3)  
843 (1996) 281–289.
- 844 [55] M. L. McAllister, T. A. A. Adcock, P. H. Taylor, T. S. Van Den Bremer,  
845 The set-down and set-up of directionally spread and crossing surface  
846 gravity wave groups, *Journal of Fluid Mechanics* 835 (2018) 131–169.