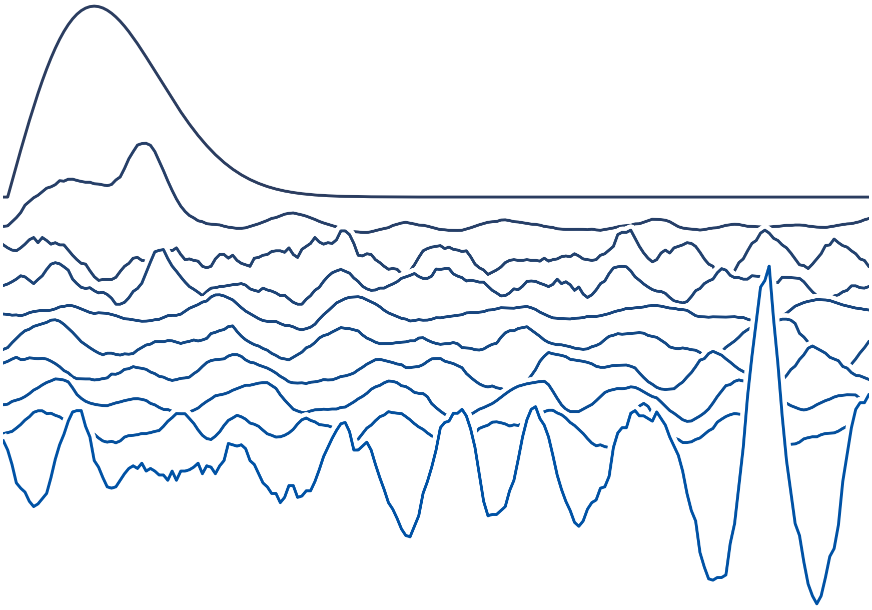
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD thesis

An Ocean of Data

Inferring the Causes of Real-World Rogue Waves

Dion Häfner

Advisor: Prof. Markus Jochum

Submitted: April 4, 2022

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Dion Häfner: *"An Ocean of Data — Inferring the Causes of Real-World Rogue Waves"*

This PhD project has been carried out at

NIELS BOHR INSTITUTE, UNIVERSITY OF COPENHAGEN

under the supervision of

PROF. DR. MARKUS JOCHUM
(Niels Bohr Institute, University of Copenhagen)

ABSTRACT — ENGLISH

Rogue waves are rare surface waves in the ocean that are significantly larger than the general wave population. Although they pose a serious threat to mariners, the causes of these waves in the real ocean are still poorly understood, and they remain hard to forecast. This is due to the lack of a high-quality observational dataset, the rarity of these waves and therefore required amounts of data, the difficulty of analyzing said data, and the lack of a principled way to infer causation. This thesis consists of a collection of 3 articles that address all of these issues through a combination of data mining, interpretable machine learning, and causal analysis based on domain knowledge. The first article describes the assembly of a comprehensive wave catalogue processing over 700 years of sea surface elevation time series from 158 buoy locations. The second article presents an analysis on the leading-order effects governing rogue wave formation based on interpretable machine learning. The third article extends this to a fully nonlinear predictive model by searching for a causally consistent neural network, and presents a path to an improved rogue wave forecast. Finally, I discuss the implications of our findings for future rogue wave research, and outline how machine learning can augment the scientific method and guide us towards scientific discovery.

ABSTRACT — DANSK

Ekstreme bølger er sjældne overfladebølger i havet, der er betydeligt større end den generelle bølgepopulation. Selvom de udgør en alvorlig trussel mod søfolk, er årsagerne til disse bølger i det virkelige hav stadig dårligt forstået, og de er stadig svære at forudsige. Dette skyldes manglen på et observationsdatasæt af høj kvalitet, sjældenheden af disse bølger og derfor nødvendige mængder af data, vanskeligheden ved at analysere nævnte data og manglen på en principiel måde at udlede årsagssammenhæng. Denne afhandling består af en samling af 3 artikler, der behandler alle disse problemstillinger gennem en kombination af datamining, fortolkelig maskinlæring og kausal analyse baseret på domæneviden. Den første artikel beskriver samlingen af et omfattende bølgekatalog, der behandler over 700 års havoverfladehøjdetidsserier fra 158 bøjeplaceringer. Den anden artikel præsenterer en analyse af de førende ordenseffekter, der styrer dannelsen af ekstreme bølger, baseret på fortolkelig maskinlæring. Den tredje artikel udvider dette til en ikke-lineær prædiktiv model ved at søge efter et kausalt konsistent neuralt netværk og præsenterer en vej til en forbedret ekstrem bølge-prognose. Til sidst diskuterer jeg implikationerne af vores resultater for fremtidig ekstrem bølge-forskning og skitserer, hvordan maskinlæring kan forstærke den videnskabelige metode og guide os mod videnskabelig opdagelse.

> “*The human realm is ruled by three elements: time, space, and probability.*”
>
> —Haruki Murakami

Contents

Prologue

0

This PhD thesis is about the combination of rogue waves and machine learning. And what a combination that is! One is an unexpected menace, preying on anyone brave or stupid enough to enter its domain, with the sole purpose of pulling them into the abyss. The other is an unusually big wave in the ocean.

Figure 0.1: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"a huge ocean wave unsplash"*.

Machine learning, and especially deep learning, has rapidly transformed the world. Since the first large-scale applications of artificial neural networks in the early 2010's we have seen unparalleled machine performance in natural language processing (Brown et al., 2020; Devlin et al., 2018), image recognition (Krizhevsky, Sutskever, and Hinton, 2017), recommender systems (Ma et al., 2020), generative art (Fig. 0.1; Esser, Rombach, and Ommer, 2021), chess (Silver et al., 2017), Go (Silver et al., 2016), and StarCraft II (Vinyals et al., 2019). Entire industries have been transformed, to the point that previously "analogue" companies like Walmart and Home Depot are now publishing machine learning research (e. g. Kouki et al., 2020; Xu et al., 2019).

And yet, despite almost unprecedented levels of enthusiasm in its adoption (and funding), it is still surprisingly difficult to use machine learning to answer scientific questions.

In fact, there are few examples where machine learning has directly led to scientific *discovery* (Succi and Coveney, 2019). That is in part because the goals of machine learning — at least the kind that has led to the successes described above — are often opposed to the goals of science. Science is about finding universal laws that encapsulate a causal relationship in our world so that we understand it better. Machine learning on the other hand performs best when "good enough" is an acceptable outcome, and where it doesn't matter how the algorithm arrives at its answer. This disconnect has not gone unnoticed in machine learning research (Marcus, 2022), especially since some areas where machine learning *should* do well (like driverless cars or medical diagnoses, Varoquaux and Cheplygina, 2021) prove so far elusive.

One promising line of research to address this is causality, where statistical models are imbued with the capability to perform causal reasoning (see e. g. Runge et al., 2019). This allows algorithms to uncover causal connections instead of associations (causal discovery), especially if we allow agents to suggest experiments (interventions) and observe the outcome. But we are still a long way from large-scale adoption of these tools.

In the meantime, the goal of this PhD thesis is to explore how we can use machine learning *today* to understand a real physical phenomenon. Because ultimately, despite their issues, machine learning models are immensely powerful products of mathematics and engineering capable of large-scale data processing like no other. In this study we use machine learning mostly as a tool for large-scale data analysis and inference, prioritizing understanding (and discovery) over prediction. This turns out to be a powerful approach[1], but also a very challenging one that is inherently interdisciplinary and requires a solid foundation in both machine learning and the target scientific domain.

Rogue waves as a study object do not make this task any easier. Most machine learning algorithms struggle with low probabilities, and in the case of rogue waves, they are excessively low. But on the other hand, this also makes them a good target to study with machine learning: Low probabilities imply the need to process massive amounts of data that are difficult to analyze with traditional methods (let alone with human intuition). And of course, they are a fascinating phenomenon that I am proud to have been working on for the past 3 years.

[1] Voit (2019) calls this "data mining-based induction" and even postulates this to become a fundamental extension to the scienfic method.

CHAPTER CONTENTS

The State of the Art

1

1.1 Rogue Wave Research

What people think of when they hear the term “rogue wave” or “freak wave” heavily depends on their personal context, even among experts.

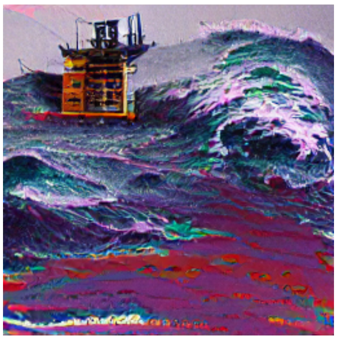


Figure 1.1: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *“offshore oil platform in a storm | rogue wave | Kodak”*.

In popular science, the term rogue wave is often used to describe any large ocean wave, and they are often credited to be responsible for the loss of ships and lives at sea (Didenkulova, 2019). This is contrary to the scientific definition, which is a relative criterion based on the observed wave height H and the height of the surrounding waves, characterized by the significant wave height H_s:

ROGUE WAVE CRITERION
$$\frac{H}{H_s} > \kappa \tag{1.1}$$

Usually, the rogue threshold κ is taken to be 2.0 or 2.2 for crest-to-trough wave heights and 1.2 for crest heights. The significant wave height H_s is defined as 4 times the standard deviation of the surface elevation (see Fig. 1.2 for an illustration of these quantities). This is roughly equivalent to the mean of the highest third of waves, which aligns with the average wave height reported by a trained observer (Holthuijsen, 2010).

The definition (1.1) immediately reveals the first fundamental issue in rogue wave research: most rogue waves are neither dangerous nor interesting. It is not noteworthy when a 50 cm wave occurs in a 20 cm sea state, but it is as

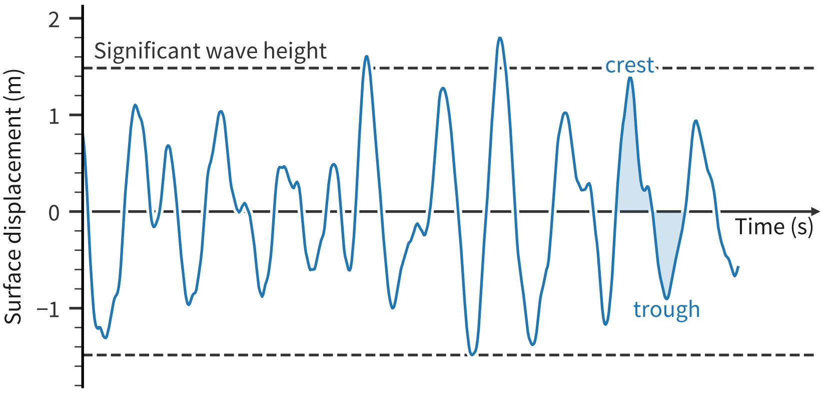


Figure 1.2: Anatomy of an Eulerian wave observation, in which the observer is fixed in space (like a seaward facing laser, or — approximately — a tightly moored wave buoy).

much a rogue wave as a 20 m wave in an 8 m sea. For this definition to make sense, it implicitly encodes 2 fundamental assumptions:

1. Small rogue waves are caused by the same generation mechanisms as big rogue waves.
2. Waves above the rogue wave threshold are somehow fundamentally different from those below it.

Both assumptions are non-trivial. In fact, the articles in Chapter 2 present evidence that the second assumption does not hold throughout most sea states in the real ocean, and Chapter 3 discusses some of the implications.

Research interest in rogue waves was originally triggered by the indisputable measurement of a rogue wave at the Draupner oil rig in the North Sea in 1995, at a wave height of 25.6 m and crest height of 18.5 m during a storm with significant wave height of 12 m (Haver, 2004; Sunde, 1995). With a relative crest height of 1.55, this event would be extremely rare under the existing theory for linear, narrow-bandwidth waves (Longuet-Higgins, 1952). This disconnect sent the research community searching for a theory that attaches a higher probability to this and similar events.

This ultimately resulted in a debate on the fundamental nature of these waves: are they themselves extremely rare, or the conditions under which they are generated? Or in the words of Hayer and Andersen: *"Freak waves: rare realizations of a typical population or typical realizations of a rare population?"* (Hayer and Andersen, 2000)[1]. The following sections outline the ideas behind both hypotheses, and present the state of the art in rogue wave research.

1.1.1 *Linear Waves*

To lowest order, the properties of a 1-dimensional wave measurement (like a time series observation at a fixed location) are fully described by its spectral density $\mathcal{S}(f)$, often just called a "wave spectrum" (see Fig. 1.3 for an example).

To see this, we adopt a simple model called the random phase-amplitude model (see e. g. Holthuijsen, 2010): We view the wave train as a superposition of independent harmonics with frequency f, where each harmonic has an amplitude depending on the corresponding value of the wave spectrum $\mathcal{S}(f)$ and an independent, uniformly random phase $\phi \in (0, 2\pi)$. After all, many processes acting on waves in the real ocean are highly stochastic — like wave generation from winds or scattering and refraction at a fractal coastline geometry — so a random phase without a preferred value makes intuitive sense. In this case, the surface elevation η is just the sum of each harmonic:

[1] This is in fact an excellent question to address with machine learning. All we need to do is to see how well a model can reliably predict rogue waves given the sea state — and hope that we have collected enough and the right kind of data.

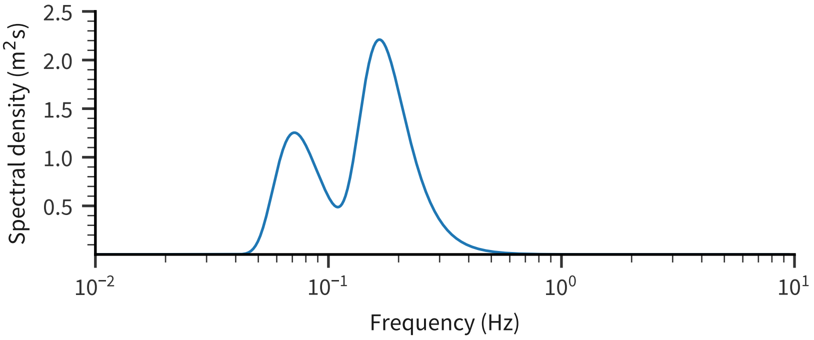


Figure 1.3: A typical bi-modal wave spectrum representing the overlap of swell and wind sea. Idealized Ochi-Hubble six-parameter wave spectrum with spectral peaks at periods 6 s and 14 s (Ochi and Hubble, 1976).

$$\eta(t) = \sum_i \sqrt{2\Delta f S(f_i)} \sin(2\pi f_i t + \phi_i) \tag{1.2}$$

with time t, frequency f, frequency resolution Δf, and random phase ϕ. In the case of a spectrum with many independent harmonics f_i (as in the real ocean), this represents the sum of a large number of random variables with finite mean and variance. So per the central limit theorem, η is a Gaussian random variable with zero mean and a variance that is fully determined by the significant wave height. This inherent stochasticity of random phases also implies that even under an identical spectrum no two wave fields will look exactly the same (Fig. 1.4).

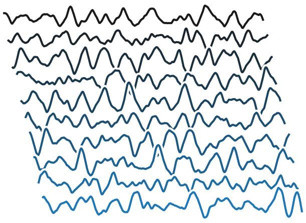


Figure 1.4: An ensemble of sea surface elevations drawn from the same wave spectrum (as shown in Fig. 1.3).

In the limit of a narrow-band spectrum, the sea surface elevation only has one maximum / minimum per wave, and the wave heights and crest heights are Rayleigh distributed (Holthuijsen, 2010; Longuet-Higgins, 1952):

RAYLEIGH WAVE DISTRIBUTION

WAVE HEIGHTS $$P(H/H_s > \kappa) = \exp(-2\kappa^2) \tag{1.3}$$

CREST HEIGHTS $$P(h/H_s > \kappa) = \exp(-8\kappa^2) \tag{1.4}$$

These probability distributions will serve as the baseline for all further comparisons. They also tells us that, under these assumptions[2], we would expect about 1 in 10 000 waves to be a rogue wave (with a threshold $\kappa = 2.0$ for waves and 1.2 for crests) through mere random linear superposition.

Unfortunately, the real ocean is not so simple. One commonly violated assumption is that of *narrow bandwidth*, which is used to derive the Rayleigh wave height distribution above. In fact, most seas do *not* have Rayleigh distributed wave heights, as we will see in Chapter 2 (not even seas that are approximately Gaussian). In particular, to create a rogue wave, both crest and trough have to be large, which makes them sensitive to the group structure of the wave train.

[2] Assumptions behind Rayleigh-distributed wave heights:

i) independent, non-interacting harmonics (linear waves);

ii) narrow spectral bandwidth.

When taking finite bandwidths into account, things are more difficult, and there are several competing wave height distributions in bandwidth-limited seas (e. g. the Boccotti, Naess, and Tayfun distributions: Boccotti, 1989; Naess, 1985; Tayfun, 1990). As an example, the Tayfun distribution is based on a parameter r (which we call *crest-trough correlation*) that is the value of the wave envelope at half the zero-crossing period (i. e., at the expected location of the trough following a crest). For large wave heights $\gtrsim H_s$ it can be approximated as (Tayfun and Fedele, 2007):

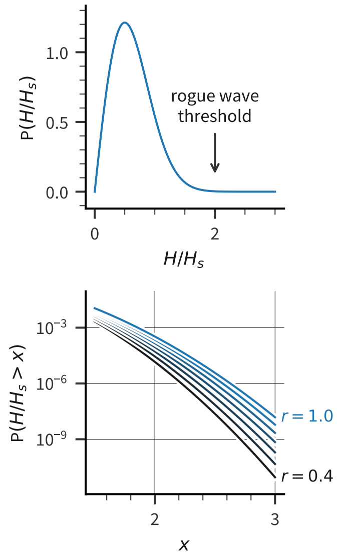


Figure 1.5: Wave height probability density (top) and survival function for large wave heights (bottom). Curves are Tayfun distributions as in (1.5) with different values of r. The case $r = 1$ is identical to the Rayleigh distribution (1.3).

$$P(H/H_s > \kappa) = \sqrt{\frac{1+r}{2r}}\left(1 + \frac{1-r^2}{4r\kappa^2}\right)\exp\left(-\frac{1}{4(1+r)}\kappa^2\right) \tag{1.5}$$

with $r \in [0, 1]$. In the limit $r \longrightarrow 1$ this reduces to the Rayleigh distribution for wave heights (Fig. 1.5).

1.1.2 *The Stokes Wave*

So far we have only considered waves and crests with independently random phases. This assumption is not fulfilled anymore as soon as waves are allowed to interact with each other, which couples the phases of different harmonics. Stokes theory extends this to weakly nonlinear waves with low characteristic steepness $\varepsilon = kH$ (with wave number k).

As in virtually all problems in fluid dynamics, an appropriate starting point is with the incompressible Navier-Stokes equations and the continuity equation (encoding momentum balance and mass conservation, respectively):

NAVIER-STOKES EQUATIONS

$$\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla)\vec{u} - \nu\nabla^2\vec{u} = -\frac{1}{\rho}\nabla p - g \cdot \hat{k} \tag{1.6}$$

$$\nabla \cdot \vec{u} = 0 \tag{1.7}$$

with velocity vector $\vec{u}$, viscosity ν, pressure p, density ρ, gravitational acceleration g, and unity vector in z direction $\hat{k}$.

Assuming inviscid ($\nu = 0$) and irrotational ($\nabla \times \vec{u} = 0$) fluid flow, we can introduce a velocity potential ϕ:

VELOCITY POTENTIAL

$$\nabla\phi = \vec{u} \tag{1.8}$$

This reduces the Navier-Stokes equations to the Bernoulli equation, and the continuity equation to a Laplace equation (see e. g. Holthuijsen, 2010):

$$\frac{\partial\phi}{\partial t} + \frac{1}{2}|\nabla\phi|^2\frac{p}{\rho} + gz = 0 \tag{1.9}$$

BERNOULLI EQUATION

$$\nabla^2\phi = 0 \tag{1.10}$$

A central missing ingredient is a set of boundary conditions at the top and bottom of the sea that give rise to finite surface elevations — waves — and shallow water effects. At each boundary we impose a kinematic boundary condition that ensures that water particles only move parallel to the respective surface $\eta(x, y, t)$:

$$u_z = \frac{\partial\eta}{\partial t} + u_x\frac{\partial\eta}{\partial x} + u_y\frac{\partial\eta}{\partial y} \qquad \text{at } z = \eta \tag{1.11}$$

KINEMATIC BOUNDARY CONDITION

$$\Leftrightarrow \quad \frac{\partial\phi}{\partial z} = \frac{\partial\eta}{\partial t} + \frac{\partial\phi}{\partial x}\frac{\partial\eta}{\partial x} + \frac{\partial\phi}{\partial y}\frac{\partial\eta}{\partial y} \qquad \text{at } z = \eta \tag{1.12}$$

For a flat bottom this just reduces to $\partial\phi/\partial z = 0$ at the ocean floor, but at the surface all terms are generally non-zero. At the surface we also find a dynamic boundary condition for the pressure p that we plug into the Bernoulli equation:

$$p = 0 \quad \Rightarrow \quad \frac{\partial\phi}{\partial t} + \frac{1}{2}|\nabla\phi|^2 + g\eta = 0 \qquad \text{at } z = \eta \tag{1.13}$$

DYNAMIC BOUNDARY CONDITION

This assumes that the pressure at the water surface equals a constant atmospheric pressure.

The set of equations (1.9)–(1.13) gives rise to a whole zoo of surface gravity waves in the ocean[3]. The equations are nonlinear (containing terms $\propto |\phi|^2$ and $\nabla\phi \cdot \eta$) and cannot be solved analytically without further assumptions. The linear wave solution with non-interacting harmonics (as in § 1.1.1) is recovered by dropping all nonlinear terms and using the plane wave ansatz $\eta(x, t) = a\cos(\omega t - kx)$ with amplitude a, frequency ω, and wave number k.

In the Stokes wave expansion, all nonlinear terms and unknown quantities (such as η and ω) are expanded in orders of the (assumed) small parameter $\varepsilon = ak$, the characteristic wave steepness (see e. g. Dean and Dalrymple, 1991). By keeping only terms up to a certain order n in ε, this leads to weakly nonlinear corrections of n-th order that generate wave trains with higher crests and flatter troughs than purely linear waves.

Weakly nonlinear corrections also cause a modification of the wave height distribution and enhance rogue wave probabilities, especially for rogue crests (Fedele et al., 2016, 2019; Gemmrich and Garrett, 2011). This leads to conditions that have slightly elevated rogue wave probabilities, which supports the “rare realizations of a typical population” theory of rogue waves.

[3] Excluding planetary-scale waves like Rossby and Kelvin waves, and neglecting interactions with bottom topograpy and breaking waves.

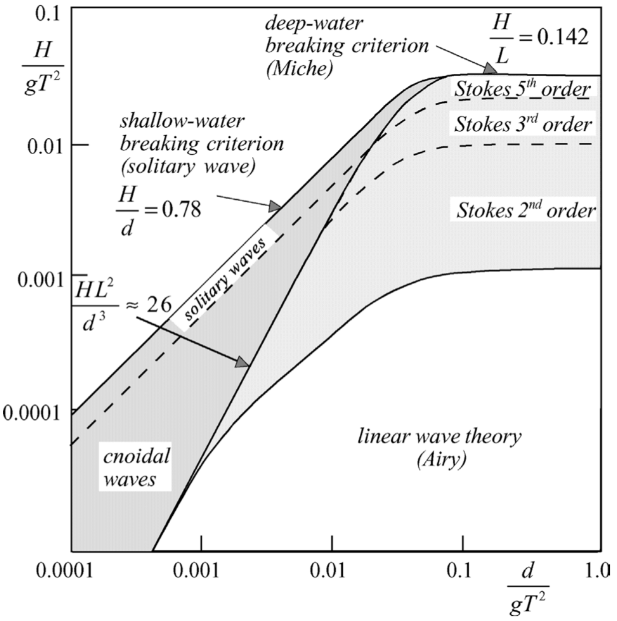


Figure 1.6: The range of applicability for different weakly nonlinear theories. From Holthuijsen (2010), originally Le Mehaute (1969). Here, T is the wave period, H wave height, d water depth, L wavelength.

1.1.3 *Cnoidal Waves*

In shallow water the Stokes expansion converges very slowly, which makes Stokes theory inapplicable in this case (see Fig. 1.6 for an overview). A characteristic parameter in this context is the Ursell number (Ursell, 1953):

$$\mathrm{Ur} = \frac{\lambda^2 H}{D^3} \tag{1.14}$$

URSELL NUMBER

with wavelength $\lambda = 2\pi/k$, wave height H, and water depth D. For high values of Ur, an expansion in the relative depth $\widetilde{D} = kD$ is more fruitful than the Stokes expansion (Dean and Dalrymple, 1991), which leads to the Korteweg-de Vries (KdV) equation and cnoidal theory (Korteweg and De Vries, 1895). Notably, cnoidal theory is the simplest theory that allows for solitary waves (solitons) — waves that travel entirely above the water level and preserve their shape. Solitons have been studied intensely as a possible mechanism for rogue wave generation (Chabchoub, Hoffmann, and Akhmediev, 2011; Clamond and Grue, 2002; Kharif and Pelinovsky, 2003).

1.1.4 *Highly Nonlinear Theory*

A large body of rogue wave research does not consider linear and weakly nonlinear solutions, as it is implicitly assumed that these mechanisms cannot be responsible for observed extreme rogue waves like the Draupner wave. Instead, these studies focus on highly nonlinear phenomena[4] such as breathers, solitons, or the modulational instability as possible creation mechanisms (e. g. Dematteis et al., 2019; Kharif et al., 2001; Kharif and Pelinovsky, 2003; Onorato et al., 2006; Onorato and Proment, 2012; Shukla et al., 2006; Toffoli et al., 2010).

A prototypical framework for these solutions is the nonlinear Schrödinger equation (NLS), which is also based on an expansion in orders of characteristic steepness ε and an expansion of the dispersion relation around a dominant wave number k_0 / frequency ω_0 (see e. g. Johnson, 1997, for a derivation). In contrast to the Stokes wave solution, the (now complex) wave amplitude $A(x,t)$ is allowed to evolve in time and space and satisfies the nonlinear Schrödinger equation (Slunyaev, Didenkulova, and Pelinovsky, 2011):

$$-2i\left(\frac{\partial A}{\partial t}+c_g\frac{\partial A}{\partial x}\right)+\frac{\omega_0}{8k_0^2}\frac{\partial^2 A}{\partial x^2}+\frac{\omega_0 k_0^2}{2}A|A|^2=0 \tag{1.15}$$

NONLINEAR SCHRÖDINGER EQUATION

with group speed c_g. This equation has solutions that grow exponentially due to energy transfer between the carrier wave and its sidebands, an effect called modulational instability or Benjamin-Feir instablity (Benjamin and Feir, 1967). These solutions are referred to as breathers, one of which is the Peregrine soliton (Peregrine, 1983, Fig. 1.7). The strength of the modulational instability is governed by the Benjamin-Feir index (Alber and Stewartson, 1978):

$$\mathrm{BFI}=\frac{k_0 A}{\Delta\omega/\omega_0} \tag{1.16}$$

BENJAMIN-FEIR INDEX

with spectral bandwidth $\Delta\omega$. The original derivation of the nonlinear Schrödinger equation assumes deep water, unidirectional propagation, and narrow-banded spectra. Modifications that relax these assumptions exist (Davey and Stewartson, 1974; Dysthe and Longuet-Higgins, 1979), and modified versions of the BFI that take shallow water and directional spreading into account have been suggested (Fedele, 2015; Serio et al., 2005). There is good evidence demonstrating the modulational instability in wave tanks (Onorato et al., 2006), but studies considering real ocean conditions have so far not confirmed an enhancement of extreme waves (Gramstad and Trulsen, 2007; Xiao et al., 2013).

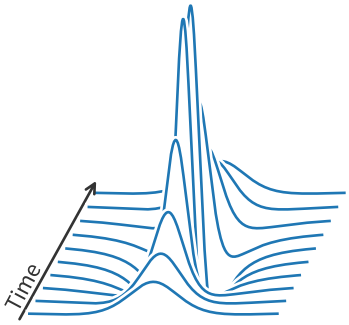


Figure 1.7: The Peregrine solution. Shown is the evolution of the wave height envelope A in space and time, with a clear localized maximum.

The nonlinear Schrödinger equation is not the only nonlinear wave equation with unstable solutions. In general, waves transfer energy via nonlinear

[4] Highly nonlinear in the sense that these waves are not just small perturbations to the linear wave profile, but entirely new solutions with unique properties.

four-wave interactions (Hasselmann, 1966). This is accounted for explicitly in the Zakharov equation (Zakharov, 1968), which can be studied to derive higher-order corrections to the wave height distribution (e. g. as in Janssen, 2003).

1.1.5 *Other Causes of Rogue Waves*

There are several other hypothesized causes for rogue waves that we have not considered so far (see Adcock and Taylor, 2014; Dudley et al., 2019; Slunyaev, Didenkulova, and Pelinovsky, 2011, for reviews). While the Bernoulli equations (1.9) and associated boundary conditions are very general in terms of the permitted dynamics *within* the fluid, most of the real-world complexities *outside* the fluid are neglected. Examples for this include:

- Interactions with non-uniform topography such as abrupt transitions in water depth or waves on top of a slope (Trulsen, Zeng, and Gramstad, 2012);
- The non-stationarity of the sea state, i. e., its evolution in time (Trulsen, 2018);
- The interaction between waves and currents (Didenkulova, Talipova, and Pelinovsky, 2021; Mallory, 1974; Onorato, Proment, and Toffoli, 2011);
- Direct wind-wave interactions (Adcock and Taylor, 2011);
- Wave breaking, e. g. the influence of crossing seas on the onset and shape of breaking waves (McAllister et al., 2019).

All of these effects impact the formation of large waves, but they are also inherently *local*, which causes them to be averaged out of bulk statistics (such as buoy measurements from many different locations).

1.2 PHYSICS AND MACHINE LEARNING

The are many examples of studies that apply machine learning to physical problems, most of which aim for improvements in computational efficiency or predictive performance of simulations (e. g. Bar-Sinai et al., 2018; Cranmer et al., 2020b, 2021; Kochkov et al., 2021; Li et al., 2020; Pestourie et al., 2021). These efforts undoubtedly contribute tremendous value. Yet, better *predictions* are not the same as improved *understanding*, the foundation of all science. Ideally, machine learning would lead to advances on both fronts, but unfortunately, process understanding seems much harder to come by, in part also due to the immense complexity of real-world data and governing processes (Fig. 1.9; Reichstein et al., 2019).

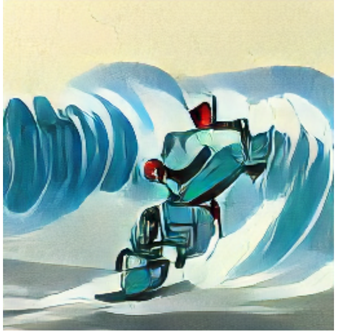


Figure 1.8: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"a cartoon robot surfing on a big wave"*.

One necessary ingredient for true understanding is the robust identification of causal connections over mere association. The emerging field of causality has formalized the identifiability of causal connections from data and provides tools for both causal inference and causal discovery (see e. g. Peters, Janzing, and Schölkopf, 2017). There are first promising applications of these methods (for example in climate: Hannart et al., 2016; Kretschmer et al., 2016; Runge et al., 2019), but it is still a long way to go before we will be able to identify arbitrary causal connections in real-world spatiotemporal systems. Nevertheless, explicitly encoding or enforcing causal relationships in models is a promising way to make machine learning a more dependable tool for scientific discovery.

Causal connections in physical systems are typically representable by simple mathematical relationships[5]. Machine learning can exploit this through sym-

[5] An observation dubbed *"The unreasonable effectiveness of mathematics in the natural sciences"* (Wigner, 1960).

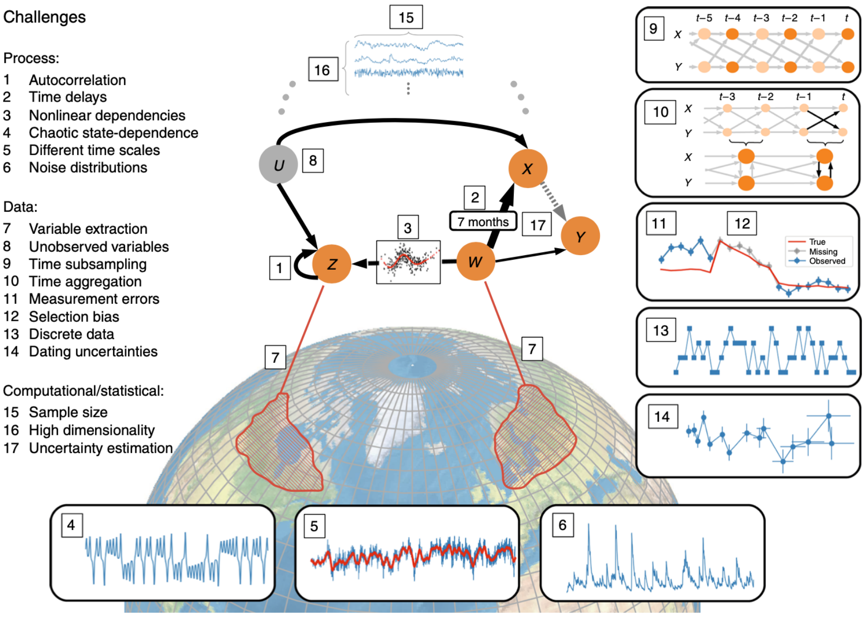


Figure 1.9: Challenges when applying machine learning to earth system data. Figure from Reichstein et al. (2019).

bolic regression, a method that aims to fit sparse mathematical expressions to data. While the idea itself is not new (traditionally based on genetic programming, Schmidt and Lipson, 2009), it is now elevated through increasingly sophisticated machine learning algorithms, with some first successes (Cranmer et al., 2020a; Lemos et al., 2022; Udrescu and Tegmark, 2020; Zanna and Bolton, 2020).

An in physics ubiquitous special case is systems of differential equations, which can be replaced or augmented with neural networks (neural ODEs / UDEs, Chen et al., 2019; Kidger, 2022; Rackauckas et al., 2021). In combination with symbolic regression, this leads to methods for the automated discovery of differential equations (and thus system dynamics) from data, an approach that shows huge potential (Bakarji et al., 2022; Brunton, Proctor, and Kutz, 2016; Champion et al., 2019; Long, Lu, and Dong, 2019; Reinbold et al., 2021) but is still very much a matter of active research, and still struggles with observational noise.

A methodologically much simpler approach is "data-mining inspired induction" (Voit, 2019), where interpretable machine learning (Molnar, 2020) and data mining guide the scientist towards the formation of hypotheses that can then be independently verified — as opposed to setting out with a specific hypothesis to test.

Data-mining inspired induction addresses a common challenge in modern science: data volumes have increased exponentially in the last decades (Fig. 1.10), while human resources are approximately fixed. The central idea is to combine the strengths of machine learning models (large-scale data analysis taking into account orders of magnitude more data than the human mind could) and humans (causal reasoning and interpretation) into a modern scientific workflow. The remainder of this thesis is a concrete application of this approach and serves as a case study on its feasibility, challenges, and opportunities on a real physical problem.

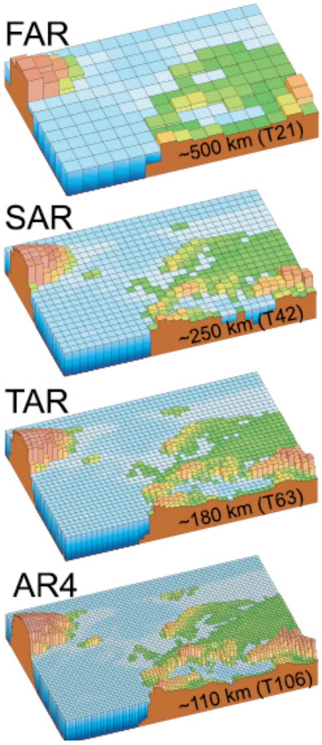


Figure 1.10: The increasing resolution of climate models over time leads to exponentially increasing data volumes. Horizontal resolution over Europe from IPCC (Intergovernmental Panel on Climate Change) reports FAR (1990), SAR (1996), TAR (2001a), and AR4 (2007). Figure from IPCC (2007).

CHAPTER CONTENTS

2 Inferring the Causes of Real-World Rogue Waves

The overarching goal of this thesis is to infer the causes of measured oceanic rogue waves from data. This could provide some much needed evidence to the field, because there are several plausible hypotheses on the generation mechanisms of rogue waves, but no consensus regarding which ones are dominant in the ocean and where to put the main research focus (see § 1.1).

Figure 2.1: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"offshore oil platform in a storm | rogue wave | illustration"*.

The central idea to tackle this is to infer how rogue wave occurrence depends on the sea state (i. e., which parameters govern rogue wave generation). We can then tie this back to a generation mechanism by interpreting the identified dependencies in light of the comprehensive corpus of theoretical literature.

To perform this inference we use "black box" machine learning methods such as deep neural networks and random forest classifiers, in connection with more traditional Bayesian methods and a causal analysis. The immense success of machine learning has often falsely led it to be considered a silver bullet that can and should be applied to any data problem, even though traditional statistics and data analysis methods may offer better or more interpretable results. But in the case of rogue waves, modern machine learning is able to handle a unique set of challenges that traditional methods are not equipped to tackle.

Algorithmic requirements to study probabilistic extreme events:

1) Scales to massive amounts of data;
2) Quantifies uncertainty;
3) Captures nonlinearity and interactions.

Rogue waves are exceedingly rare events, so we must collect and process massive amounts of data. Also, rogue waves may occur with a non-zero probability in any condition (i. e., the classes "rogue seas" and "non-rogue seas" are not separable). This means that we cannot discard non-events, and only relative event rates — *rogue wave probabilities* — are meaningful. To make things worse, the fact that they are rare events forces us to carefully consider the effects of limited data volumes. This makes for an enormously challenging task for learning algorithms: They need to scale to *big data* while providing uncertainty quantification and robustness that is mostly used in the context of *little data* — and typically computationally expensive, e. g. when using Bayesian methods based on sampling from a posterior distribution. At the same time, there is no universally accepted parametric form on how rogue wave probabilities depend on the sea state (or even which parameters are sufficient to fully characterize a sea state). This means that any chosen method will have to be *flexible*, ideally supporting arbitrary nonlinear connections between parameters. Scaling and flexibility are inherent strengths of machine learning, and many methods can be augmented to quantify uncertainties. This makes them an excellent fit for this task.

The full objectives of this study are:

OBJECTIVES

i) Assemble a dataset that contains enough data to study rogue waves throughout a wide regime of sea states.

ii) Explore the unique challenges in large-scale wave data curation for machine learning applications (since this is the first study at this scale).

iii) Analyze how rogue wave occurrence depends on sea state parameters, taking uncertainties due to limited data into account.

iv) Determine whether there are sea states of significantly higher rogue wave risk (which could settle the question whether rogue waves are rare realizations of common sea states or common realizations of rare sea states).

v) Infer the dominant creation mechanisms of rogue waves.

vi) Suggest a way forward for rogue wave *prediction*.

The central part of this thesis consists of 3 research articles that address all of these issues.

2.1 ARTICLE I — FOWD: A FREE OCEAN WAVE DATASET FOR DATA MINING AND MACHINE LEARNING

This first article (Häfner, Gemmrich, and Jochum, 2021a) serves as the foundation of all further work by assembling a large catalogue of waves and sea states that we call FOWD (Free Ocean Wave Dataset).

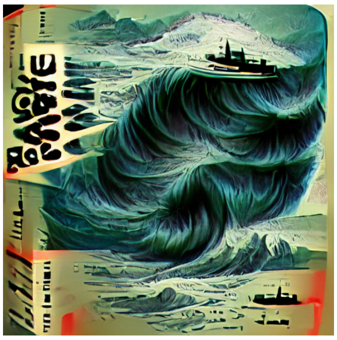


Figure 2.3: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"rogue wave book cover"*.

Since we need as much data as we can possibly get to study extreme wave statistics (which requires many thousands of rogue wave events), we decided to process the entire data catalogue of the Coastal Information Data Program (CDIP, Behrens et al., 2019). CDIP operates a network of more than 150 waverider buoys along the US coast and in US overseas territories (Fig. 2.2), and supplies raw surface elevations in its outputs. This is a huge dataset — some of these buoys have been measuring almost continuously since the 1990s at a sampling frequency of 1.28 Hz, which leads to a combined time series length of over 700 years.

We were also concerned that the usual approach of splitting the time series into equal-time chunks and observed maximum wave height (as e. g. in Casas-Prat and Holthuijsen, 2010; Christou and Ewans, 2014) would not be appropriate for machine learning, because the presence of a rogue wave biases sea state parameters that are sensitive to outliers and leads to confounding (where label information leaks into parameter space)[1]. Therefore, we process the history of every wave in the record separately (over 4 billion waves total) with a running window (Fig. 2.4).

As we have to process billions of sea states, this approach comes with a considerable computational demand. We solve this through a memory-efficient implementation that allows us to process many stations in parallel. The final output dataset is about 1 TB in size and freely available for download.

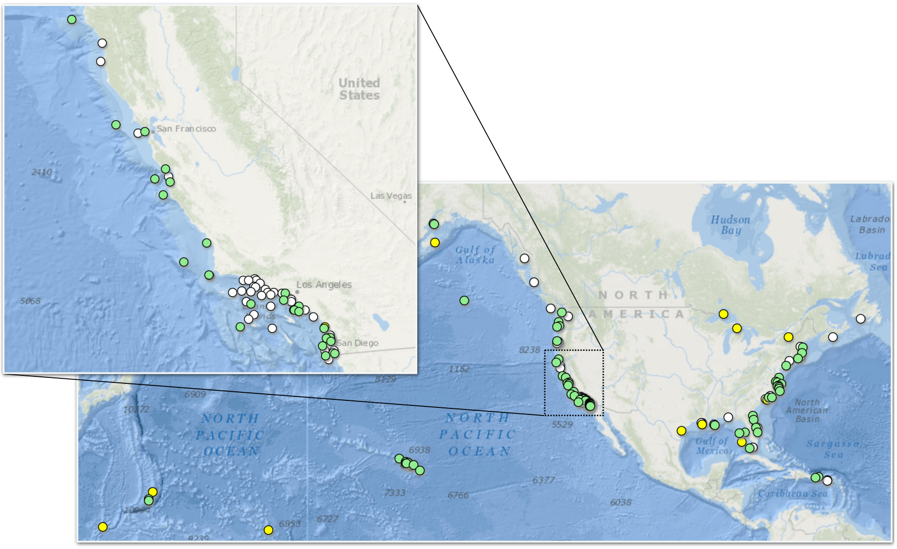
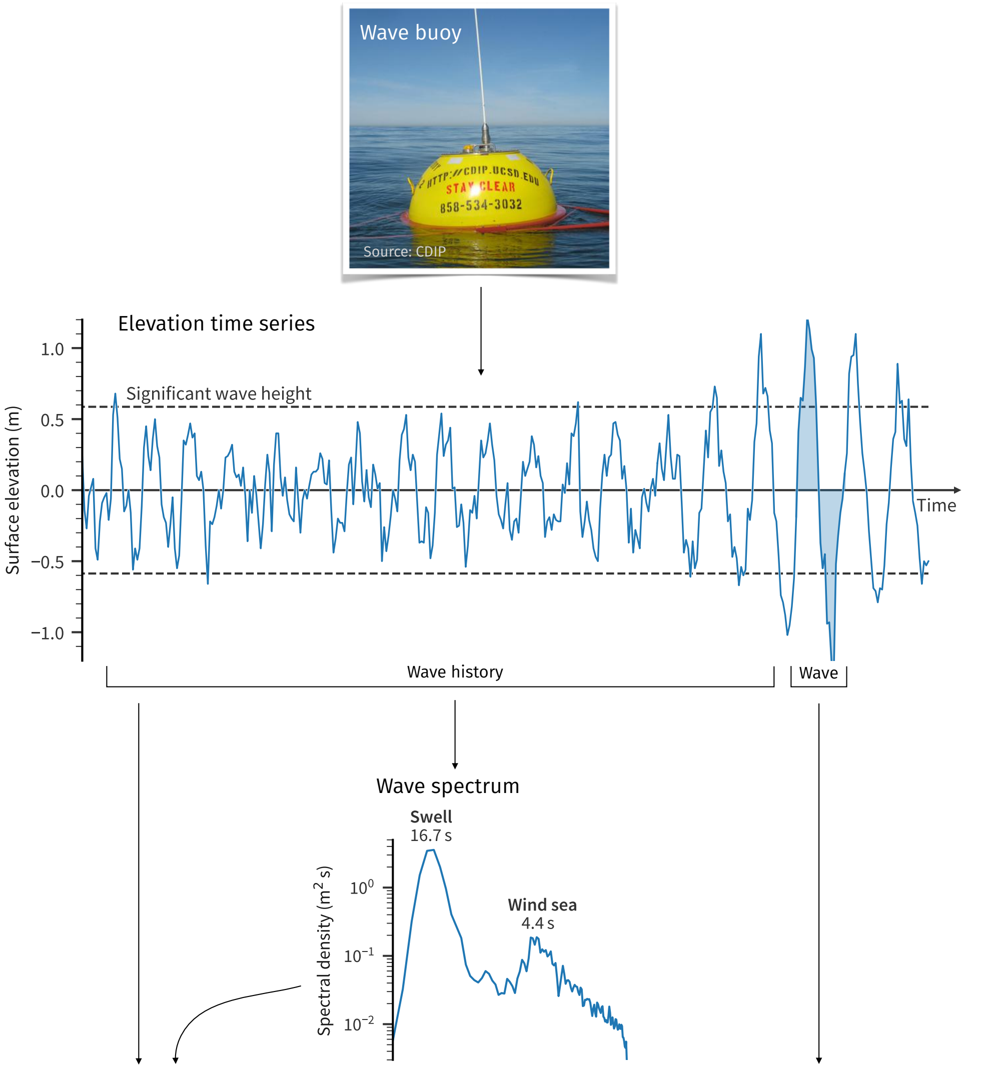


Figure 2.2: The locations of all CDIP buoys. Most buoys are located in relatively shallow water off the Southern Californian coast, but some deep water buoys with long time records exist (e. g. around Hawaii). Based on the CDIP station map at `cdip.ucsd.edu`.

[1] As it turns out, rightfully so — see our findings on surface elevation kurtosis in article 2.



Sea state parameters

```
{
    "start_time": 0.0,
    "end_time": 1829.6875,
    "significant_wave_height_direct": 1.0925,
    "significant_wave_height_spectral": 1.1734,
    "mean_period_direct": 7.7227,
    "mean_period_spectral": 6.3043,
    "maximum_wave_height": 2.12,
    "rel_maximum_wave_height": 1.8067,
    "skewness": 0.0495,
    "kurtosis": 0.2568,
    "valid_data_ratio": 1.0,
    "peak_wave_period": 15.8922,
    "peak_wavelength": 393.0131,
    "steepness": 0.0066,
    "bandwidth_peakedness": 0.1962,
    "bandwidth_narrowness": 0.905,
    "benjamin_feir_index_peakedness": 0.0254,
    "benjamin_feir_index_narrowness": 0.0055,
    "crest_trough_correlation": 0.699,
    "energy_in_frequency_interval": [
        10.4343,
        633.7447,
        120.9921,
        99.3144,
        244.8272
    ],
    "rel_energy_in_frequency_interval": [
        0.0121,
        0.7331,
        0.14,
        0.1149,
        0.2832
    ]
}
```

Wave parameters

```
{
    "start_time": 1782.8125,
    "end_time": 1800.0,
    "id_local": 0.0,
    "zero_crossing_period": 16.4193,
    "zero_crossing_wavelength": 418.8372,
    "maximum_elevation_slope": 0.768,
    "crest_height": 1.22,
    "trough_depth": -1.31,
    "height": 2.53,
    "ursell_number": 0.0555,
    "raw_elevation": [
        0.41,
        0.65,
        0.63,
        0.87,
        1.22,
        1.13,
        0.99,
        0.93,
        0.6,
        0.1,
        -0.37,
        -0.55,
        -0.45,
        -0.94,
        -0.93,
        -1.2,
        -1.31,
        -0.52,
        -0.32,
        -0.16,
        -0.06
    ]
}
```

Figure 2.4: A real example of how FOWD processes a wave record, here containing a rogue wave ($H/H_s = 2.16$) in a swell-dominated sea. Sea state and wave parameters are computed for every zero-crossing wave in the measured elevation time series with a running window.

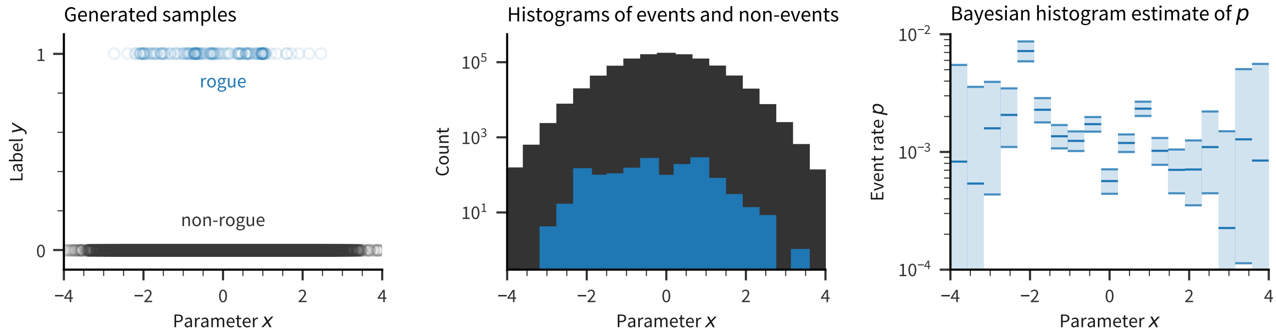
To get a first idea of how rogue waves depend on the sea state, we need to find how the rogue wave probability p varies with each sea state parameter. We also want to achieve this without assuming a functional dependency of p on the parameters (this rules out something like logistic regression which assumes a linear connection), and quantify uncertainties in our estimates.

This led us to develop "Bayesian histograms"[2], where we apply a binning to each parameter and assume that p is identically, independently Beta-distributed within each bin (i. e., that rogue and non-rogue samples within each bin are drawn randomly with a constant rogue wave probability p). Choosing an appropriate conjugate prior for p, this gives us a non-parametric way to estimate p and its uncertainty depending on each parameter without expensive Monte Carlo sampling (Fig. 2.5).

When applying Bayesian histograms to a subset of FOWD, we find that spectral bandwidth, crest-trough correlation, and surface elevation kurtosis are most informative (cause the biggest variation in p) — a result that is revisited and studied in much more detail in article 2.

[2] A Python package for Bayesian histograms is now available at `github.com/dionhaefner/bayesian-histograms`.

Figure 2.5: An example of a Bayesian histogram on generated (fake) data. The Bayesian histogram estimate of the event rate (rogue wave probability) p is based on the ratio between positive and negative samples within each bin. Uncertainties are higher in regions with less data and in regions with lower values of p.



FOWD: A Free Ocean Wave Dataset for Data Mining and Machine Learning

DION HÄFNER,[a] JOHANNES GEMMRICH,[b] AND MARKUS JOCHUM[a]

[a] *Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark*
[b] *University of Victoria, Victoria, British Columbia, Canada*

(Manuscript received 9 November 2020, in final form 24 March 2021)

ABSTRACT: The occurrence of extreme (rogue) waves in the ocean is for the most part still shrouded in mystery, because the rare nature of these events makes them difficult to analyze with traditional methods. Modern data-mining and machine-learning methods provide a promising way out, but they typically rely on the availability of massive amounts of well-cleaned data. To facilitate the application of such data-hungry methods to surface ocean waves, we developed the Free Ocean Wave Dataset (FOWD), a freely available wave dataset and processing framework. FOWD describes the conversion of raw observations into a catalog that maps characteristic sea state parameters to observed wave quantities. Specifically, we employ a running-window approach that respects the nonstationary nature of the oceans, and extensive quality control to reduce bias in the resulting dataset. We also supply a reference Python implementation of the FOWD processing toolkit, which we use to process the entire Coastal Data Information Program (CDIP) buoy data catalog containing over 4 billion waves. In a first experiment, we find that, when the full elevation time series is available, surface elevation kurtosis and maximum wave height are the strongest univariate predictors for rogue wave activity. When just a spectrum is given, crest–trough correlation, spectral bandwidth, and mean period fill this role.

SIGNIFICANCE STATEMENT: Rogue waves are ocean waves that are at least 2 times as high as the surrounding waves. They tend to strike without warning, often damaging ocean-going vessels and offshore structures. Because of their inherent randomness and rarity, there is no satisfying forecasting method for rogue wave risk, nor do we know under which conditions they preferably occur. Modern machine-learning methods provide a promising new alternative, but they require vast amounts of clean data. Here, we provide a way to create such a dataset from ocean surface measurements. We demonstrate our method by processing a buoy dataset containing over 4 billion wave measurements; the result is freely available for download. In a first experiment, we show that it *is* possible to extract risk factors for rogue waves from data, with some conditions producing 10–100 times more rogue waves than others. This work paves the way to a better physical understanding of and better forecasting methods for these dangerous events.

KEYWORDS: Wave properties; Waves, oceanic; Data mining; Data processing; Data quality control; Data science; Machine learning

1. Introduction

During the last 25 years, the study of extreme ocean waves (also known as "rogue waves" or "freak waves") has experienced a renaissance, triggered by the observation of the 25.6-m-high New Year wave at the Draupner oil rig in 1995 (Haver 2004). By now, there are several known mechanisms to generate much higher waves than predicted by linear theory (Adcock and Taylor 2014; Kharif and Pelinovsky 2003; Slunyaev et al. 2011; Dysthe et al. 2008), most of which rely on either highly nonlinear effects like Benjamin–Feir instability (e.g., Gramstad et al. 2018) or weakly nonlinear corrections to the Rayleigh wave height distribution (e.g., Toffoli et al. 2010).

However, while there is plenty of experimental evidence for these mechanisms in wave tanks and simulations, the relative importance of these processes in the real ocean is still unknown. This is evidenced by the rich spectrum of studies emphasizing different physical causes of rogue waves (Janssen and Bidlot 2009; Toffoli et al. 2010; Gemmrich and Garrett 2011; Xiao et al. 2013; Fedele et al. 2016; Gramstad et al. 2018; McAllister et al. 2019). This has the consequence that, so far, there is no reliable forecast for rogue wave risk (see also Dudley et al. 2019), although there have been some recent efforts (Barbariol et al. 2019).

There are several studies that aim to relate sea state parameters to rogue wave occurrence (Cattrell et al. 2018; Casas-Prat and Holthuijsen 2010; Karmpadakis et al. 2020; Gemmrich and Garrett 2011), but they are limited by the analyzed amount of data (often only one or several storms), their coverage of parameter space (often only look at 1 or 2 parameters), or sophistication of analysis (often no uncertainty analysis). To our knowledge, no study has been able to show the dependence of rogue wave occurrence on sea state (or show that it does not exist) with statistical significance throughout a wide regime of sea states.

We attribute this shortcoming to a lack of sufficient amounts of well-curated, accessible data on one hand, and a lack of a

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JTECH-D-20-0185.s1.

Corresponding author: Dion Häfner, dion.haefner@nbi.ku.dk

DOI: 10.1175/JTECH-D-20-0185.1

sophisticated analysis framework that handles nonlinearities and feature interactions on the other hand. In this study, we address the first issue and present the Free Ocean Wave Dataset (FOWD).

Particularly since the advent of machine-learning competitions—e.g., via the platform "Kaggle" (kaggle.com), where teams compete to find the best-performing machine-learning solutions to domain-specific problems—freely available, high-quality datasets have become an invaluable resource both as benchmarks for machine-learning researchers and as study objects for domain experts. Enabling easy access to domain-specific data allows even non–domain experts to participate in model building, to the benefit of the whole research community. We therefore also see this work as an important stepping-stone toward opening extreme wave research to a wider, potentially more machine-learning-literate, audience.

While we will be using rogue waves as a motivating example throughout this publication, other researchers can and should of course use FOWD to study phenomena other than extreme wave/crest heights (e.g., wave steepness or characteristic shape). In essence, FOWD relates aggregated sea state parameters to individual wave measurements. Applications are therefore plentiful.

As a primary data source for this version of FOWD we use the Coastal Data Information Program (CDIP) buoy data catalog. CDIP is a buoy network consisting primarily of Datawell Directional Waverider buoys for wave monitoring around the coasts of the United States (see, e.g., Behrens et al. 2019). The CDIP catalog (as of November 2020) contains measurements at 161 locations along the west and east coasts of North America and U.S. overseas states and territories like Hawaii, Guam, Puerto Rico, and the Marshall Islands.

Section 2 describes FOWD in detail, particularly which parameters are included, how they are computed, and which quality control processes we employ to validate the results. Section 3 outlines our Python reference implementation that allows us to efficiently process massive amounts of raw data, and section 4 describes the processing of the CDIP buoy data catalog. Section 5 gives an example application in which we look at how rogue wave probabilities vary depending on various sea state parameters. Section 6 gives a summary and conclusive remarks.

The FOWD–CDIP dataset is freely available for download (https://doi.org/10.17894/ucph.c589422c-64fd-4585-af31-4571497bcbe5; see also the data availability statement).

2. The FOWD specification

At its core, FOWD describes a mechanism to process raw observations (elevation time series and, optionally, directional spectra) into a catalog that maps parameters describing the current sea state x to observed wave or crest parameters y.

By "wave" we denote the series of surface elevations (relative to the 30-min mean elevation) from a given zero upcrossing to the next zero upcrossing. The crest and trough are then the maximum and minimum elevation of the wave, respectively, and the wave height is the sum of its crest height and trough depth. Some waves might be excluded by quality control criteria; see section 2c.

Throughout this study, we characterize extreme waves on the basis of their abnormality index AI = H/H_S, with wave height H and spectral significant wave height $H_S = 4(m_0)^{1/2}$, where m_0 is the zeroth moment of the spectral density [see also section 2a(2)].

FOWD output files are in netCDF4 format, which is widely used throughout the sciences and allows additional metadata to be attached. Every row in the resulting netCDF4 file represents a single wave and the sea state in which it was recorded.

Section 2a introduces the various quantities included in FOWD output and gives a more in-depth description of the computation of some parameters (where estimation is nonobvious or ambiguous). Section 2b describes the running-window processing approach we use in FOWD. Section 2c lists our quality control (QC) criteria, and section 2d outlines the steps we take to ensure reproducibility of FOWD output files.

a. Computed quantities

We group all output quantities into four categories:

1) *Station metadata* are anything that is specific to the sensor (and is not directly related to waves or the sea state). This includes both metadata describing the raw data source (to ensure reproducibility; more in section 2d) and the conditions in which it was recorded (latitude/longitude and water depth).
2) *Wave-specific parameters* are all quantities that describe a single wave, such as wave height or maximum slope. A typical study using FOWD aims to determine how a wave-specific parameter depends on one or several sea state parameters.
3) *Aggregated sea state parameters* describe the circumstances in which each wave occurred; that is, they relate to the past sea state of each wave. They are computed from the immediate 10- and 30-min history prior to (but not including) the current wave (see also section 2b for more on this running-window approach). Quantities are computed using only the raw sea surface elevation as input (either directly or by computing a spectrum first).
4) *Directional sea state parameters*: Some sensors (like the CDIP buoys) might include additional directional information that is not computable from the raw surface elevation time series. When such directional information (in form of a directional spectrum) is given, FOWD computes some directional parameters from it and includes them in the output. Note that this does *not* use the same running-window approach as the aggregated sea state parameters. Instead, each wave is mapped to the nearest (in time) available directional measurement. I.e., directional information usually includes some information relating to the *future* of the wave. But since directional information is robust to the influence of individual extreme events, we do not consider this a problem.

A complete overview of all computed quantities is shown in Table A1 in the appendix. Here, we outline some important quantities (as suggested in literature) and how they are estimated from the observed time series.

1) Space–time domain transformations

Since FOWD only processes (one dimensional) point measurements, we need some mechanism to transform information from the time domain back to the spatial domain. We relate frequencies f to wavenumbers k (and by extension, periods to wavelengths) through the dispersion relation for linear waves:

$$f^2 = \frac{gk}{(2\pi)^2}\tanh(kD), \tag{1}$$

with water depth D and gravitational acceleration $g = 9.81\,\mathrm{m\,s^{-2}}$. This also assumes the absence of currents.

To determine the wavenumber for a given frequency, we use an approximate inverse of (1) as given in Fenton (1988):

$$k \approx \frac{\alpha + \beta^2\cosh^{-2}\beta}{D(\tanh\beta + \beta\cosh^{-2}\beta)}, \tag{2}$$

with

$$a = (2\pi f)^2\frac{D}{g} \quad \text{and} \tag{3}$$

$$\beta = \frac{\alpha}{\sqrt{\tanh\alpha}}. \tag{4}$$

2) Spectral density estimation

To compute spectral quantities, we need to estimate the spectral density $\mathcal{S}(f)$ from the raw surface elevation time series. There is no unique way to do this, and any given method is a trade-off between spectral resolution, bias, and variance (noise).

In FOWD, we chose to use Welch's method (Welch 1967) with a window length of 180 s and a window overlap of 50% using a Hann window (also known as a Hanning window). This corresponds to about 230 measurements per segment in the case of CDIP data with sampling frequency 1.28 Hz. This implies that the 30-min spectra are an average of 20 individual segments and the 10-min spectra are an average of 7 segments. All segments are zero padded to the next highest power of 2. This gives a spectral resolution of 0.005 Hz and a maximum (Nyquist) frequency of 0.64 Hz for 1.28-Hz CDIP data.

We can then compute moments of $\mathcal{S}$ by integrating

$$m_n = \int_0^\infty f^n\mathcal{S}(f)\,df. \tag{5}$$

We numerically approximate all integrals in FOWD through a trapezoidal rule (with second-order accuracy).

3) Wave period and steepness

There are several popular approaches to define a dominant wave period for a given sea state. Depending on the application, either peak period, spectral mean period, or mean zero-crossing period may be more appropriate. Also, since we only have access to a noisy estimate of the true spectral density $\mathcal{S}$, some ways to compute the mean period from the spectrum are more accurate than others, depending, for example, on the frequency resolution of the sensor.

Therefore, we include several estimates of dominant wave period/frequency in FOWD:

$$\text{spectral peak period} \quad \overline{T}_p = \frac{\int_0^\infty \mathcal{S}(f)^4\,df}{\int_0^\infty f\mathcal{S}(f)^4\,df}, \tag{6}$$

$$\text{mean zero-crossing period (spectral)} \quad \overline{T}_{s,0} = \sqrt{m_0/m_2}, \quad \text{and} \tag{7}$$

$$\text{mean zero-crossing period (direct)} \quad \overline{T}_{d,0} = \frac{1}{N}\sum_{i=0}^{N} t_i, \tag{8}$$

where t_i refers to the zero-crossing periods of all waves in the corresponding surface elevation slice (zero crossings determined by linear interpolation) and the expression for $\overline{T}_p$ is taken from Young (1995).

For the characteristic wave steepness ϵ we use the peak wavenumber k_p, approximated from the peak period (6) and dispersion relation (1), following Serio et al. (2005):

$$\epsilon = \sqrt{2m_0}k_p. \tag{9}$$

4) Spectral bandwidth and Benjamin–Feir index

The computation of spectral bandwidth follows Serio et al. (2005). As is the case with wave period, there is more than one way to estimate spectral bandwidth from data; in fact, there are at least three common quantities:

$$\text{broadness} \quad \sigma_B = \sqrt{1 - \frac{m_2^2}{m_0 m_4}},$$

$$\text{narrowness} \quad \sigma_N = \sqrt{\frac{m_0 m_2}{m_1^2} - 1}, \quad \text{and}$$

$$\text{peakedness} \quad \sigma_Q = \frac{m_0^2}{2\sqrt{\pi}}\left[\int_0^\infty f\mathcal{S}(f)^2\,df\right]^{-1}. \tag{10}$$

Some authors also refer to peakedness as "quality factor."

Broadness is problematic because of the occurrence of m_4, the fourth moment of the spectral density $\mathcal{S}$. Because of the f^4 term occurring in its estimation, broadness is extremely sensitive to the high-frequency tail of $\mathcal{S}$, which renders it an unacceptably noisy quantity at lower sampling rates (such as CDIP's 1.28 Hz). Therefore, FOWD only includes narrowness and peakedness as spectral bandwidth estimates.

The Benjamin–Feir index (BFI) was introduced in Janssen (2003) and is a central parameter quantifying the strength of nonlinear interactions. Following Serio et al. (2005), we compute the BFI from steepness ϵ, bandwidth σ (which could be any of the three definitions above), peak wavenumber k_p, and depth D as

$$\mathrm{BFI} = \frac{\epsilon\nu}{\sigma}\sqrt{\max\{\beta/\alpha, 0\}}, \tag{11}$$

with

$$\nu = 1 + \frac{2k_pD}{\sinh(2k_pD)}, \tag{12}$$

$$\alpha = 2 - \nu^2 + 8(k_pD)^2\frac{\cosh(2k_pD)}{\sinh^2(2k_pD)}, \quad \text{and} \tag{13}$$

$$\beta = \frac{8 + \cosh(4k_pD) - 2\tanh^2(k_pD)}{8\sinh^4(k_pD)} - \frac{\left[2\cosh^2(k_pD) + \frac{\nu}{2}\right]^2}{\sinh^2(2k_pD)\left[\frac{k_pD}{\tanh(k_pD)} - \frac{\nu}{2}\right]^2}. \tag{14}$$

In FOWD, we compute the BFI twice, with spectral bandwidth σ estimated through both narrowness and peakedness [as defined in (10)].

5) Crest–trough correlation

Tayfun (1990) suggests another key parameter to describe wave height distributions, the correlation coefficient r between squared crest height A_0^2 and squared trough depth A_1^2, which we refer to as ‘‘crest–trough correlation.’’ This parameter r is closely related to spectral bandwidth (as, for narrowband seas, crests and troughs are approximately of the same size, becoming increasingly chaotic/uncorrelated as more harmonics are added). By extension, it is also a measure for the tendency of the sea state to form wave groups (Fig. 1).

The estimation of crest–trough correlation from the spectral density $\mathcal{S}$ is further elaborated in Tayfun and Fedele (2007). Following these lines, we compute r via

$$r = \frac{1}{m_0}\sqrt{\rho^2 + \lambda^2}, \tag{15}$$

with

$$\rho = \int_0^\infty \mathcal{S}(\omega)\cos\left(\omega\frac{\overline{T}}{2}\right)d\omega \quad \text{and} \tag{16}$$

$$\lambda = \int_0^\infty \mathcal{S}(\omega)\sin\left(\omega\frac{\overline{T}}{2}\right)d\omega, \tag{17}$$

where $\overline{T} = m_0/m_1$ is the spectral mean period and $\omega = 2\pi f$ is the angular frequency.

6) Spectral partitioning

To characterize processes that act mostly on short or long waves, spectral energy content is often more indicative than quantities based on the whole spectrum (such as mean period). Therefore, FOWD includes the relative energy content $\mathcal{E}$ over several spectral bands, computed as a definite integral over the spectral density $\mathcal{S}$:

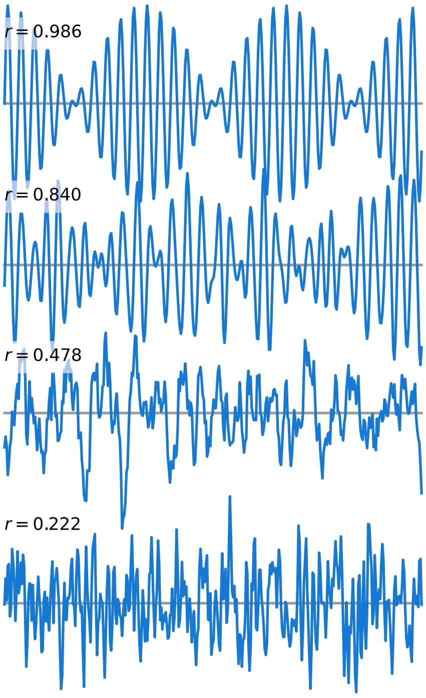


FIG. 1. The crest–trough correlation r is higher in ‘‘groupy,’’ low-bandwidth sea states. Shown are surface elevations generated from Ochi–Hubble spectra (Ochi and Hubble 1976) with increasing spectral bandwidth (from top to bottom) and the corresponding value of r.

$$\mathcal{E}_i = \frac{\int_{f_i}\mathcal{S}(f)\,df}{\int_0^\infty \mathcal{S}(f)\,df} = \frac{1}{m_0}\int_{f_i}\mathcal{S}(f)\,df. \tag{18}$$

We use five distinct spectral bands (with limits f_i), each characteristic for a different physical regime (Table 1). [This is a crude way to perform spectral partitioning as compared with more-sophisticated approaches that take directionality into account (Portilla-Yandún et al. 2016; Portilla-Yandún 2018). However, this simple integral is straightforward to compute and interpret, and can be estimated using only a surface displacement time series].

Similarly to the relative energy content, we also compute the total energy density contained in each frequency band (in joules per meter squared):

$$P_i = \rho g\int_{f_i}\mathcal{S}(f)\,df, \tag{19}$$

with approximate density of seawater $\rho = 1024\,\mathrm{kg\,m^{-3}}$ and gravitational acceleration $g = 9.81\,\mathrm{m\,s^{-2}}$.

7) Angular integrals

To make it possible to investigate the dependence of waves on phenomena like swell-wind sea crossing angles, we also split directional quantities into five distinct frequency bands, analogously to spectral energy content (Table 1). Since directional spread and wave direction are measured as an angle, we need

TABLE 1. Frequency bands used by FOWD and their approximate corresponding physical regime [as, e.g., given in Holthuijsen (2010)]. Here, and elsewhere ID is identifier.

Band ID	Frequency range	Corresponding wave regime
1	<0.05 Hz	Tides and seiches
2	0.05–0.1 Hz	Swell
3	0.1–0.25 Hz	Long-wave wind sea
4	0.25–1.5 Hz	Short-wave wind sea
5	0.08–0.5 Hz	Entire local wind sea

to take special care when averaging these quantities. Furthermore, we want to weight the directional value at each frequency with the corresponding spectral energy at that frequency, to ensure that the resulting average represents the dominant angle within this frequency band.

To achieve this, we compute the integral of a directional quantity q (which can be either dominant direction or directional spread) component-wise in Cartesian coordinates, weighted with the spectral density $\mathcal{S}$:

$$\overline{x} = \int_{f_i} \mathcal{S}(f) \sin q(f)\, df \quad \text{and} \tag{20}$$

$$\overline{y} = \int_{f_i} \mathcal{S}(f) \cos q(f)\, df, \tag{21}$$

where f_i again demarcates the boundaries of each frequency band. Then we transform the resulting Cartesian components back to an angle:

$$\overline{q} = \arctan(\overline{x}/\overline{y}), \tag{22}$$

which is the desired weighted angular average.

8) Directionality index

A key parameter to characterize the influence of directional spread on the wave dynamics is the "directionality index" R (as introduced in Fedele 2015). It is commonly defined as

$$R = \frac{\sigma_\theta^2}{2\nu^2}, \tag{23}$$

where σ_θ is the directional spread (in radians), and ν denotes the spectral bandwidth [we use narrowness, as in Fedele et al. (2019)]. This factor R makes it possible to compute various directionality-corrected versions of, for example, the Benjamin–Feir index and kurtosis (Fedele 2015; Fedele et al. 2019). In FOWD, we estimate R by computing the narrowness of the spectrum as provided by CDIP. Directional spread is computed as outlined above, which we integrate over all frequencies to obtain σ_θ.

b. *Running-window processing*

Usually, studies that investigate extreme wave observations divide all data into blocks of equal length in time, e.g., 30-min chunks, that are then analyzed separately (e.g., Casas-Prat and Holthuijsen 2010; Cattrell et al. 2018). However, the transient nature of the ocean has long been identified as a potential source for systematic error (Adcock and Taylor 2014; Gemmrich and Garrett 2011; Gemmrich et al. 2016), as it is not clear that the wave height distribution is constant within each chunk.

A related consideration is that the estimated quantities must be *agnostic of the future*—that is, look-aheads must be impossible. This property is critical for machine-learning applications, where future state leaking into the training data may completely invalidate the generalization abilities of a machine-learning algorithm.

We have therefore decided to use a running-window approach in FOWD. Here, we iterate through the raw data one zero-upcrossing at a time, computing the characteristic sea state parameters based on the immediate history of every wave. This implies that there is no time gap between the end of the aggregation period and the current wave, at the expense of additional computation time (since the sea state has to be recomputed for every wave).

Picking a window length is always a trade-off between bias (longer windows are more prone to nonstationarity) and variance (shorter windows leave us with less data with which to work). Therefore, all parameters are computed three times:

- The parameters are calculated twice using fixed 30- and 10-min windows. This makes it possible to investigate the stationarity of the current sea state by comparing the values obtained from each window length.
- The parameters are calculated one more time using a variable, data-dependent window as suggested in Boccotti (2000) and used in Fedele et al. (2019). We define the optimal window size n to be the one that minimizes

$$\mathrm{std}\left(\frac{\sigma_{n,i+1}}{\sigma_{n,i}} - 1\right), \tag{24}$$

where $\sigma_{n,i}$ is the standard deviation of the sea surface elevation in the ith chunk with length n, applied to the past 12 h of time series.

To make this process more robust, we recompute (24) 10 times for each candidate window with a different time offset. FOWD tries a total of 11 different windows lengths between 10 and 60 min and selects the one that minimizes the sum of (24) across all trials. This process tends to generate time windows longer than 40 min in most conditions but is also capable of reducing the window size if needed (Fig. 2).

Because the standard deviation of the sea surface elevation σ is directly related to significant wave height, we expect this to yield near-optimal window sizes for significant wave height and other slowly drifting quantities (such as mean period and energy content), but suboptimal results for faster drifting parameters (such as steepness, peak period, and kurtosis).

c. *Quality control*

FOWD uses a combination of QC flags, most of which are inspired by the process suggested in Christou and Ewans (2014). A measurement is discarded if any of the following conditions are met when applied to the past 30-min surface elevation:

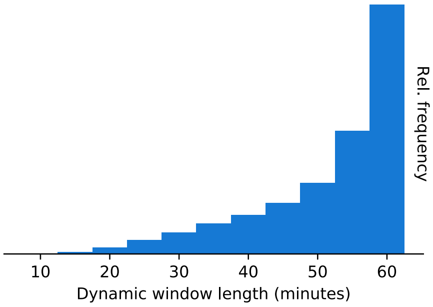


FIG. 2. Most dynamic windows are longer than 30 min. Shown is a histogram of the determined optimal window size across all Hawaiian CDIP stations.

1) There are any waves with zero-crossing period >25 s.
2) The rate of change of the surface elevation η exceeds the limit rate of change by a factor of 2 or more at any point; that is,

$$\left|\frac{\partial \eta}{\partial t}\right| > 2U_{\text{lim}}. \tag{25}$$

The limit rate of change U_{lim} is defined as

$$U_{\text{lim}} = 2\pi \frac{\text{std}(\eta)}{\langle T_{d,0}\rangle}\sqrt{2\ln N}, \tag{26}$$

with standard deviation std, mean observed zero-crossing periods $\langle T_{d,0}\rangle$, and number of waves in the record N. This criterion removes records containing waves that are much steeper than the average rate of change $\text{std}(\eta)/\langle T_{d,0}\rangle$—that is, records with single, very steep waves—but leaves sea states with many steep waves intact.
3) There are 10 consecutive data points of the same value.
4) There is any absolute crest or trough elevation that is greater than 8 times the normalized median absolute deviation (MADN) of the surface elevation; that is,

$$|h| > 8\kappa\, \text{median}[|\eta - \text{median}(\eta)|], \tag{27}$$

with $\kappa = 1.483$, which ensures that MADN converges to standard deviation for Gaussian distributed η with growing sample size (see, e.g., Huber and Ronchetti 2009). This criterion permits crest heights and trough depths of up to about 2 times the significant wave height, which should be more than enough for any real signal. [In a linear sea, a crest exceeding $2H_S$ would have a probability of $\exp(-32) \approx 10^{-14}$].
5) Surface elevations are not equally spaced in time (but they may contain "NaN" values).
6) The ratio of missing (NaN) data to valid data exceeds 5%.
7) There are less than 100 individual zero crossings.

All waves that fail QC and are larger than 2 times the significant wave height are written to a log file to allow for manual inspection. In addition, all waves that are larger than 2.5 times the significant wave height are written to the log file, regardless of whether they pass QC. This enables us to evaluate the QC process and tweak thresholds or exclude faulty subdatasets as needed. A brief evaluation of this QC process when applied to the CDIP data is given in section 4b.

d. *Additional metadata and reproducibility*

All FOWD output files are self-documenting in the sense that they include all relevant metadata as netCDF4 attributes, both for each variable and the dataset as a whole. Apart from the static metadata documenting the coordinates and parameters (which is the same for every FOWD output file), we also include some metadata related to the processing environment and raw data source to ensure reproducibility. Specifically, each wave record includes the time stamp, file name, and a unique file identifier (UUID) of the raw source file from which it came (see Table A1). The output files also include the exact version of the FOWD processing implementation used to create the file in form of a "git" tag, along with a UUID. That way, we enable users to reproduce any result by allowing them to use the exact same processing version and input file.

3. Reference implementation

As part of this work, we supply a Python reference implementation of the FOWD processing toolkit. It makes use of the popular Python packages xarray, numpy, and scipy to process large amounts of input data efficiently. The implementation processes either CDIP netCDF4 files or generic input files in a fixed netCDF4 format. Multiple CDIP deployments (within the same station) can be processed in parallel.

a. *Memory efficiency*

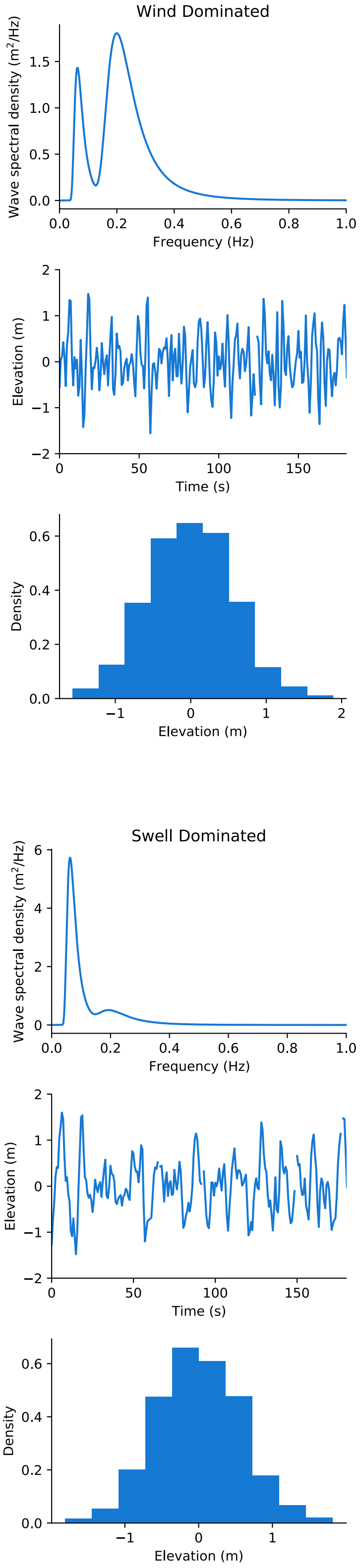
Because of FOWD's running-window approach (see section 2b), FOWD output datasets are about 10 times as big as the input surface elevation time series (since every wave results in about 80 output features). This demands that the processing implementation does not store entire output files in memory.

We achieve this by keeping only the immediate 30-min history of the current processing time in memory. Each new record is flushed to disk using Python's "pickle" format. After the processing has finished, these pickle files are read back by the main process in chunks, reformatted to the netCDF4 output format, and flushed to disk again. This ensures that the main process uses only a negligible amount of memory while each worker process only keeps the input data in memory. In other words, if the input data fit in memory, processing will succeed.

b. *Testing strategy*

In software engineering, automated tests are an invaluable tool to ensure proper functionality of a product. Unfortunately, writing automated tests for processing workflows of physical data is often impossible or infeasible because of the lack of ground-truth answers with which to compare. On the other hand, faulty results are often easy to detect for humans when they fall outside of reasonable physical limits or show the wrong scaling behavior. We have therefore opted for semiautomated *sanity checks* instead of fully automated unit tests for the core processing.

Each sanity check test case generates a random surface elevation time series from a different ground-truth wave spectrum



⇒

```
{
    "estimated_sea_state": {
        "bandwidth_narrowness": 0.489,
        "bandwidth_peakedness": 0.404,
        "benjamin_feir_index_narrowness": 0.25,
        "benjamin_feir_index_peakedness": 0.302,
        "crest_trough_correlation": 0.415,
        "energy_in_frequency_interval": [
            37.275, 505.533, 1713.038, 916.344, 2701.6
        ],
        "kurtosis": -0.134,
        "maximum_wave_height": 2.958,
        "mean_period_direct": 4.422,
        "mean_period_spectral": 4.157,
        "peak_wave_period": 5.099,
        "peak_wavelength": 40.59,
        "rel_energy_in_frequency_interval": [
            0.012, 0.159, 0.54, 0.289, 0.852
        ],
        "rel_maximum_wave_height": 1.316,
        "significant_wave_height_direct": 2.052,
        "significant_wave_height_spectral": 2.248,
        "skewness": 0.063,
        "steepness": 0.123,
        "valid_data_ratio": 0.99
    },
    "spectral_parameters": {
        "peak_period_swell": 16,
        "peak_period_wind": 5,
        "shape_swell": 1,
        "shape_wind": 1,
        "swh_swell": 1,
        "swh_wind": 2
    },
    "water_depth": 500
}
```

⇒

```
{
    "estimated_sea_state": {
        "bandwidth_narrowness": 0.755,
        "bandwidth_peakedness": 0.421,
        "benjamin_feir_index_narrowness": 0.018,
        "benjamin_feir_index_peakedness": 0.032,
        "crest_trough_correlation": 0.569,
        "energy_in_frequency_interval": [
            141.391, 1967.554, 856.077, 267.045, 1643.
        ],
        "kurtosis": -0.1,
        "maximum_wave_height": 3.287,
        "mean_period_direct": 7.514,
        "mean_period_spectral": 6.791,
        "peak_wave_period": 14.956,
        "peak_wavelength": 349.251,
        "rel_energy_in_frequency_interval": [
            0.044, 0.609, 0.265, 0.083, 0.508
        ],
        "rel_maximum_wave_height": 1.449,
        "significant_wave_height_direct": 2.055,
        "significant_wave_height_spectral": 2.269,
        "skewness": 0.048,
        "steepness": 0.014,
        "valid_data_ratio": 0.99
    },
    "spectral_parameters": {
        "peak_period_swell": 16,
        "peak_period_wind": 5,
        "shape_swell": 1,
        "shape_wind": 1,
        "swh_swell": 2,
        "swh_wind": 1
    },
    "water_depth": 500
}
```

FIG. 3. Sanity check test cases allow us to verify manually that computed parameters are reasonable. Shown are test (left) inputs and (right) outputs for (top) high-frequency and (bottom) low-frequency seas. Estimated sea state parameters are defined in Table A1. Spectral parameters are input parameters of the Ochi–Hubble spectrum used to generate each test case (as shown in upper-left panels).

and runs it through the FOWD processing. Here, only the spectral shape is prescribed externally, surface elevations are drawn as harmonics with random phases from the spectrum. The resulting output parameters can then be inspected manually.

Two example sanity check spectra are bimodal Ochi–Hubble spectra (Ochi and Hubble 1976) that are either swell dominated (low-frequency peak is dominant) or wind dominated (high-frequency peak is dominant). We would expect that the wind dominated spectrum leads to lower period, higher steepness and BFI, and shorter wavelength. In both cases, we expect to find a spectral significant wave height of

$$\mathrm{SWH}_{\mathrm{total}} = \sqrt{\mathrm{SWH}^2_{\mathrm{swell}} + \mathrm{SWH}^2_{\mathrm{wind}}} \qquad (28)$$

and excess kurtosis and skewness around 0. Directly estimated significant wave height $H_{1/3}$ is usually slightly lower than its spectral counterpart H_{m_0}, and vice versa for wave period.

Indeed, all of these expectations are met for this particular test case (Fig. 3). Other sanity checks feature idealized spectra, for example, containing just a single harmonic, that allow us to validate parameters that are more difficult to interpret like crest–trough correlation, or idealized directional spectra. Because of these sanity checks, we are confident that the FOWD core processing produces meaningful results.

4. Processing of CDIP buoy data

The following sections describe the CDIP input and FOWD output data, analyze QC performance and the impact of FOWD's running-window processing, and discuss some caveats that apply when using buoy data for extreme wave studies.

a. Input data and processing

In total, the CDIP catalog spans about 750 years of continuous surface elevation measurements (almost all at sampling rates of 1.28 Hz) and is available in netCDF4 format through a THREDDS server. This amounts to about 270 GByte of raw data.

While CDIP data files also include horizontal displacements and a number of derived quantities (like significant wave height, peak period, and others), we use only the raw vertical surface displacement, station metadata, and directional quantities for processing. This ensures that FOWD is applicable to any instrument that delivers a surface displacement time series (including radar or laser sensors).

We applied only minimal preprocessing to the data, which consists of removing all data that have an error flag set and subtracting the 30-min running mean from the raw vertical surface elevation. After that, we processed all data in about 72 h on 10 cluster nodes in parallel (using the FOWD reference implementation described in section 3). The resulting output dataset has a total (compressed) size of 1.1 TB. We create one output file per CDIP station, with individual file sizes ranging between 1.7 MByte and 38 GByte.

In total, FOWD contains about 4.2 billion individual waves and sea states. An interactive map indicating all data locations and some key statistics is available in the online supplemental material.

TABLE 2. The number of times each QC flag was triggered for the whole CDIP catalog. See section 2c for a definition of flags a–g. Note that multiple flags can be active for the same wave.

Flag	Count
a	31 547
b	18 465
c	39 470
d	47 544
e	0
f	11 915
g	4089
Failed waves	77 371

b. Quality control and filtering

As outlined in section 2c, FOWD automatically logs waves failing QC that are higher than 2 significant wave heights, and all waves higher than 2.5 significant wave heights (whether they pass QC or not). This allows us to assemble some higher-order statistics to get an idea of how prevalent quality issues are in the CDIP data and to verify that FOWD's QC system works as intended.

In total, just under 80 000 waves fail QC (Table 2). About 80% of these QC failures occur in only 5 CDIP locations (of 161). This suggests that relatively few deployments with general quality problems cause a majority of QC failures.

To investigate this further and isolate faulty deployments, the FOWD implementation includes a postprocessing command that produces plots of all records in the QC logs. These

TABLE 3. Blacklisted CDIP deployments that failed visual inspection.

CDIP ID	Excluded deployments
045p1	d01, d02, d03, d13, d15, d17, d19, d21
094p1	d01, d02, d03, d04, d05
096p1	d04
100p1	d11
106p1	d02
109p1	d05, d06
111p1	d06
132p1	d01
141p1	d03
142p1	d02, d15, d18
144p1	d01
146p1	d01, d02
158p1	d02, d04
162p1	d07
163p1	d01, d05
167p1	d01
172p1	d01
177p1	All deployments
196p1	d04
201p1	d03
205p1	All deployments
206p1	All deployments
261p1	All deployments
430p1	d06
431p1	d02

TABLE 4. Characteristic scale used to normalize root-mean-square residual for each parameter (Fig. 4).

Parameter	Typical range	Resulting scale
sea_state_30m_bandwidth_peakedness	0–0.6	0.6
sea_state_30m_benjamin_feir_index_peakedness	0–0.6	0.6
sea_state_30m_crest_trough_correlation	0.2–1.0	0.8
sea_state_30m_kurtosis	From −0.5 to 1.5	2.0
sea_state_30m_mean_period_direct	4–15 s	11 s
sea_state_30m_mean_period_spectral	4–15 s	11 s
sea_state_30m_peak_wave_period	4–20 s	16 s
sea_state_30m_peak_wavelength	0–600 m	600 m
sea_state_30m_rel_energy_in_frequency_interval_1	0–0.2	0.2
sea_state_30m_rel_energy_in_frequency_interval_2	0–1	1
sea_state_30m_rel_energy_in_frequency_interval_3	0–1	1
sea_state_30m_rel_energy_in_frequency_interval_4	0–0.4	0.4
sea_state_30m_rel_energy_in_frequency_interval_5	0–1	1
sea_state_30m_rel_maximum_wave_height	1.2–2.2	1
sea_state_30m_significant_wave_height_direct	0.5–8.0 m	7.5 m
sea_state_30m_significant_wave_height_spectral	0.5–8.0 m	7.5 m
sea_state_30m_skewness	From −0.5 to 0.5	1
sea_state_30m_steepness	0–0.12	0.12

plots show the raw surface elevation of the failing wave and its immediate 30-min history.

After inspecting each of these plots, we decided to blacklist 38 deployments and 4 entire CDIP stations that showed obvious quality problems like frequent spikes, extreme oscillations, unphysical values, or jumps (Table 3). On top of excluding these blacklisted CDIP deployments, we also removed all records in conditions in which buoys are known to be unreliable [similar to McAllister and van den Bremer (2020)]:

1) records with 30-min significant wave height smaller than 1 m,
2) records with spectral mean frequency higher than 1/3.2 of the Nyquist frequency (for 1.28-Hz data, this is equivalent to filtering all records with a mean wave period below 5 s), and
3) records where the relative energy content of frequency band 1 exceeds 10% (extensive low-frequency drift).

After filtering, the final dataset contains about 1.4 billion waves and sea states (about 67% filtered, most due to the minimum significant wave height requirement).

Since FOWD is also intended for use by non–wave experts, it is essential to provide access to a precleaned dataset. Therefore, the filtered FOWD–CDIP dataset is available for download along with the unfiltered one (see the data availability statement).

c. Impact of running-window processing

After processing the CDIP data, we can now investigate how large of a difference FOWD's running-window processing (as described in section 2b) makes in practice, relative to the usual fixed-window approach.

To this end, we divide the FOWD catalog for one particular CDIP station (with ID 188p1, containing about 30 million waves) into 30-min chunks. The last measurement in each of these chunks (concerning the past 30-min sea state) then represents what would have been obtained for all waves if FOWD did not use running windows.

We can then quantify the influence of the running-window approach by computing the root-mean-square (RMS) difference between this last measurement of every chunk and all other data points in it. To make it easier to compare the different parameters, we divide each by a characteristic scale to obtain a normalized RMS (Table 4).

The resulting distribution of the normalized RMS in each chunk shows that, while deviations are typically below 10% of the characteristic scale, they can reach up to 50% in extreme cases (Fig. 4). As expected, some parameters (such as kurtosis and maximum wave height) are much more prone to drift than others (such as significant wave height and spectral energy). However, this result is sensitive to which characteristic scale we choose, so comparisons between parameters remain qualitative.

A particularly important quantity in this context is the significant wave height. If the significant wave height is underestimated with an error of only 5%, a wave with true abnormality index AI = 2 is estimated as a wave with AI = 2.1, which is less than one-half as likely to occur (assuming Rayleigh-distributed waves).

We conclude that the running-window approach *can* lead to significantly different results, apart from the more important effect of preventing look-aheads (as discussed in section 2b). In other words, explicitly accounting for a drifting sea state provides an opportunity to reduce bias by a nontrivial amount—although we did not measure how much this approach influences final results or conclusions.

d. Shortcomings of buoy data

Although any dataset that provides surface elevation measurements can be processed into a FOWD dataset, buoy measurements remain a dominant data source due to their relatively large availability (at least in comparison with radar and laser measurements). Therefore, this section discusses some

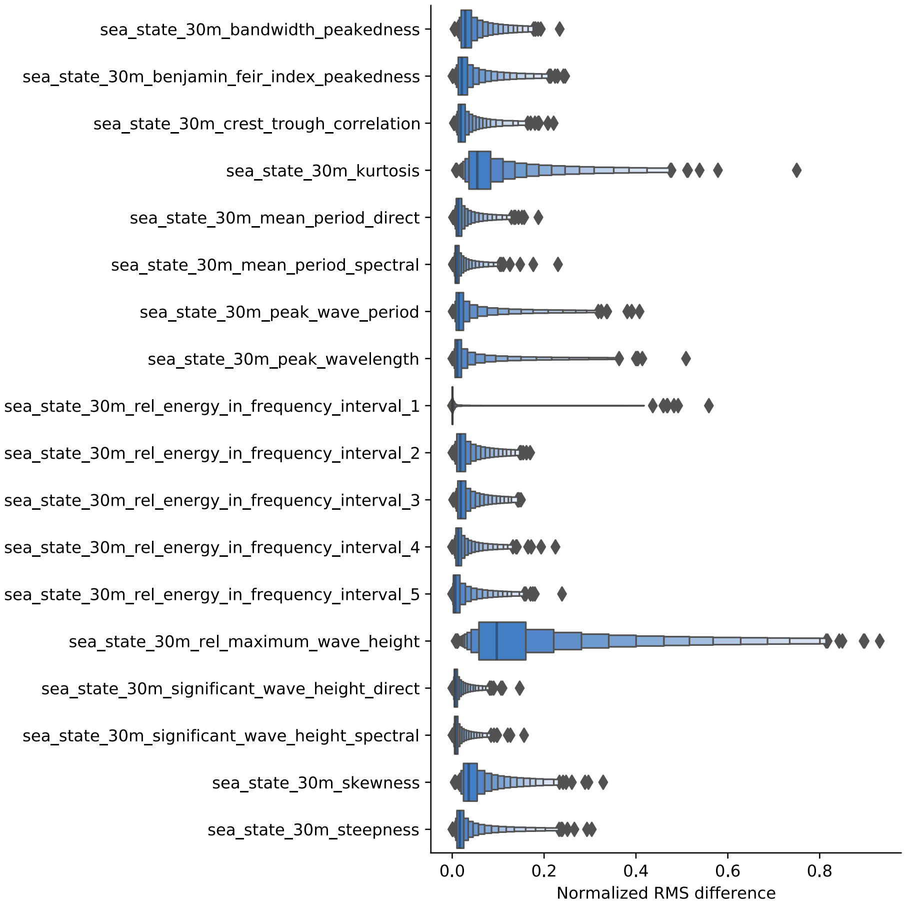


FIG. 4. In extreme cases, using running windows (instead of fixed chunks) leads to RMS differences of up to 50% of the characteristic scale of a parameter (Table 4). Shown is the distribution of normalized RMS difference between processing based on running windows and fixed chunks for some parameters.

of the known problems with buoy data, and how they carry over to FOWD and its possible applications.

First and foremost, buoys tend to linearize surface elevations to some degree [see McAllister and van den Bremer (2020, 2019) for a discussion]. This is especially problematic in rough seas with high steepness, because buoys can be dragged through a steep crest or move laterally around it and underestimate the true wave height. Combined with the inherent sampling variability of a point measurement (the two-dimensional wave has to hit the buoy at the crest to be registered at full height; see Benetazzo et al. 2015), wave estimates based on buoy data tend to be too conservative (see also Casas-Prat and Holthuijsen 2010).

This is inconvenient for studies with the goal to estimate absolute rogue wave risk, since one needs to take additional steps to correct for these biases, include other data sources, or accept that the results represent a lower bound for rogue wave risk. However, this is not a problem when estimating the *relative* importance of sea state risk factors, as buoys should be similarly inaccurate across a wide range of different sea states (after the most problematic conditions are filtered; see section 4b—perhaps with the exception of very steep seas). We therefore see no problem with using buoy data for the type of study presented in section 5.

Another issue to keep in mind is *selection bias*. Buoys tend to be placed in locations that are easy to reach and of special interest for humans. This implies that coastal areas are overrepresented, and therefore results derived from the whole dataset will be less representative for open-ocean conditions.

No reasonable amount of one-dimensional time series data can tell us about truly exceptional events. In offshore engineering contexts, an important quantity is the ‘‘10 000 year wave,’’ which is the largest expected wave in a 10 000 yr period. Events of this rarity cannot be estimated with this dataset

TABLE 5. Number of waves in the FOWD–CDIP dataset fulfilling various criteria.

Waves with AI < 2	1 383 488 167
Waves with AI ≥ 2	82 058
Waves with AI ≥ 2.2	11 849
Waves with AI ≥ 2.5	564
Waves with AI ≥ 2 within 30 s	2455

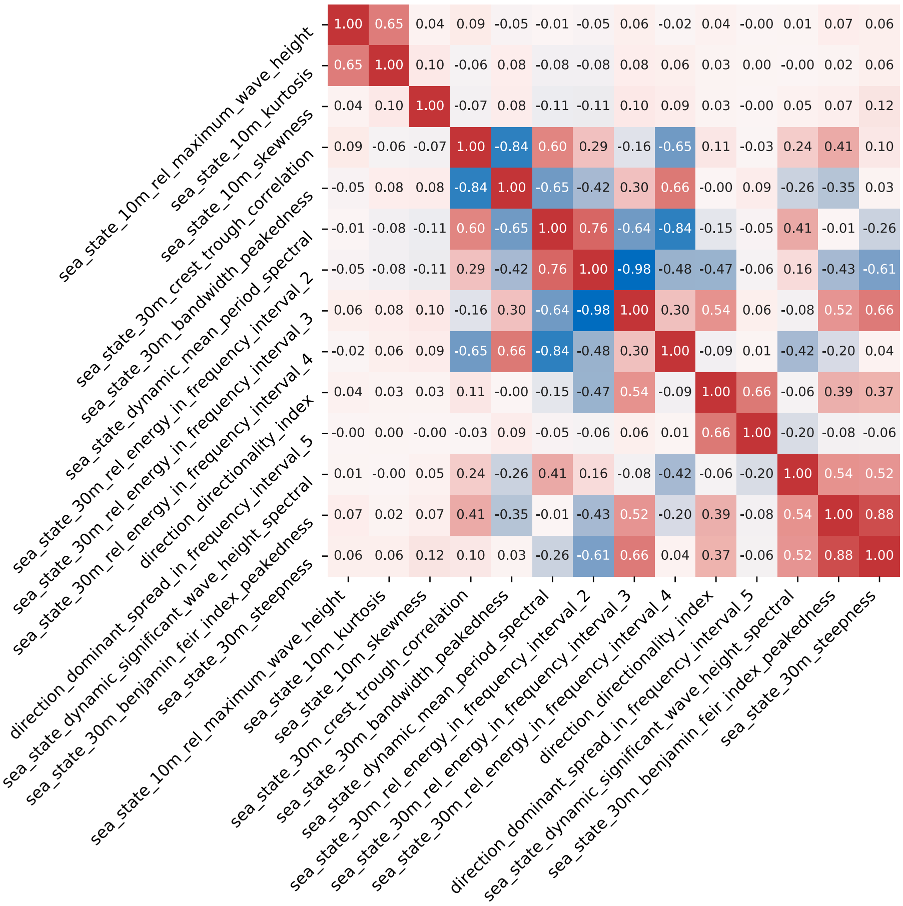


FIG. 5. Linear (Pearson) correlation matrix of selected parameters. Almost all parameters are strongly correlated with at least one other parameter, but exceptions exist (e.g., skewness, kurtosis/maximum wave height, and wind sea directional spread).

without additional work (such as further theoretical assumptions, or data augmentation via simulations).

5. Example application: Which sea state parameter is the best predictor for rogue wave occurrence?

As an example of an application of FOWD, we look at the connection between sea state and the occurrence of rogue waves to find which sea state parameter is the best predictor for rogue wave activity (where we find the largest change in rogue wave probability when varying the parameter).

In this context, we define rogue waves as any wave whose height exceeds 2 times the significant wave height, i.e., AI > 2. For any given sea state with wave height distribution $P(\mathrm{AI})$ we would expect the next wave to be a rogue wave with probability

$$p = \int_2^\infty P(\mathrm{AI})\, d\mathrm{AI}. \tag{29}$$

From linear superposition of random waves with narrow spectral bandwidth (Longuet-Higgins 1952), we would expect this criterion to be fulfilled for roughly 1 in 3000 waves. In the filtered FOWD–CDIP dataset, this criterion is fulfilled for about 100 000 of 1.5 billion total waves (i.e., 1 in 15 000), with about 3% of all rogue waves occurring within seconds of one another (Table 5).

This implies that the measured incidence rate of rogue waves across all sea states is lower by about a factor of 5 than is predicted by linear theory. This is not uncommon for buoy data (Casas-Prat and Holthuijsen 2010) and could to some degree be due to the underestimation of extreme waves by buoys (as discussed in section 4d). However, we suspect that this has mostly physical causes. Effects like crest–trough correlations < 1 (as we will see below) or wave breaking can severely limit the formation of rogue waves and are not accounted for in linear theory.

During the following sections, we will take a closer look under which conditions rogue waves preferably occur. For this, we use the combined data from all Hawaiian CDIP stations (stations with IDs 098p1, 106p1, 146p1, 165p1, 187p1, 188p1, 198p1, 225p1, 233p1), containing about 200 million waves.

a. *Confounding and roguish sea states*

To get a feeling for the data, we investigate correlations between some of the sea state parameters and have a look at the probability density functions of sea states in which we find rogues with AI > 2 and AI > 2.4.

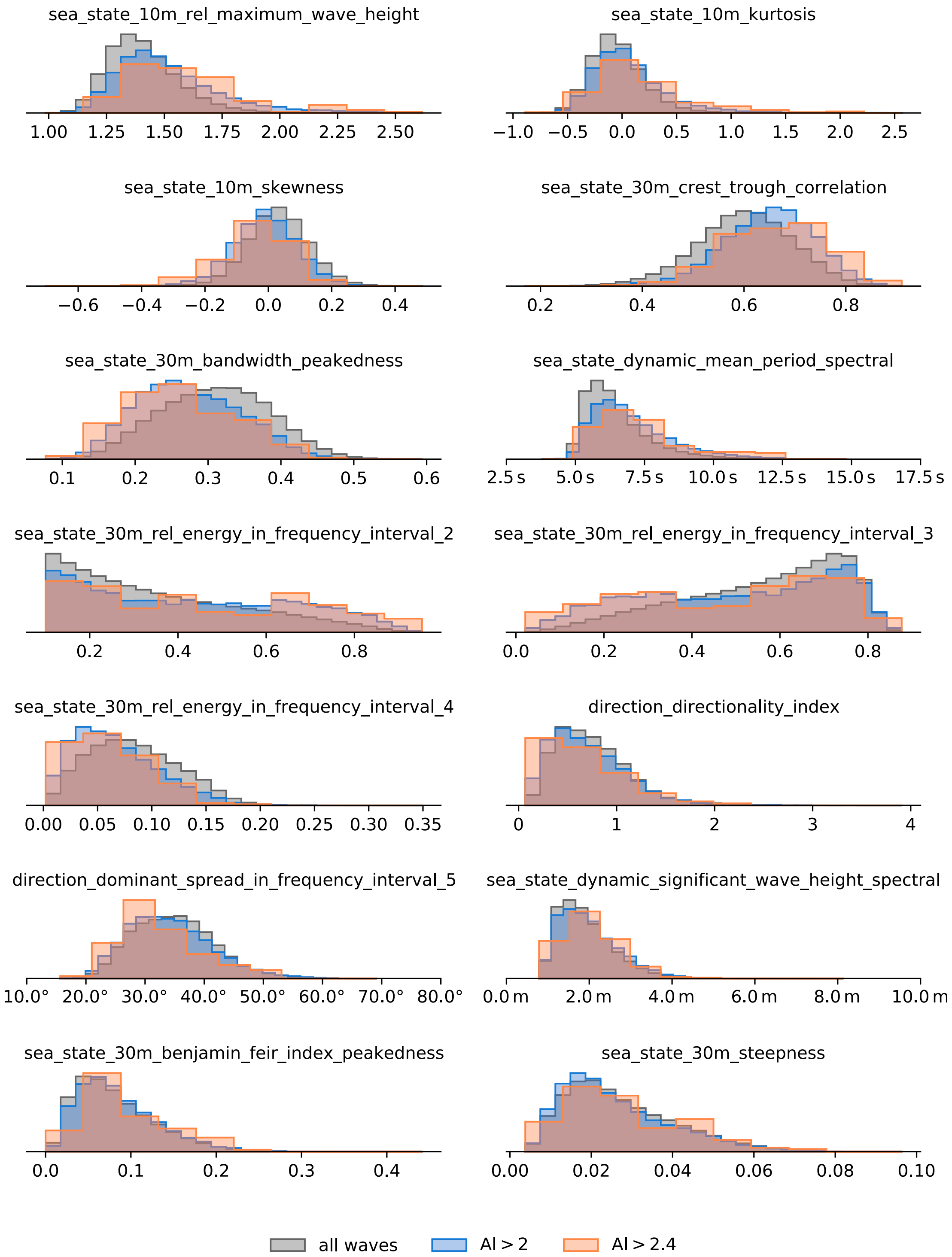


FIG. 6. Most parameters show a clear difference between the probability distributions of all sea states and those containing an extreme wave, but some just show a weak dependence (e.g., directional spread, significant wave height, and steepness). Shown are the probability density functions (PDFs) of various sea state parameters, estimated via histograms. Each parameter includes PDFs for the sea states of all waves, waves with AI > 2, and waves with AI > 2.4.

The correlation matrix of the sea state parameters (Fig. 5) provides yet another important sanity check for FOWD, since many parameters are correlated by definition (such as BFI, which is computed based on steepness and spectral bandwidth). Furthermore, it serves as an important reminder that there are many nonobvious correlations, such as the one between spectral bandwidth and mean period. Any conclusion we draw about the influence of a parameter on rogue wave activity thus has to take possible confounders into account.

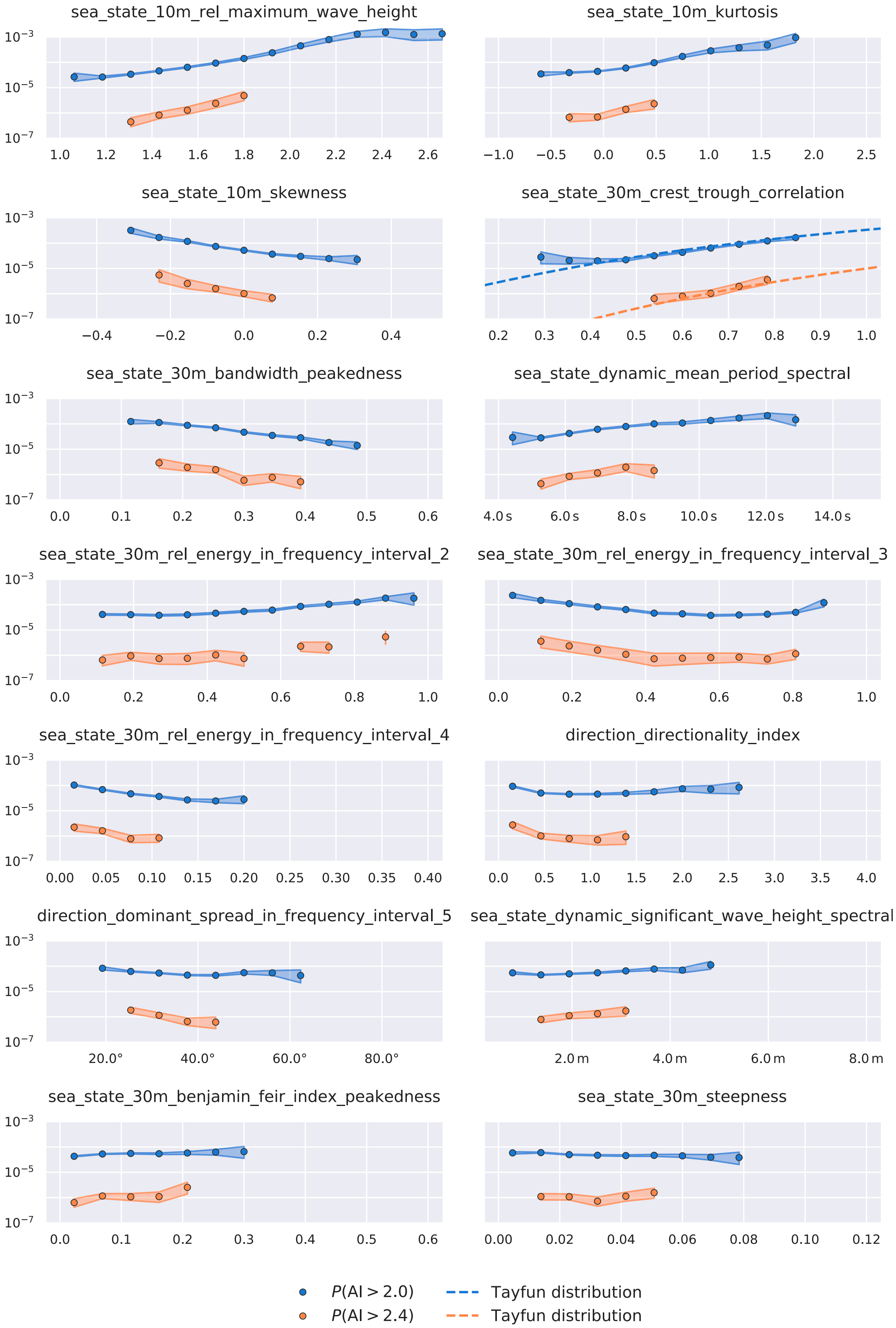


FIG. 7. Some sea state parameters are much more informative for rogue wave activity than others. Shown is the dependence of the rogue wave probability on several sea state parameters for AI $>$ 2 and AI $>$ 2.4. Symbols represent rogue wave probability posterior mean; shading represents the 95% minimum credible interval. Dashed lines indicate the values predicted by the Tayfun wave height distribution (Tayfun and Fedele 2007).

TABLE A1. All quantities included in FOWD output files. Quantities marked with a dagger are further explained throughout section 2a.

Name in output dataset	Description	Unit	Example value
	Station metadata		
meta_station_name	Name of original measurement station	—	CDIP_098p1
meta_source_file_name	File name of raw input data file	—	098p1_d01.nc
meta_source_file_uuid	UUID of raw input data file	—	CC54C8D5-7B1B-4170-9DBA-EBFD91F26F14
meta_deploy_latitude	Deploy lat of instrument	°N	21.4156
meta_deploy_longitude	Deploy lon of instrument	°E	−157.678
meta_water_depth	Water depth at deployment location	m	100.0
meta_sampling_rate	Measurement sampling frequency in time	Hz	1.28
meta_frequency_band_lower	Lower limit of frequency band	Hz	(0.0, 0.05, 0.1, 0.25, 0.08)
meta_frequency_band_upper	Upper limit of frequency band	Hz	(0.05, 0.1, 0.25, 1.5, 0.5)
	Wave-specific parameters		
wave_id_local	Incrementing wave ID for given station	—	11 726
wave_start_time	Wave start time	—	1218:44.220 000 000 10 Aug 2000
wave_end_time	Wave end time	—	1218:50.470 000 000 10 Aug 2000
wave_zero_crossing_period	Wave zero-crossing period relative to 30-m sea surface elev	s	5.644 304 276
wave_zero_crossing_wavelength†	Wave zero-crossing wavelength relative to 30-m sea surface elev	m	49.740 48
wave_raw_elevation	Raw surface elev relative to 30-m sea surface elev	m	(0.200 261, 0.889 527, 0.509 184, −0.550 564, −0.690 152, −0.270 083, −0.200 052)
wave_crest_height	Wave crest height relative to 30-m sea surface elev	m	0.889 527
wave_trough_depth	Wave trough depth relative to 30-m sea surface elev	m	−0.690 152
wave_height	Absolute wave height relative to 30-m sea surface elev	m	1.579 679
wave_ursell_number	Ursell no.	1	0.003 908
wave_maximum_elevation_slope	Max slope of surface elev in time	$\mathrm{m\,s^{-1}}$	0.921 658
	Aggregated sea state parameters		
sea_state_30m_start_time	Sea state aggregation start time	—	1148:45.000 999 936 10 Aug 2000
sea_state_30m_end_time	Sea state aggregation end time	—	1218:43.438 000 000 10 Aug 2000
sea_state_30m_significant_wave_height_spectral†	Significant wave height estimated from wave spectrum (Hm0)	m	1.798 395
sea_state_30m_significant_wave_height_direct	Significant wave height estimated from wave history (H1/3)	m	1.648 174
sea_state_30m_maximum_wave_height	Max wave height estimated from wave history	m	3.188 91
sea_state_30m_rel_maximum_wave_height	Max wave height estimated from wave history relative to spectral significant wave height	1	1.773 198
sea_state_30m_mean_period_direct	Mean zero-crossing period estimated from wave history	s	5.133 130 549
sea_state_30m_mean_period_spectral	Mean zero-crossing period estimated from wave spectrum	s	5.034 029 007
sea_state_30m_skewness	Skewness of sea surface elev	1	0.010 083
sea_state_30m_kurtosis	Excess kurtosis of sea surface elev	1	−0.076 898
sea_state_30m_valid_data_ratio	Ratio of valid measurements to all measurements	1	1.0
sea_state_30m_peak_wave_period†	Dominant wave period	s	6.841 089 249
sea_state_30m_peak_wavelength†	Dominant wavelength	m	73.070 08
sea_state_30m_steepness†	Dominant wave steepness	1	0.054 674
sea_state_30m_bandwidth_peakedness†	Spectral bandwidth estimated through spectral peakedness (quality factor)	1	0.312 186
sea_state_30m_bandwidth_narrowness†	Spectral bandwidth estimated through spectral narrowness	1	0.435 69

TABLE A1. (*Continued*)

Name in output dataset	Description	Unit	Example value
sea_state_30m_benjamin_feir_index_peakedness†	Benjamin–Feir index estimated through steepness and peakedness	1	0.164 307
sea_state_30m_benjamin_feir_index_narrowness†	Benjamin–Feir index estimated through steepness and narrowness	1	0.117 731
sea_state_30m_crest_trough_correlation	Crest–trough correlation parameter r estimated from spectral density	1	0.608 416
sea_state_30m_energy_in_frequency_interval†	Total energy density contained in frequency band	J m^{-2}	(1.935 885, 106.749 48, 1620.2413, 301.649, 1926.3574)
sea_state_30m_rel_energy_in_frequency_interval†	Relative energy contained in frequency band	1	(0.000 953, 0.052 571, 0.797 922, 0.148 553, 0.948 675)
Sea state parameters are repeated analogously for 10-min (_10m_) and dynamic (_dynamic_) window sizes			
Directional sea state parameters			
direction_sampling_time	Time at which directional quantities are sampled	—	1211:52.000 000 000 10 Aug 2000
direction_dominant_spread_in_frequency_interval†	Dominant directional spread in frequency band	°	(57.965 824, 38.118 546, 31.545 62, 39.302 81, 33.078 98)
direction_dominant_direction_in_frequency_interval†	Dominant wave direction in frequency band	°	(83.074, 136.024 32, 74.008 62, 77.266 02, 74.895 02)
direction_peak_wave_direction	Peak wave direction relative to normal-north	°	70.468 75
direction_directionality_index†	Directionality index R (squared ratio of directional spread and spectral bandwidth)	1	0.924 404

The probability density functions of roguish seas (Fig. 6) indicate several potential controlling parameters for rogue wave occurrence, where the distribution of seas containing a rogue wave differs substantially from that of all waves (with, e.g., skewness, spectral bandwidth, and maximum wave height being promising candidates). This analysis, while intuitively approachable, yields little quantitative insight into the relative importance of each parameter, and it neglects the influence of sample size effects. The following section addresses this through a simple analytical Bayesian parameter estimation.

b. *Estimation of rogue wave probabilities with uncertainties*

A major challenge when dealing with rare events like rogue waves is to determine whether there actually are enough data points to make a statement. We will therefore quantify this uncertainty through Bayesian credible intervals on the rogue wave probability p. As the first step, we assume that the occurrence of n^+ rogue waves and n^- nonrogue waves in a given sea state is drawn randomly with some rogue wave probability p. Then n^+ follows a binomial distribution:

$$n^+ \sim \mathrm{Binom}(n^+ + n^-, p). \tag{30}$$

The goal of this analysis is to estimate p from measurements of n^+ and n^-. For p, we encode prior information by assuming a beta prior, given by

$$p_{\mathrm{prior}} \sim \mathrm{Beta}(\alpha_0, \beta_0), \tag{31}$$

with parameters α_0 and β_0, which we choose as $\alpha_0 = 1$ and $\beta_0 = 10\,000$, roughly representing the expected order of magnitude $O(p) \approx 10^{-4}$ (this is just a weakly informative prior to constrain p to the right order of magnitude—the exact values have no influence on the conclusions of this analysis).

Applying Bayes's theorem,

$$P(p|X) = \frac{P(X|p)P(p)}{P(X)}, \tag{32}$$

we find the posterior of the rogue wave probability as

$$p \sim \mathrm{Beta}(n^+ + \alpha_0, n^- + \beta_0), \tag{33}$$

that is, another beta distribution (since the chosen beta prior for p is conjugate to the binomial likelihood of n^+).

This posterior is simple to evaluate analytically. In particular, we can use widely available library functions to compute the minimum credible interval (highest posterior credible interval) for p. This gives us the possibility to quantify our uncertainty in p based on the number of available samples, expressed as, for example, the 95% credible interval.

To finally investigate the influence of the sea state on the rogue wave probability p, we split each sea state parameter into 15 equally sized bins. We assume that, within each bin, p is independently and identically distributed (iid) with a distribution according to (33), and we evaluate the mean and credible interval of p independently for each bin. We also exclude bins that contain less than 10 rogue wave events (i.e., where $n^+ < 10$) to eliminate overly uncertain estimates. As a result, we can study how p behaves as a function of each sea state parameter and quantify our uncertainty based on how much data we have in each regime.

We stress that this uncertainty is based on the assumption that p is iid. Beta distributed within each bin, which is clearly not the case if we acknowledge that p depends on more than one parameter. Therefore, these uncertainties can only serve as an indicator whether or not there are enough data to make a statement about this marginalized version of the true, multivariate distribution of p. In other words, they indicate how confident we can be in the best estimate of p for this dataset if we can only measure one parameter at a time.

The results of this process show a clear, highly significant dependence of the rogue wave probability on some sea state parameters, and the lack of such a dependence on others (Fig. 7). In particular, we find the following:

1) Surface elevation kurtosis, relative maximum wave height, and skewness are the strongest predictors for rogue wave risk. For relative maximum wave height, $P(\mathrm{AI} > 2)$ ranges between 2.9×10^{-5} and 1.0×10^{-3}. So if an up-to-date, in situ surface elevation time series is available, these parameters are able to quantify rogue wave risk with a factor of about 35 in variation.
2) Crest–trough correlation and spectral bandwidth (peakedness) are the strongest spectral predictors, with $P(\mathrm{AI} > 2)$ varying between 2.4×10^{-5} and 1.4×10^{-4} for crest–trough correlation—that is, almost one order of magnitude in variation from the spectrum alone.
3) The Tayfun wave height distribution (Tayfun 1990; Tayfun and Fedele 2007) seems to be an excellent baseline for rogue wave activity.
4) There is, at this level of detail, only a minor dependency of rogue wave occurrence on directional spread, Benjamin–Feir index, significant wave height, and steepness.

So, in this first analysis, it seems that bandwidth effects are the dominant modifier of rogue wave risk, whereas nonlinear effects (at least those governed by steepness and BFI) seem to play a minor corrective role in comparison with that. However, it is important to keep in mind that we are only looking at one set of stations and only one sea state parameter at a time.

6. Conclusions

FOWD is a free ocean wave dataset that relates wave point measurements to the conditions in which the wave occurred and that is optimized for use in data-mining and machine-learning applications. In the previous sections, we describe which quantities are included in our wave catalog FOWD and how they are computed, and which steps we take to ensure quality and reproducibility (section 2). We describe the reference implementation and the steps we take to be able to process massive amounts of data at the terabyte scale (section 3). We summarize the processing of the CDIP buoy data catalog and analyze the quality of the resulting catalog (section 4). We apply additional filtering to remove problematic measurements. By visual inspection, we find that the resulting dataset is of high quality. Last, we study the occurrence probability of rogue waves depending on the sea state in an example application, where we have been able to demonstrate that certain parameters are much better predictors than others (section 5). We find that, based on analyzing only one sea state parameter at a time, rogue wave risk can vary by at least one order of magnitude. The estimated rogue wave probabilities are consistent with those found in earlier studies based on observations and simulations (e.g., Fedele et al. 2016, 2017).

The strongest parameters in this analysis are surface elevation skewness/kurtosis, and maximum relative wave height of the past record. This is of little surprise when taking into account how many rogue waves occur in rapid succession of each other (Table 5), but the importance of kurtosis and skewness could also be evidence for the role of second- and third-order (weakly) nonlinear contributions (Mori and Janssen 2006; Gemmrich and Garrett 2011; Christou and Ewans 2014). The most important spectral parameters are spectral bandwidth and crest–trough correlation, which is compatible with the finding in Cattrell et al. (2018) that spectral bandwidth is important (although we disagree with the conclusion that rogue waves *cannot* be predicted from characteristic parameters).

On the other hand, we were unable to detect any noteworthy dependency of rogue wave risk on directional spread [hypothesized, e.g., by Gramstad et al. (2018) and McAllister et al. (2019)], wave steepness (which is evidence against the importance of weakly nonlinear corrections), or Benjamin–Feir index (one of two parameters used by ECMWF's freak wave forecast; see Janssen and Bidlot 2009). This does of course *not* prove that such dependencies do not exist, just that it is not detectable in this limited dataset (of Hawaiian stations) and by univariate analysis (i.e., considering one parameter at a time). A more sophisticated analysis is needed, which is precisely what we want to enable with FOWD.

We believe that this work represents an important motivation and contribution to enable physical insight into ocean waves through sophisticated data-driven methods. Downstream studies can either process their own raw data—because of the flexibility of the FOWD specification and reference implementation—or make use of the already processed CDIP data.

Extreme probabilistic events such as rogue waves are notoriously difficult to analyze statistically in a robust, meaningful way. By lowering the bar of entry for non–wave experts, we hope to enable new, powerful descriptive and predictive approaches to ocean wave phenomena.

Acknowledgments. Dion Häfner received funding from the Danish Hydrocarbon Research and Technology Centre (DHRTC). We thank Øyvind Breivik and Mika Malila for their valuable comments during the drafting stage of FOWD. We thank three anonymous reviewers for their constructive and insightful remarks. Raw data were furnished by the Coastal Data Information Program (CDIP), Integrative Oceanography Division, operated by the Scripps Institution of Oceanography, under the sponsorship of the U.S. Army Corps of Engineers and the California Department of Parks and Recreation. Computational resources were provided by DC^3, the Danish Center for Climate Computing.

Data availability statement. Filtered and unfiltered versions of the the FOWD–CDIP data are available for download at https://doi.org/10.17894/ucph.c589422c-64fd-4585-af31-4571497bcbe5. The exact version of the FOWD reference implementation used throughout this study (v0.5.2) is available at https://doi.org/10.5281/zenodo.4628203. The current version can be found at https://github.com/dionhaefner/FOWD. The scripts used to generate the plots and statistics in this paper are available at https://gist.github.com/dionhaefner/51ef93980a87d6b6bb557599b79582da.

APPENDIX

Complete Overview of All FOWD Quantities

See Table A1 for an exhaustive list of all quantities included in FOWD.

REFERENCES

Adcock, T. A. A., and P. H. Taylor, 2014: The physics of anomalous (‘rogue’) ocean waves. *Rep. Prog. Phys.*, **77**, 105901, https://doi.org/10.1088/0034-4885/77/10/105901.

Barbariol, F., J.-R. Bidlot, L. Cavaleri, M. Sclavo, J. Thomson, and A. Benetazzo, 2019: Maximum wave heights from global model reanalysis. *Prog. Oceanogr.*, **175**, 139–160, https://doi.org/10.1016/j.pocean.2019.03.009.

Behrens, J., J. Thomas, E. Terrill, and R. Jensen, 2019: CDIP: Maintaining a robust and reliable ocean observing buoy network. *2019 IEEE/OES 12th Current, Waves and Turbulence Measurement*, San Diego, CA, IEEE/OES, https://doi.org/10.1109/CWTM43797.2019.8955166.

Benetazzo, A., F. Barbariol, F. Bergamasco, A. Torsello, S. Carniel, and M. Sclavo, 2015: Observation of extreme sea waves in a space–time ensemble. *J. Phys. Oceanogr.*, **45**, 2261–2275, https://doi.org/10.1175/JPO-D-15-0017.1.

Boccotti, P., 2000: *Wave Mechanics for Ocean Engineering*. Elsevier, 520 pp.

Casas-Prat, M., and L. H. Holthuijsen, 2010: Short-term statistics of waves observed in deep water. *J. Geophys. Res.*, **115**, C09024, https://doi.org/10.1029/2009JC005742.

Cattrell, A. D., M. Srokosz, B. I. Moat, and R. Marsh, 2018: Can rogue waves be predicted using characteristic wave parameters? *J. Geophys. Res. Oceans*, **123**, 5624–5636, https://doi.org/10.1029/2018JC013958.

Christou, M., and K. Ewans, 2014: Field measurements of rogue water waves. *J. Phys. Oceanogr.*, **44**, 2317–2335, https://doi.org/10.1175/JPO-D-13-0199.1.

Dudley, J. M., G. Genty, A. Mussot, A. Chabchoub, and F. Dias, 2019: Rogue waves and analogies in optics and oceanography. *Nat. Rev. Phys.*, **1**, 675–689, https://doi.org/10.1038/s42254-019-0100-0.

Dysthe, K., H. E. Krogstad, and P. Müller, 2008: Oceanic rogue waves. *Annu. Rev. Fluid Mech.*, **40**, 287–310, https://doi.org/10.1146/annurev.fluid.40.111406.102203.

Fedele, F., 2015: On the kurtosis of deep-water gravity waves. *J. Fluid Mech.*, **782**, 25–36, https://doi.org/10.1017/jfm.2015.538.

——, J. Brennan, S. Ponce de León, J. Dudley, and F. Dias, 2016: Real world ocean rogue waves explained without the modulational instability. *Sci. Rep.*, **6**, 27715, https://doi.org/10.1038/srep27715.

——, C. Lugni, and A. Chawla, 2017: The sinking of the El Faro: Predicting real world rogue waves during Hurricane Joaquin. *Sci. Rep.*, **7**, 11188, https://doi.org/10.1038/s41598-017-11505-5.

——, J. Herterich, A. Tayfun, and F. Dias, 2019: Large nearshore storm waves off the Irish coast. *Sci. Rep.*, **9**, 15406, https://doi.org/10.1038/s41598-019-51706-8.

Fenton, J. D., 1988: The numerical solution of steady water wave problems. *Comput. Geosci.*, **14**, 357–368, https://doi.org/10.1016/0098-3004(88)90066-0.

Gemmrich, J., and C. Garrett, 2011: Dynamical and statistical explanations of observed occurrence rates of rogue waves. *Nat. Hazards Earth Syst. Sci.*, **11**, 1437–1446, https://doi.org/10.5194/nhess-11-1437-2011.

——, J. Thomson, W. E. Rogers, A. Pleskachevsky, and S. Lehner, 2016: Spatial characteristics of ocean surface waves. *Ocean Dyn.*, **66**, 1025–1035, https://doi.org/10.1007/s10236-016-0967-6.

Gramstad, O., E. Bitner-Gregersen, K. Trulsen, and J. C. Nieto Borge, 2018: Modulational instability and rogue waves in crossing sea states. *J. Phys. Oceanogr.*, **48**, 1317–1331, https://doi.org/10.1175/JPO-D-18-0006.1.

Haver, S., 2004: A possible freak wave event measured at the Draupner Jacket 1 January 1995. *Rogue Waves 2004*, Brest, France, IFREMER, http://www.ifremer.fr/web-com/stw2004/rw/fullpapers/walk_on_haver.pdf.

Holthuijsen, L. H., 2010: *Waves in Oceanic and Coastal Waters*. Cambridge University Press, 404 pp.

Huber, P. J., and E. M. Ronchetti, 2009: *Robust Statistics*. 2nd ed. John Wiley and Sons, 380 pp.

Janssen, P. A. E. M., 2003: Nonlinear four-wave interactions and freak waves. *J. Phys. Oceanogr.*, **33**, 863–884, https://doi.org/10.1175/1520-0485(2003)33<863:NFIAFW>2.0.CO;2.

——, and J.-R. Bidlot, 2009: On the extension of the freak wave warning system and its verification. ECMWF Tech. Memo. 588, 44 pp., https://doi.org/10.21957/uf1sybog.

Karmpadakis, I., C. Swan, and M. Christou, 2020: Assessment of wave height distributions using an extensive field database. *Coastal Eng.*, **157**, 103630, https://doi.org/10.1016/j.coastaleng.2019.103630.

Kharif, C., and E. Pelinovsky, 2003: Physical mechanisms of the rogue wave phenomenon. *Eur. J. Mech.*, **22B**, 603–634, https://doi.org/10.1016/j.euromechflu.2003.09.002.

Longuet-Higgins, M. S., 1952: On the statistical distribution of the height of sea waves. *J. Mar. Res.*, **11**, 245–266.

McAllister, M. L., and T. S. van den Bremer, 2019: Lagrangian measurement of steep directionally spread ocean waves: Second-order motion of a wave-following measurement buoy. *J. Phys. Oceanogr.*, **49**, 3087–3108, https://doi.org/10.1175/JPO-D-19-0170.1.

——, and ——, 2020: Experimental study of the statistical properties of directionally spread ocean waves measured by buoys. *J. Phys. Oceanogr.*, **50**, 399–414, https://doi.org/10.1175/JPO-D-19-0228.1.

——, S. Draycott, T. A. Adcock, P. H. Taylor, and T. S. Bremer, 2019: Laboratory recreation of the Draupner wave and the role of breaking in crossing seas. *J. Fluid Mech.*, **860**, 767–786, https://doi.org/10.1017/jfm.2018.886.

Mori, N., and P. A. E. M. Janssen, 2006: On kurtosis and occurrence probability of freak waves. *J. Phys. Oceanogr.*, **36**, 1471–1483, https://doi.org/10.1175/JPO2922.1.

Ochi, M. K., and E. N. Hubble, 1976: Six-parameter wave spectra. *15th Int. Conf. on Coastal Engineering*, Honolulu, HI, ASCE, 301–328, https://doi.org/10.1061/9780872620834.018.

Portilla-Yandún, J., 2018: The global signature of ocean wave spectra. *Geophys. Res. Lett.*, **45**, 267–276, https://doi.org/10.1002/2017GL076431.

——, A. Salazar, and L. Cavaleri, 2016: Climate patterns derived from ocean wave spectra. *Geophys. Res. Lett.*, **43**, 11 736–11 743, https://doi.org/10.1002/2016GL071419.

Serio, M., M. Onorato, A. R. Osborne, and P. A. E. M. Janssen, 2005: On the computation of the Benjamin-Feir index. *Nuovo Cimento*, **28C**, 893–903, https://doi.org/10.1393/ncc/i2005-10134-1.

Slunyaev, A., I. Didenkulova, and E. Pelinovsky, 2011: Rogue waters. *Contemp. Phys.*, **52**, 571–590, https://doi.org/10.1080/00107514.2011.613256.

Tayfun, M. A., 1990: Distribution of large wave heights. *J. Waterw. Port Coastal Ocean Eng.*, **116**, 686–707, https://doi.org/10.1061/(ASCE)0733-950X(1990)116:6(686).

——, and F. Fedele, 2007: Wave-height distributions and nonlinear effects. *Ocean Eng.*, **34**, 1631–1649, https://doi.org/10.1016/j.oceaneng.2006.11.006.

Toffoli, A., O. Gramstad, K. Trulsen, J. Monbaliu, E. Bitner-Gregersen, and M. Onorato, 2010: Evolution of weakly nonlinear random directional waves: Laboratory experiments and numerical simulations. *J. Fluid Mech.*, **664**, 313–336, https://doi.org/10.1017/S002211201000385X.

Welch, P., 1967: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, **15**, 70–73, https://doi.org/10.1109/TAU.1967.1161901.

Xiao, W., Y. Liu, G. Wu, and D. K. P. Yue, 2013: Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution. *J. Fluid Mech.*, **720**, 357–392, https://doi.org/10.1017/jfm.2013.37.

Young, I. R., 1995: The determination of confidence limits associated with estimates of the spectral peak frequency. *Ocean Eng.*, **22**, 669–686, https://doi.org/10.1016/0029-8018(95)00002-3.

2.2 ARTICLE II — REAL-WORLD ROGUE WAVE PROBABILITIES

Now that we have access to a large high-quality dataset, we can search for parameter combinations with significantly enhanced rogue wave activity. We address this in the second article (Häfner, Gemmrich, and Jochum, 2021b).

Figure 2.6: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"a ship hit by a rogue wave in the distance:80 | unsplash:20"*.

So far we have only looked at a subset of the full FOWD data, and at each parameter in isolation. To remedy the former we aggregate the FOWD dataset into chunks of 100 waves in which we assume the sea state to be constant, which allows us to analyze the entire dataset at once. We once again use Bayesian histograms for the univariate analysis. To control for correlations between parameters we study conditional probabilities through the same approach, which gives some first evidence on the causal structure of the problem (through observed conditional independencies). For this we introduce the quantity *predictive power*, which measures by how much p changes as each parameter is varied.

We also need to determine whether there are significant interactions between parameters (because if there are, a univariate analysis will be highly misleading), again in a non-parametric way and with uncertainties. There is no off-the-rack machine learning algorithm that fits these requirements considering our data volume and low event rates. This led us to develop our own method for this task.

For the multivariate analysis, we use a shallow decision tree surrogate model to cluster the rogue wave probability p (estimated through a deep random forest classifier) into high-dimensional rectangular regions with approximately constant p (Fig. 2.7). We then interpret the results of this process separately for each cluster in the same way as in the univariate case, which also gives us uncertainties of p within each cluster (based on unseen validation data to mitigate overfitting).

This approach, similar to Bayesian histograms, is quite conservative and not the strongest learner, since it assumes no dependence between neighboring bins / clusters whatsoever, but in our case we have enough data to be able to value robustness over exploitation. Our multivariate analysis shows that feature interactions are generally weak, so we focus on the univariate analysis in the article.

Our analysis reveals that crest-trough correlation is the dominant causal parameter behind rogue wave formation. Surface elevation kurtosis, which we found to be an important parameter in article 1, has no predictive quality in the aggregated data, which suggests that it can only indicate whether a rogue wave is already forming and cannot be used for forecasting.

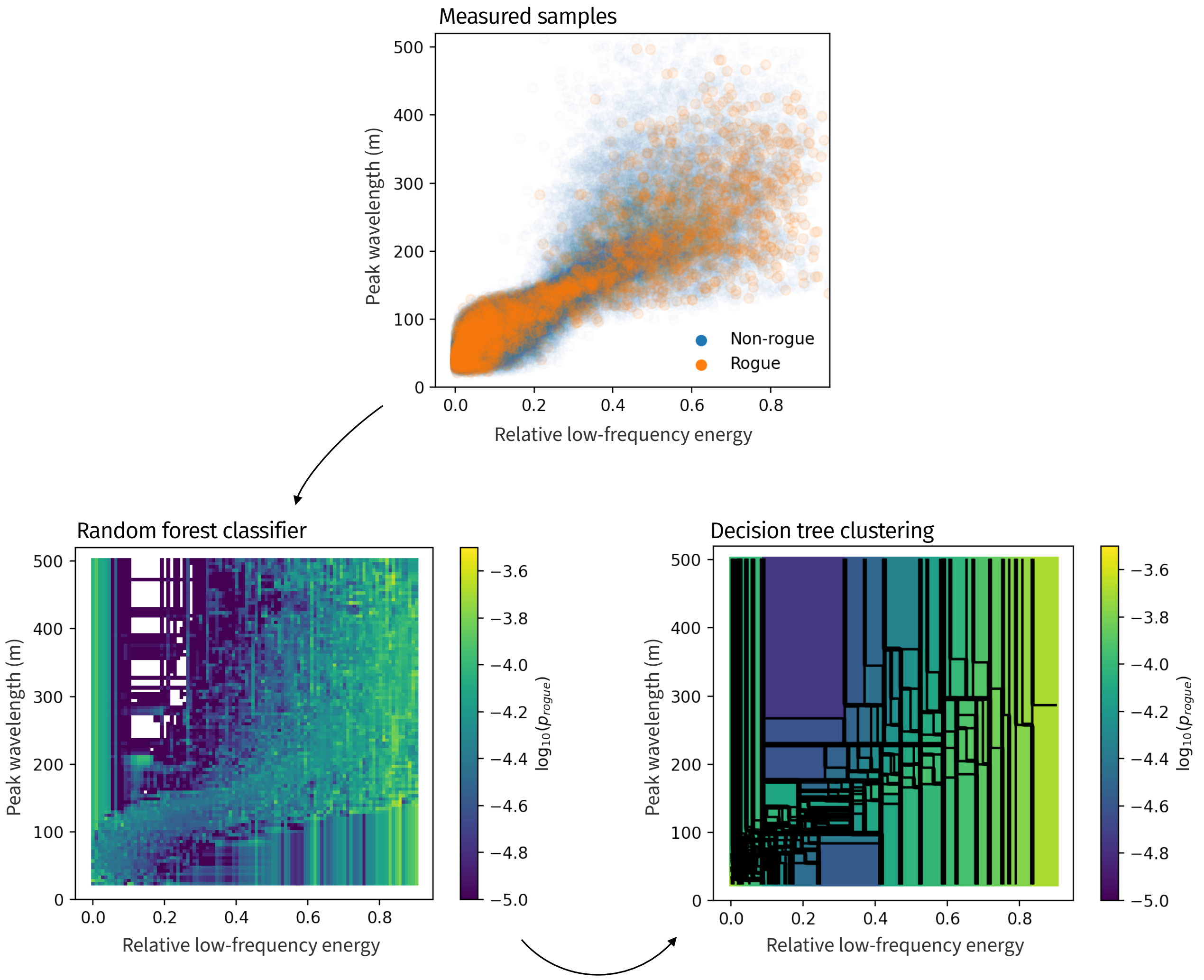


Figure 2.7: Decision tree clustering for rare events. In the first step, raw samples are converted into (noisy) probability estimates, e. g. through a random forest classifier. In the second step, a surrogate decision tree model is trained on the output of the first model with (high) constant number of samples per leaf. The resulting leaves represent a rectangular partition in feature space where p is approximately constant in each cluster. This allows us to analyze each cluster separately based on the number of rogue and non-rogue samples in it, just like in the univariate case.

scientific reports

Check for updates

OPEN

Real-world rogue wave probabilities

Dion Häfner[1]✉, Johannes Gemmrich[2] & Markus Jochum[1]

Rogue waves are dangerous ocean waves at least twice as high as the surrounding waves. Despite an abundance of studies conducting simulations or wave tank experiments, there is so far no reliable forecast for them. In this study, we use data mining and interpretable machine learning to analyze large amounts of *observational data* instead (more than 1 billion waves). This reveals how rogue wave occurrence depends on the sea state. We find that traditionally favored parameters such as surface elevation kurtosis, steepness, and Benjamin–Feir index are weak predictors for real-world rogue wave risk. In the studied regime, kurtosis is only informative within a single wave group, and is *not* useful for forecasting. Instead, crest-trough correlation is the dominating parameter in all studied conditions, water depths, and locations, explaining about a factor of 10 in rogue wave risk variation. For rogue crests, where bandwidth effects are unimportant, we find that skewness, steepness, and Ursell number are the strongest predictors, in line with second-order theory. Our results suggest that linear superposition in bandwidth-limited seas is the main pathway to "everyday" rogue waves, with nonlinear contributions providing a minor correction. This casts some doubt whether the common rogue wave definition as any wave exceeding a certain height threshold is meaningful in practice.

An extreme ocean wave ("rogue wave" or "freak wave") is commonly defined as any wave that is higher than 2 or 2.2 times the significant wave height H_S, and they pose a substantial threat to seafaring vessels and offshore structures[1].

Despite having been in research focus for almost 25 years, they are still being studied extensively[2–7]. By now, we know several ways to produce truly exceptional waves in wave tanks and simulations[8–10]. However, things are more difficult in the real ocean, where theoretical assumptions (such as unidirectionality) break down. The causes of real-world rogue waves are therefore still unknown, and heavily debated[11–18].

In recent years, more and more studies approached the problem from a different angle: by inferring the dependence of rogue wave occurrence on the sea state from observed field data[3,5,11,18]. However, no study has so far quantified the probability to encounter a rogue wave depending on the sea state throughout a wide regime of conditions, taking into account more than one parameter at a time, and in a statistically robust fashion. Here, we aim to fill this gap.

In this study, we use FOWD (the Free Ocean Wave Dataset)[19], a wave catalogue based on data recorded by buoys in 158 different locations around the US coasts and overseas territories, based on raw data from CDIP[20] (Coastal Data Information Program). We use the pre-filtered version of FOWD-CDIP (v0.4.4) containing about 1.5 billion individual waves (of which about 100,000 exceed $2H_S$), which has already removed faulty deployments and waves recorded during conditions where buoys are unreliable.

We create an aggregated version of the full dataset that bundles together 100 waves at a time (see "Methods"), and are thus able to analyze all sea states simultaneously using robust Bayesian statistics and machine learning. By finding the conditions that show the highest rogue wave probability, we aim to test some common hypotheses concerning rogue waves and their creation mechanisms. To this end, we include only a subset of 12 sea state parameters that we can meaningfully tie to a (hypothesized) cause of rogue waves or crests (Table 1).

We identify the key control parameters for real-world rogue wave risk via careful examination of the correlation between these parameters and measured rogue wave occurrences. Because many of the parameters are also correlated with each other, we have to account for possible confounding along every step (correlation matrix shown in Supplementary Figure S1).

The upcoming sections present the results of this analysis, followed by a discussion of possible limitations and conclusive remarks.

[1]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. [2]University of Victoria, Victoria, BC, Canada. ✉email: dion.haefner@nbi.ku.dk

Parameter	Physical meaning	References
Crest-trough correlation	Correlation coefficient between wave crest heights and trough depths	[18,21,22]
Spectral bandwidth	Spectral peak width, controls wave group dynamics	[11]
Mean period	Mean wave period	[2]
Rel. low-frequency energy	Relative low-frequency (swell) energy content	[2,4,23,24]
Directional spread	Short-crestedness of waves	[6]
Ursell number ($\log_{10}$)	Non-linear shallow water effects	[25]
Benjamin–Feir index	Degree of non-linearity, modulational instability	[26–28]
Excess kurtosis	Proneness to outliers of sea surface elevation	[13,26,29]
Steepness	Weakly nonlinear corrections, wave breaking	[15,17]
Significant wave height	Reference wave height, total energy	[14]
Skewness	Shape asymmetry between wave crests and troughs	[13,30]
Relative depth ($\log_{10}$)	Shallow-water effects	[27]

Table 1. The sea state parameters examined in this study. See Table 2 for more information about the estimation of each parameter.

Results

Throughout the following sections, we characterize the extremeness of a wave or crest by its abnormality index (AI for waves and CAI for crests). This is defined as $\mathrm{AI} = \mathrm{H}/\mathrm{H}_\mathrm{S}$ and $\mathrm{CAI} = \eta/\mathrm{H}_\mathrm{S}$, where H is the measured zero-crossing wave height, η the measured crest height, and H_S the 30 min spectral significant wave height.

Unless stated otherwise, all analysis is based on the full, aggregated FOWD-CDIP dataset (or stratified versions of it).

The following sections present the 4 main results of this study.

Bandwidth effects are the dominant pathway to rogue waves. To quantify how the rogue wave probability p depends on the sea state, we first examine how p changes when varying one sea state parameter at a time. Here, p is defined as the probability of any given wave to exceed the rogue wave threshold, i.e., $p = \Pr[\mathrm{AI} > \mathrm{y}]$ with $y = 2.0$ and, where we have enough data, also $y = 2.4$.

We split each sea state parameter x (Table 1) evenly into N bins, and assume that the associated wave height measurements are independently, identically distributed within each bin (see "Methods"). The "predictive power" $\mathbb{P}_x$ of a parameter x then quantifies the logarithmic ratio between the highest and lowest binned value of $p(x)$. For example, a value of $\mathbb{P}_x = 2$ implies that $p(x)$ changes by 2 orders of magnitude as x is varied.

Applying this binning, we find that crest-trough correlation has the highest univariate predictive power out of all parameters (Fig. 1), explaining about 1 order of magnitude in variation of p (with values ranging between $3 \cdot 10^{-5}$ and $2 \cdot 10^{-4}$ for $\mathrm{AI} = 2$). Spectral bandwidth, mean period, and low-frequency energy content are also informative with $\mathbb{P}$ between 0.5 and 0.8, but these parameters are strongly correlated with crest-trough correlation, so we have to control for possible confounding.

To examine whether spectral bandwidth or crest-trough correlation is the real causal factor, we stratify our analysis on each of these parameters. When stratifying on spectral bandwidth, crest-trough correlation is still the most informative parameter with $\mathbb{P} \approx 0.5$, while all other parameters drop to $\mathbb{P} < 0.2$. When stratifying on crest-trough correlation, all other parameters become unimportant with most values of $\mathbb{P}$ between 0 and 0.2, depending on which value of crest-trough correlation we condition on (see also Supplementary Figure S2).

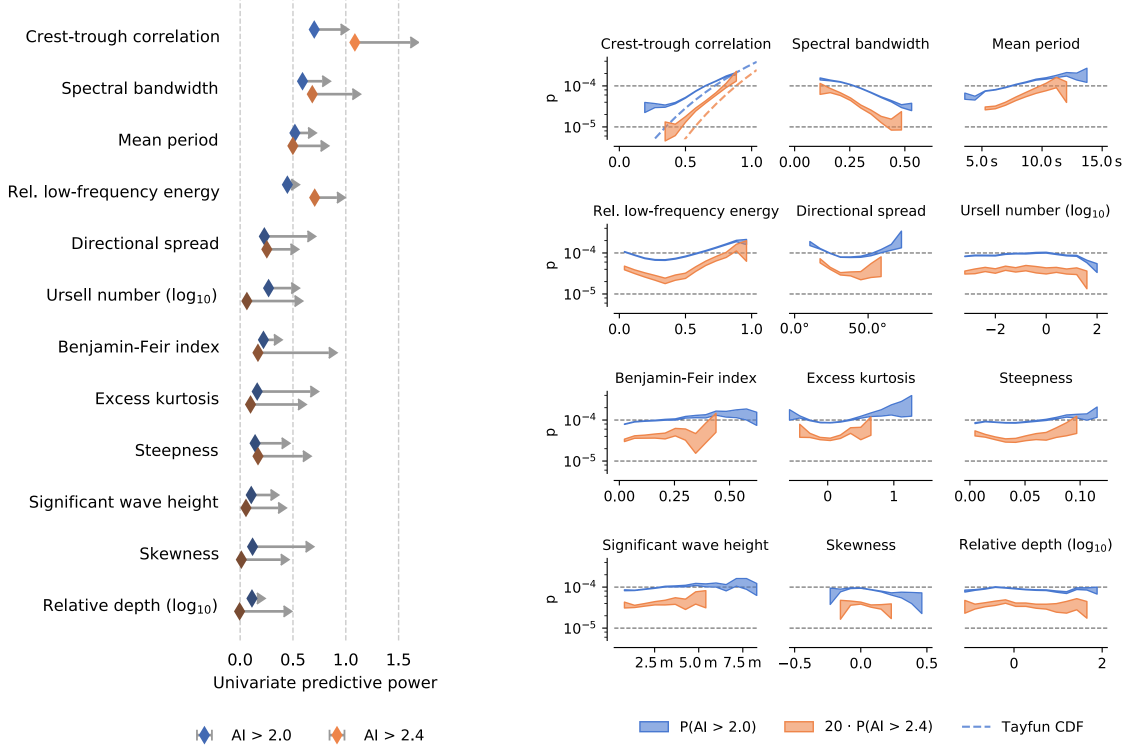
This implies that spectral bandwidth (and most other parameters) act *through* their correlation with crest-trough correlation. This is strong evidence that crest-trough correlation is the key control parameter for rogue waves, with some other factors serving as minor corrections.

When we take the full, multivariate parameter space into account, things are more difficult to analyze, because interactions between parameters could possibly create "hot corners" of elevated rogue wave activity that are not detectable by univariate analysis. To discover whether this is the case, we run a clustering algorithm that identifies rectangular regions in parameter space where we find higher rogue wave probabilities than in any univariate bin (see "Methods" section).

This multivariate analysis reveals that crest-trough correlation is still the most important parameter in all found clusters, where all cluster populations have crest-trough correlations above 0.75 (Fig. 2). All of the clusters are also located in swell-dominated conditions with high mean period, low directional spread, and low steepness. We examine the role of wave period and steepness further below.

Surface elevation kurtosis does *not* predict rogue waves. The kurtosis (fourth standardized moment) of the sea surface elevation is a commonly studied parameter in connection with rogue waves[13,29,31], and a central ingredient of ECMWF's rogue wave forecast[26]. However, some authors have expressed doubt whether a high kurtosis is the *cause* or *effect* of extreme waves[32,33], as kurtosis is a measure for tail-heaviness of a distribution, and rogue waves are extreme outliers by definition. In other words, we examine the question: is a sea state that is more prone to outliers in the recent past also prone to more outliers (rogue waves) now?

We examine this by studying how the predictive power of kurtosis depends on the time lag between the end of the aggregation period (based on which the sample kurtosis is computed) and the observed wave height. Because



(a) Lower bound (2.5th percentile) predictive power of each parameter. Arrows indicate location of upper bound (97.5th percentile).

(b) The underlying scaling of rogue wave probability p with each parameter. Shading indicates 95 % credible interval of p. Curves for $P(\mathrm{AI} > 2.4)$ are scaled by a factor of 20.

Figure 1. When looking at one sea state parameter at a time, some are better predictors for rogue wave occurrence than others. In particular, crest-trough correlation and spectral bandwidth are much more informative than e.g. Benjamin–Feir index and steepness. (**a**) Shows the predictive power of each parameter, which is computed from the range spanned by the curves in (**b**) (the variation of the rogue wave probability with each parameter).

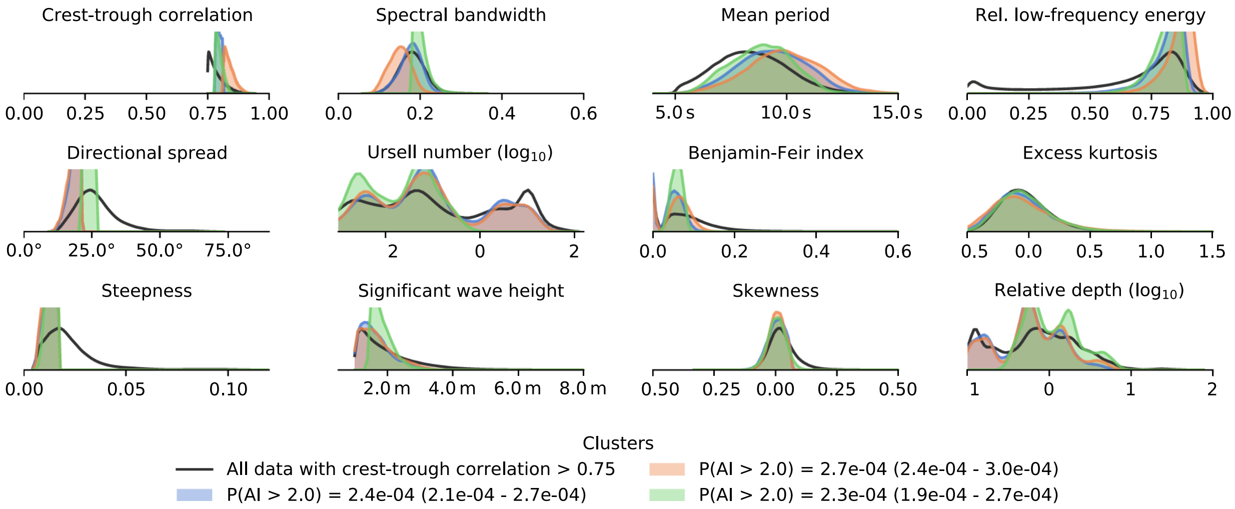


Figure 2. "Hot corners" of rogue wave activity have high crest-trough correlation, strong swells, and low steepness. Shown is the distribution of each cluster population in parameter space, and the distribution of all waves with high crest-trough correlation for comparison. Clusters are computed through decision-tree based clustering (see "Methods"), taking all parameters into account at the same time. All clusters show a higher rogue wave incidence than any univariate bin. Ranges in legend indicate 95% credible interval.

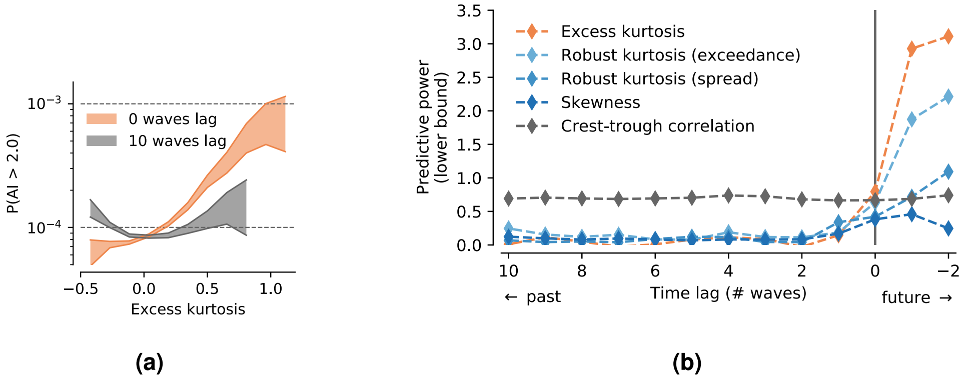


Figure 3. Past sea surface elevation kurtosis is a poor predictor for rogue wave occurrence in the future. Shown is the scaling of the rogue wave probability p with kurtosis for 2 different values of time lag (**a**) and the resulting *predictive power* of various quantities depending on time lag (**b**). Here, *time lag* refers to the time between the end of the aggregation period used to compute each sea state parameter and the start of the observed wave.

we can only study this in non-time aggregated data, which requires 100 times more resources than aggregated data, we need to restrict this analysis to a subset of the full dataset. We use the FOWD data from all Hawaiian CDIP stations (098p1, 106p1, 146p1, 165p1, 187p1, 188p1, 198p1, 225p1, 233p1), containing 160 million waves.

We also include two robust kurtosis estimators in this analysis (based on quantile spread and expected exceedance probabilities[34]), as the sample kurtosis based on the fourth moment of the sea surface elevation is a noisy quantity that is highly sensitive to single extreme measurements. These robust alternatives should be more accurate estimators for the true kurtosis of the sea state (as can be obtained through simulations or very long, controlled experiments under identical conditions).

Results show that even a small time lag of only 3 waves between the end of the aggregation period and observed wave height reduces the predictive power of kurtosis to its (low) background value (Fig. 3). If the kurtosis is computed including future state (negative time lag), it is extremely informative as expected, since rogue wave occurrence *causes* very high values of kurtosis. But even for a time lag of 0, where the end of the aggregation period lies right before the current wave, we discover a substantially elevated predictive power.

We explain this with the common occurrence of multiple rogue waves within the same wave group, where measuring the first rogue wave gives an elevated probability of encountering a second one right after. Indeed, the FOWD dataset contains a relatively high number of multiple rogue waves in rapid succession (about 2500 waves with $\mathrm{AI} > 2$ within 30 s of each other, which corresponds to about 3% of all rogues)[19].

We also find that the robust kurtosis estimators are not more informative than straightforward sample kurtosis, even though they are indeed less affected by time lag.

We conclude therefore that surface elevation kurtosis is a short-ranged predictor that is only useful within a single wave group, and has little predictive quality otherwise. This has an important implication. If outliers in the past are a poor predictor for outliers in the future, one sensible interpretation is that the encounter of a rogue wave is indeed mostly up to chance (and thus unlikely to elevate the general proneness to outliers in the whole sea state).

The effects of steepness and Benjamin–Feir index depend on wave period. If we look at how the rogue wave probability depends on spectral energy content (Fig. 1), we notice something curious: p attains a local maximum for both very high-frequency and very low-frequency seas. To investigate this, we re-run our analysis for high-frequency and low-frequency conditions.

As low-frequency/high-frequency seas we take all data where the relative energy content in the spectral band 0.05 Hz to 0.1 Hz (representing swell) lies in the interval (0, 0.1) and (0.8, 0.85), respectively.

This reveals a fundamental difference between these regimes (Fig. 4). Low-frequency seas have naturally higher values of p, even for similar values of crest-trough correlation. High-frequency seas show a lower baseline p, but are able to reach almost the same maximum p through an additional dependency on steepness and Benjamin–Feir index (BFI) that is absent in the low-frequency case. In fact, this relationship is inverted in low-frequency seas, where p is *lower* for higher steepness and BFI.

To understand this, it is important to keep in mind that steepness acts on extreme waves in multiple ways. On one hand, steepness is the key parameter in weakly nonlinear modifications to the wave height distribution[17]. On the other hand, steepness also governs wave breaking, an effect that tends to *remove* tall waves[35,36]. Depending on the physical regime, either effect might take over, and fundamentally change the way steepness influences extreme waves.

High-frequency seas can under certain, rare conditions reach about the same rogue wave probabilities as low-frequency seas. The strongest multivariate cluster has a lower bound p of $1.6 \cdot 10^{-4}$ for $\mathrm{AI} = 2$ (Supplementary Figure S3). Therefore, the chance to encounter a rogue wave *within a certain time window* is greatest under these conditions (so far, we have only considered the probability *per wave*).

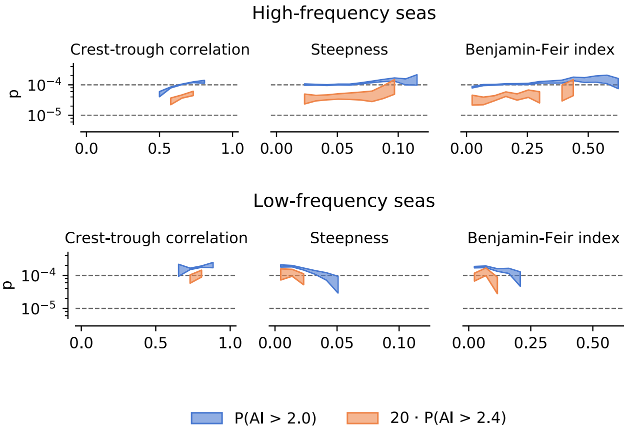


Figure 4. Low-frequency seas have naturally higher rogue wave activity for similar crest-trough correlations, but scale negatively with steepness and BFI. Shown is the scaling of the rogue wave probability p with some sea state parameters. Low-frequency/high-frequency conditions are all seas with relative low-frequency energy in the interval (0.8, 0.85) and (0, 0.1), respectively. Curves for $P(\mathrm{AI} > 2.4)$ are scaled by a factor of 20.

Rogue crests are governed by skewness, steepness, and Ursell number. Crest heights differ in some fundamental ways from wave heights, since they are affected by second-order nonlinearities that cancel out for wave heights[22], and they are (by definition) *not* affected by crest-trough correlation. Therefore, we re-run our full analysis for rogue crests.

We find that crest-trough correlation and spectral bandwidth are indeed of very low predictive power (Fig. 5). Instead, surface elevation skewness, steepness, and Ursell number are the strongest parameters, with predictive powers between 0.5 and 1.0. Our multivariate analysis fails to reveal any regions with higher rogue wave probability than the most extreme univariate bin (where $\log_{10}(\text{Ursell number}) \in (1.8, 2.2)$).

A positive skewness indicates steeper crests and flatter, more rounded troughs, and is frequently cited as a proxy for second-order bound nonlinear corrections[13,30,37]. Steepness and Ursell number are the central parameters of the Forristall crest height distribution[25]. Therefore, it seems that rogue crest heights are well explained by second-order theory at this level of detail, but further corrections of up to fourth order may be needed for extremely rare rogue crests[17].

Discussion

The results presented during the previous sections are robust to analysis parameter choices and sample size effects (all statements are based on 95% credible intervals).

In particular, we find that our results are stable with regard to sensor location and water depth. To investigate this, we re-ran the analysis on several subsets of the full data, grouped by geographic region (Southern California, Hawaii, US East Coast, West Pacific), relative water depth, and single stations. We did not detect any notable deviations from the dependencies of p on the sea state presented above (wherever such comparisons were possible due to the reduced amount of data).

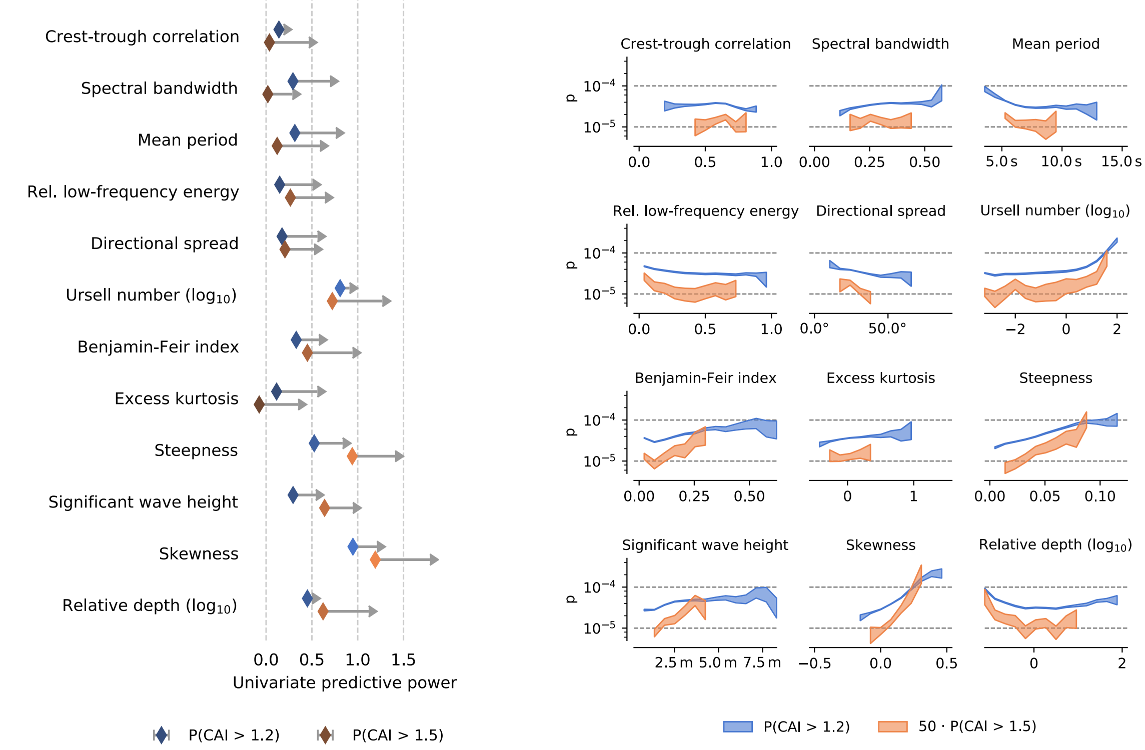
This is surprising, as shallow-water effects and interactions with bathymetry are one hypothesized cause of rogue waves[38,39]. On the other hand, these effects typically require special topographic conditions, which might simply not be present in our data.

The weak dependence of p on significant wave height seems to imply that large rogue waves tend to be governed by the same dynamics as small rogue waves. As with location and water depth, we investigated this to some degree by re-running the analysis using only conditions with a significant wave height > 4 m. Again, we did not observe notably different scalings of p with the sea state (where there were enough data).

We identified crest-trough correlation as the most important parameter for rogue wave formation. It is well understood that bandwidth effects are an important parameter for wave heights[11,21,22]. Perhaps more surprising is the absence of a strong dependency on steepness and BFI, which are central ingredients in current rogue wave prediction[26], even though we did detect a small positive influence in high-frequency seas (such as storms). For more discussion on the implications of these findings see "Conclusion" section.

Regarding the reliability of our results, some caveats still apply. As the underlying data are supplied by buoys in mostly coastal regions, there are some considerations that might limit the applicability of these results.

Wave buoys are known to underestimate extreme crests through several mechanisms, such as lateral movements around the crest, being dragged through the crest, or linearization of the sea state due to their Lagrangian motion. Even though these effects were found to be of minor importance[40], we cannot rule out that our conclusions are potentially biased by this. Therefore, our buoy data could underestimate the total number of rogue waves to some degree, and the influence of second-order effects on wave crests might be even higher if measured by a different sensor (this does not affect wave heights, though).



(a) Lower bound (2.5th percentile) predictive power of each parameter. Arrows indicate location of upper bound (97.5th percentile).

(b) The underlying scaling of rogue wave probability p with each parameter. Shading indicates 95 % credible interval of p. Curves for $P(\mathrm{CAI} > 1.5)$ are scaled by a factor of 50.

Figure 5. For rogue crests, skewness, steepness, and Ursell number are the most informative parameters. Plots are identical to Fig. 1, except that they refer to crest instead of wave heights.

The location of the buoys is another biasing factor. Overall, we are confident that our findings are robust in the studied regime of coastal and island regions in shallow and deep water at moderate significant wave heights, but they might be different in other regions and conditions where we did not have data.

We also do not include any parameters that are not measurable from the sea surface elevation, such as atmospheric conditions (winds), ocean currents, or local topography. There is good evidence that these factors can be important in certain situations[39,41,42], but since they depend on localized features we do not expect them to be very good predictors in aggregated data from different locations (with the possible exception of winds).

Overall, it is important to keep in mind that our results relate to the rogue wave probability per wave at one given location in space. For extended periods of time and large objects such as oceangoing vessels, the total risk to encounter a rogue wave will be dramatically higher than the probabilities we present here.

Conclusion

By analyzing over 1 billion wave measurements from buoys, we find that the by far most important parameter for rogue wave occurrence is crest-trough correlation (parameter r in the Tayfun distribution[22]). This suggests that, in most conditions, the Rayleigh distribution for Gaussian seas[43] is in fact an upper bound for real-world rogue waves, as the Tayfun distribution converges to the Rayleigh distribution for $r \to 1$. Characteristic steepness, BFI, and swell strength provide minor corrections to this. On the other hand, sea surface elevation kurtosis, which is taken as an important indicator for rogue wave activity in many studies[13,29,31], appears to have no detectable predictive quality when controlling for the fact that rogue waves naturally cause higher kurtosis.

We interpret this as evidence that almost all “freaks” are actually rare realizations of linear or weakly nonlinear seas that are fairly well described by available wave height statistics[22,25]. A similar conclusion has been reached by other, simulation-based studies[13,17].

This implies that the term *rogue wave* should perhaps be reserved for waves that are truly a “different breed” (such as those caused by modulational instability and other nonlinear effects, or those occurring during a storm), not just any wave that exceeds an arbitrary abnormality index threshold.

Rogue crests seem to be reasonably well-described by second-order, weakly nonlinear theory[25,30,37], as we found the most important parameters to be skewness, steepness, and Ursell number. However, we did focus on waves during our analysis, so there might be more to uncover—e.g., when conditioning on skewness or different depth regimes.

We also see this work as a demonstration how machine learning methods can be helpful in extreme wave research. Some previous studies have attempted to perform binary classification on rogue wave data[44,45] (i.e.,

Parameter	Related FOWD variable(s)	Estimation
Crest-trough correlation	`sea_state_30m_crest_trough_correlation`	See (1). This represents the envelope of the autocorrelation function at time lag of 1/2 mean zero-crossing period for linear waves
Spectral bandwidth	`sea_state_30m_bandwidth_peakedness`	Peakedness (quality factor) of wave spectral density[46]
Mean period	`sea_state_30m_mean_period_spectral`	$\sqrt{m_0/m_2}$, with n-th moment of wave spectral density m_n
Rel. low-frequency energy	`sea_state_30m_rel_energy_in_frequency_interval`	$m_0^{-1} \int S(f)$ df in the frequency interval 0.05 Hz to 0.1 Hz, with wave spectral density $S(f)$
Directional spread	`direction_dominant_spread_in_frequency_interval`, `sea_state_30m_rel_energy_in_frequency_interval`	Average over frequency-dependent directional spread weighted with energy in each frequency band
Ursell number ($\log_{10}$)	`sea_state_30m_steepness`	Ursell number $U = \epsilon/\tilde{D}^3$, with relative water depth $\tilde{D}$ and characteristic steepness ϵ
Benjamin–Feir index	`sea_state_30m_benjamin_feir_index_peakedness`	Through characteristic steepness and spectral bandwidth (peakedness)[46]
Excess kurtosis	`sea_state_30m_kurtosis`	Fourth standardized moment of surface elevation time series
Steepness	`sea_state_30m_steepness`	Characteristic steepness $\epsilon = \sqrt{2m_0}k_p$ with spectral peak wavenumber k_p
Significant wave height	`sea_state_30m_significant_wave_height_spectral`	Significant wave height $H_S = 4\sqrt{m_0}$
Skewness	`sea_state_30m_skewness`	Third standardized moment of surface elevation time series
Relative depth ($\log_{10}$)	`sea_state_30m_peak_wavelength`, `meta_water_depth`	Relative depth $\tilde{D} = D/\lambda_p$ with water depth D and peak wavelength λ_p

Table 2. Overview of how each sea state parameter is estimated from the sea surface elevation.

to predict whether a rogue wave will occur in some block of data or not). We believe that due to the inherently stochastic nature of ocean waves, predicting *rogue wave probabilities* is a better way forward, and have demonstrated that this can lead to tangible insights.

Finally, our statistical and machine learning-based analysis in this study has been purely descriptive. We believe that this work also has important implications for rogue wave *prediction*. Crest-trough correlation can be computed from the wave spectrum, which is routinely forecast globally by agencies like ECMWF. This provides a strong baseline for a rogue wave risk forecast. Combined with more sophisticated machine learning algorithms that are not piecewise constant and take the actual wave height into account (not just binary classification), we are confident that wave height distribution tails will become much more forecastable in the future.

Methods

Parameter estimation. Most parameters are taken directly from FOWD without modification. The only exceptions are peak relative depth, Ursell number, and dominant directional spread, which are not part of FOWD, but can be computed based on other FOWD parameters.

An overview over how each parameter is estimated is shown in Table 2. All parameters are based on a 30 min aggregation window.

Since the crest-trough correlation r is a central parameter to this article, we give the full expression here[19,22]:

$$r = \frac{1}{m_0}\sqrt{\rho^2 + \lambda^2} \quad \text{with} \quad \rho = \int_0^\infty \mathrm{S}(\omega)\cos\left(\omega\frac{\overline{T}}{2}\right)\,\mathrm{d}\omega, \quad \lambda = \int_0^\infty \mathrm{S}(\omega)\sin\left(\omega\frac{\overline{T}}{2}\right)\,\mathrm{d}\omega \tag{1}$$

where $S(\omega)$ is the wave spectral density, m_n its n-th moment, ω the angular frequency, and $\overline{T} = m_0/m_1$ the spectral mean period.

Data preprocessing. We apply the following preprocessing steps to the FOWD wave catalogue:

1. To account for the sampling variability of our relatively low-frequency buoy data, we correct all wave/crest heights and trough depths (and quantities directly derived from them) based on the mean wave period $\overline{T}$ and sampling frequency f_0[18]:

$$h' = h \cdot \left(1 - \frac{\pi^2}{6(f_0\overline{T})^2}\right)^{-1} \tag{2}$$

   As FOWD filtering already removes all records with mean period lower than 5 s for 1.28 Hz CDIP data, this correction factor is quite conservative (maximum possible value of 4.2%).

2. To reduce the 800 GB FOWD-CDIP dataset to a manageable size, we aggregate records into chunks by mapping each 100th sea state to the maximum measured wave height in the upcoming 100 waves.

   This is notably different from the traditional approach to create fixed-*time* chunks (usually 20 min[11,18]). Having a fixed number of waves allows us to directly translate the probability of finding at least one rogue wave within the aggregation window (p_{100}) to the rogue wave probability for any given wave (p), assuming that all wave heights are identically, independently distributed (*iid.*) within the aggregation period:

$$p = 1 - (1 - p_{100})^{1/100} \tag{3}$$

	α_0	β_0
AI > 2	1	10,000
AI > 2.4	1	1,000,000
CAI > 1.2	1	10,000
CAI > 1.4	1	1,000,000

Table 3. Beta prior parameters for p for different wave (AI) and crest (CAI) height thresholds.

> This process also removes the influence of multiple rogue waves occurring back-to-back, because we only measure the probability that at least one wave in the record is a rogue wave. This has an additional regularizing effect that prevents the analysis from over-emphasizing conditions which have a tendency for multiple rogue waves.

All preprocessed data are freely available for download (see data availability statement).

Univariate binning. In the univariate case, we split all wave height observations into N equal-sized bins for each sea state parameter x. Our analysis then hinges on the assumption that all binary samples within a bin (consisting of n^+ rogue and n^- non-rogue observations) are identically, independently distributed (*iid.*) according to a binomial distribution with rogue wave probability p as the only parameter. Our goal is to estimate p, which we interpret in Bayesian fashion as a random variable, from measurements of n^+ and n^- within each bin (we introduced this process in the initial publication of FOWD[19]).

For p we assume a Beta distributed prior with parameters α_0, β_0 (Table 3). The role of this prior is to constrain p to a reasonable order of magnitude, while being weakly informative so the exact choice of parameters does not influence final results.

Because the Beta prior is conjugate to the binomial likelihood, we obtain for the posterior of p:

$$P(p \mid n^+, n^-) = \mathrm{Beta}(n^+ + \alpha_0,\ n^- + \beta_0) \tag{4}$$

Since this is just another Beta distribution, the posterior for p is easy to evaluate with any modern statistical software. Specifically, we quantify our best estimate for p through the median of (4), and our uncertainty by the 95% credible interval (based on quantiles of the posterior).

The assumption that measurements are iid. within each univariate bin is obviously not fulfilled if p depends on more than one sea state parameter, so the uncertainties obtained through this process can only give an indication of our confidence in the marginal rogue wave probability when we can only measure one parameter at a time. We also need to pick small enough bins such that the variance of the true $p(x)$ is small within each bin.

In the case of aggregated data, we model p_{100} instead of p via (4), where n^+/n^- relate to the number of 100-wave chunks containing a rogue wave/no rogue waves, and with β_0 reduced by a factor of 100. After estimating the desired statistical properties of p_{100} (median and quantile-based credible interval), we translate those into the corresponding values of p via (3) (all reported quantities are *per wave*).

Predictive power. We define the "predictive power" $\mathbb{P}_x$ of a parameter x as:

$$\mathbb{P}_x = \log_{10}\left(\frac{p_{i_\mathrm{max}}}{p_{i_\mathrm{min}}}\right) \tag{5}$$

$$i_\mathrm{max} = \operatorname*{argmax}_i \left[Q_{0.025}(p_i)\right] \quad \text{(bin index with highest lower bound } p\text{)} \tag{6}$$

$$i_\mathrm{min} = \operatorname*{argmin}_i \left[Q_{0.975}(p_i)\right] \quad \text{(bin index with lowest upper bound } p\text{)} \tag{7}$$

where p_i denotes the value of p in the i-th bin of x, and $Q_q(p_i)$ denotes the q-th quantile of p_i. This measures how much of the variation of p is explained by x (if we can only consider this one parameter) in a way that is robust to sample size effects. We also quantify our uncertainty in $\mathbb{P}_x$ through Monte Carlo sampling, based on the known distributions of p_{i_max} and p_{i_min} as given in (4).

High-dimensional clustering. To account for interactions between sea state parameters, we use a decision-tree based clustering algorithm to identify rectangular regions in feature space where the rogue wave probability is higher than any probability obtained via univariate analysis.

At its core, the algorithm is a two-step process:

1. Fit a deep random forest classifier to binary data to obtain $\tilde{p}(X)$, which is a rough, noisy estimate of $p(X)$. Here, X denotes the vector of *all* sea state parameters x.

2. Fit a shallow decision tree regressor to $\log \tilde{p}(X)$ (with mean squared error criterion). The leaves of this surrogate model then represent the desired clusters wherein $p(X)$ is approximately constant. We find and retain the 12 leaves with the highest (significant) imbalance between classes.

As this process represents a model search it is vulnerable to overfitting. Therefore, we only use 34% of all available data to identify clusters, and the remaining 66% of the data to analyze the conditions within the cluster (i.e., they determine the final reported rogue wave probability).

This is a conservative process, where all estimators are piecewise constant, which severely limits their learning capabilities. On the other hand, this process should be robust to overfitting, its outputs are easy to analyze (since they just represent another rectangular bin in feature space), and the efficient computation of decision trees ensures that it can scale to billions of data points.

For the decision tree and random forest algorithms, we used the implementations by scikit-learn[47]. The full implementation of our analysis is available as a Jupyter notebook (see Data availability section) that can be used to reproduce all plots in this publication.

Data availability

All preprocessed input data are available at https://doi.org/10.17894/ucph.99bab774-2c97-4e9f-871f-3c349cc0d510. The Jupyter notebook used to generate the results and figures in this report is available at https://doi.org/10.5281/zenodo.4724496.

Received: 30 November 2020; Accepted: 21 April 2021
Published online: 12 May 2021

References

1. Didenkulova, E. Catalogue of rogue waves occurred in the World Ocean from 2011 to 2018 reported by mass media sources. *Ocean Coast. Manag.* https://doi.org/10.1016/j.ocecoaman.2019.105076 (2019).
2. Wang, L., Li, J., Liu, S. & Ducrozet, G. Statistics of long-crested extreme waves in single and mixed sea states. *Ocean Dyn.* https://doi.org/10.1007/s10236-020-01418-9 (2020).
3. Orzech, M. D. & Wang, D. Measured rogue waves and their environment. *J. Mar. Sci. Eng.* **8**, 890. https://doi.org/10.3390/jmse8110890 (2020).
4. Støle-Hentschel, S., Trulsen, K., Nieto Borge, J. C. & Olluri, S. Extreme wave statistics in combined and partitioned windsea and swell. *Water Waves* https://doi.org/10.1007/s42286-020-00026-w (2020).
5. Karmpadakis, I., Swan, C. & Christou, M. Assessment of wave height distributions using an extensive field database. *Coast. Eng.* https://doi.org/10.1016/j.coastaleng.2019.103630 (2020).
6. McAllister, M. L. & van den Bremer, T. S. Experimental study of the statistical properties of directionally spread ocean waves measured by buoys. *J. Phys. Oceanogr.* **50**, 399–414. https://doi.org/10.1175/JPO-D-19-0228.1 (2019).
7. McAllister, M. L., Draycott, S., Adcock, T. A. A., Taylor, P. H. & Bremer, T. S. V. D. Laboratory recreation of the Draupner wave and the role of breaking in crossing seas. *J. Fluid Mech.* **860**, 767–786. https://doi.org/10.1017/jfm.2018.886 (2019).
8. Cousins, W. & Sapsis, T. P. Reduced-order precursors of rare events in unidirectional nonlinear water waves. *J. Fluid Mech.* **790**, 368–388. https://doi.org/10.1017/jfm.2016.13 (2016).
9. Chabchoub, A., Hoffmann, N. P. & Akhmediev, N. Rogue wave observation in a water wave tank. *Phys. Rev. Lett.* **106**, 204502. https://doi.org/10.1103/PhysRevLett.106.204502 (2011).
10. Toffoli, A. *et al.* Evolution of weakly nonlinear random directional waves: Laboratory experiments and numerical simulations. *J. Fluid Mech.* **664**, 313–336. https://doi.org/10.1017/S002211201000385X (2010).
11. Cattrell, A. D., Srokosz, M., Moat, B. I. & Marsh, R. Can rogue waves be predicted using characteristic wave parameters?. *J. Geophys. Res. Oceans* **123**, 5624–5636. https://doi.org/10.1029/2018JC013958 (2018).
12. Benetazzo, A. *et al.* On the shape and likelihood of oceanic rogue waves. *Sci. Rep.* **7**, 8276. https://doi.org/10.1038/s41598-017-07704-9 (2017).
13. Fedele, F., Brennan, J., Ponce de León, S., Dudley, J. & Dias, F. Real world ocean rogue waves explained without the modulational instability. *Sci. Rep.* **6**, 27715. https://doi.org/10.1038/srep27715 (2016).
14. Cavaleri, L. *et al.* The Draupner wave: A fresh look and the emerging view. *J. Geophys. Res. Oceans* **121**, 6061–6075. https://doi.org/10.1002/2016JC011649 (2016).
15. Adcock, T. A. A. & Taylor, P. H. The physics of anomalous ('rogue') ocean waves. *Rep. Prog. Phys.* **77**, 105901. https://doi.org/10.1088/0034-4885/77/10/105901 (2014).
16. Xiao, W., Liu, Y., Wu, G. & Yue, D. K. P. Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution. *J. Fluid Mech.* **720**, 357–392. https://doi.org/10.1017/jfm.2013.37 (2013).
17. Gemmrich, J. & Garrett, C. Dynamical and statistical explanations of observed occurrence rates of rogue waves. *Nat. Hazards Earth Syst. Sci.* **11**, 1437–1446. https://doi.org/10.5194/nhess-11-1437-2011 (2011).
18. Casas-Prat, M. & Holthuijsen, L. H. Short-term statistics of waves observed in deep water. *J. Geophys. Res. Oceans* https://doi.org/10.1029/2009JC005742 (2010).
19. Häfner, D., Gemmrich, J. & Jochum, M. FOWD: A Free Ocean Wave Dataset for data mining and machine learning (2021, submitted). Preprint available at arXiv:2011.12071.
20. Behrens, J., Thomas, J., Terrill, E., Jensen, R. & CDIP: maintaining a robust and reliable ocean observing buoy network. In IEEE/OES Twelfth Current. *Waves and Turbulence Measurement (CWTM)* **1–5**, 2019. https://doi.org/10.1109/CWTM43797.2019.8955166 (2019).
21. Tayfun, M. A. Distribution of large wave heights. *J. Waterway Port Coastal Ocean Eng.* **116**, 686–707. https://doi.org/10.1061/(ASCE)0733-950X(1990)116:6(686) (1990).
22. Tayfun, M. A. & Fedele, F. Wave-height distributions and nonlinear effects. *Ocean Eng.* **34**, 1631–1649. https://doi.org/10.1016/j.oceaneng.2006.11.006 (2007).
23. Rodriguez, G., Soares, C. G., Pacheco, M. & Pérez-Martell, E. Wave height distribution in mixed sea states. *J. Offshore Mech. Arctic Eng.* **124**, 34–40. https://doi.org/10.1115/1.1445794 (2002).
24. Gramstad, O. & Trulsen, K. Can swell increase the number of freak waves in a wind sea?. *J. Fluid Mech.* **650**, 57–79. https://doi.org/10.1017/S0022112009993491 (2010).
25. Forristall, G. Z. Wave crest distributions: Observations and second-order theory. *J. Phys. Oceanogr.* **30**, 1931–1943 https://doi.org/10.1175/1520-0485(2000)030<1931:WCDOAS>2.0.CO;2 (2000).

26. Janssen, P. & Bidlot, J.-R. On the Extension of the Freak Wave Warning System and Its Verification https://doi.org/10.21957/uf1sybog (2009).
27. Kharif, C. & Pelinovsky, E. Physical mechanisms of the rogue wave phenomenon. *Eur. J. Mech. B/Fluids* **22**, 603–634. https://doi.org/10.1016/j.euromechflu.2003.09.002 (2003).
28. Janssen, P. A. E. M. Nonlinear Four-Wave Interactions and Freak Waves. *J. Phys. Oceanogr.* **33**, 863–884 https://doi.org/10.1175/1520-0485(2003)33<863:NFIAFW>2.0.CO;2 (2003).
29. Mori, N. & Janssen, P. A. E. M. On kurtosis and occurrence probability of freak waves. *J. Phys. Oceanogr.* **36**, 1471–1483. https://doi.org/10.1175/JPO2922.1 (2006).
30. Fedele, F. & Tayfun, M. A. On nonlinear wave groups and crest statistics. *J. Fluid Mech.* **620**, 221–239. https://doi.org/10.1017/S0022112008004424 (2009).
31. Gramstad, O., Bitner-Gregersen, E., Trulsen, K. & Nieto Borge, J. C. Modulational instability and rogue waves in crossing sea states. *J. Phys. Oceanogr.* **48**, 1317–1331. https://doi.org/10.1175/JPO-D-18-0006.1 (2018).
32. Christou, M. & Ewans, K. Field measurements of rogue water waves. *J. Phys. Oceanogr.* **44**, 2317–2335. https://doi.org/10.1175/JPO-D-13-0199.1 (2014).
33. Stansell, P. Distributions of freak wave heights measured in the North Sea. *Appl. Ocean Res.* **26**, 35–48. https://doi.org/10.1016/j.apor.2004.01.004 (2004).
34. Kim, T.-H. & White, H. On more robust estimation of skewness and kurtosis. *Finance Res. Lett.* **1**, 56–73. https://doi.org/10.1016/S1544-6123(03)00003-5 (2004).
35. Perlin, M., Choi, W. & Tian, Z. Breaking waves in deep and intermediate waters. *Ann. Rev. Fluid Mech.* **45**, 115–145. https://doi.org/10.1146/annurev-fluid-011212-140721 (2013).
36. Banner, M. L., Gemmrich, J. R. & Farmer, D. M. Multiscale measurements of ocean wave breaking probability. *J. Phys. Oceanogr.* **32**, 3364–3375 https://doi.org/10.1175/1520-0485(2002)032<3364:MMOOWB>2.0.CO;2 (2002).
37. Tayfun, M. A. Narrow-band nonlinear sea waves. *J. Geophys. Res. Oceans* **85**, 1548–1552. https://doi.org/10.1029/JC085iC03p01548 (1980).
38. Janssen, T. T. & Herbers, T. H. C. Nonlinear wave statistics in a focal zone. *J. Phys. Oceanogr.* **39**, 1948–1964. https://doi.org/10.1175/2009JPO4124.1 (2009).
39. Trulsen, K., Zeng, H. & Gramstad, O. Laboratory evidence of freak waves provoked by non-uniform bathymetry. *Phys. Fluids* **24**, 097101. https://doi.org/10.1063/1.4748346 (2012).
40. McAllister, M. L. Lagrangian measurement of steep directionally spread ocean waves: Second-order motion of a wave-following measurement buoy. *J. Phys. Oceanogr.* **49**, 3087–3108. https://doi.org/10.1175/JPO-D-19-0170.1 (2019).
41. Onorato, M., Proment, D. & Toffoli, A. Triggering rogue waves in opposing currents. *Phys. Rev. Lett.* **107**, 184502. https://doi.org/10.1103/PhysRevLett.107.184502 (2011).
42. Onorato, M. & Proment, D. Approximate rogue wave solutions of the forced and damped nonlinear Schrödinger equation for water waves. *Phys. Lett. A* **376**, 3057–3059. https://doi.org/10.1016/j.physleta.2012.05.063 (2012).
43. Longuet-Higgins, M. S. On the statistical distribution of the height of sea waves. *JMR* **11**, 245–266 (1952).
44. Cattrell, A. *Increasing Maritime Safety with Improved Understanding of Rogue Waves*. Ph.D. thesis, University of Southampton (2020).
45. Teutsch, I., Weisse, R., Moeller, J. & Krueger, O. A statistical analysis of rogue waves in the southern North Sea. *Nat. Hazards Earth Syst. Sci.* **20**, 2665–2680. https://doi.org/10.5194/nhess-20-2665-2020 (2020).
46. Serio, M., Onorato, M., Osborne, A. R. & Janssen, P. A. On the computation of the Benjamin–Feir Index. *Nuovo Cimento della Societa Italiana di Fisica C* **28**, 893–903. https://doi.org/10.1393/ncc/i2005-10134-1 (2005).
47. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

Dion Häfner was supported by the Danish Hydrocarbon Research and Technology Centre (DHRTC). Raw data were furnished by the Coastal Data Information Program (CDIP), Integrative Oceanography Division, operated by the Scripps Institution of Oceanography, under the sponsorship of the U.S. Army Corps of Engineers and the California Department of Parks and Recreation. Computational resources were provided by DC^3, the Danish Center for Climate Computing. We thank 3 anonymous reviewers for their insightful comments.

Author contributions

M.J. and J.G. conceived the project. D.H. drafted, implemented, and executed the analysis. All authors interpreted the results. D.H. drafted the manuscript. All authors reviewed the manuscript.

Competing interests

Dion Häfner's work has been funded by the Danish Hydrocarbon Research and Technology Centre (DHRTC). Johannes Gemmrich and Markus Jochum declare no potential competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-89359-1.

Correspondence and requests for materials should be addressed to D.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.3 ARTICLE III — A CAUSAL PREDICTIVE MODEL FOR REAL-WORLD ROGUE WAVE PROBABILITIES

Figure 2.8: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"the great wave"*.

The final article in this series (to be submitted) puts the previous findings on a more rigorous causal foundation, and demonstrates how we can use our results to arrive at a better rogue wave forecast.

For a well-performing predictive model we need to relax the assumption of independent regions in parameter space (as we used in the previous articles to study how the rogue wave probability p depends on the sea state). Instead, we would now like to interpolate between data points through an artificial neural network, but this introduces a considerable risk of overfitting. To mitigate this, we perform a causal analysis based on the state-of-the-art in rogue wave research (as presented in § 1.1) to include only *direct causes* of rogue waves in the model. To limit the number of possible parameter interactions and combat overfitting we employ a multi-head neural network, where only parameters sharing the same input head can interact non-additively with each other.

To quantify how well the trained model captures the causal structure we apply a procedure that is inspired by invariant causal prediction (ICP; Peters, Bühlmann, and Meinshausen, 2016; Peters, Janzing, and Schölkopf, 2017). We search for a model that stays approximately invariant under re-training on different environments (such as summer vs. winter conditions, or deep water vs. shallow water). This allows us to identify a model that represents a good trade-off between predictive performance and invariance.

Additionally, we visualize the prediction surface of this model and compare to the findings in article 2, which largely confirms earlier results, but also leads to some new insights on the nature of higher-order corrections due to parameter interactions. On top of this, we identify a crucial interaction between crest-trough correlation and directionality index that has (to our knowledge) not been described before.

A Causal Predictive Model for Real-World Rogue Wave Probabilities

Dion Häfner*

Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

Johannes Gemmrich

University of Victoria, Victoria, British Columbia, Canada

Markus Jochum

Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

Extreme waves in the ocean ("rogue waves") are well studied in theory and lab experiments under idealized conditions, but relatively little research is based on direct observations. Therefore, it is still unclear how well common approximations hold up in the real ocean. Here, we present a predictive model that combines parameters representing both linear and nonlinear wave dynamics through an artificial neural network, trained directly on observations from wave buoys. By imposing strong architectural constraints we arrive at a model that is approximately invariant under retraining across a wide regime of sea states, suggesting causal consistency with the dominant rogue wave generation processes, and that we can analyze in detail. We find that a combination of crest-trough correlation, characteristic steepness, directionality index, and relative water depth has high causal consistency, outputs well-calibrated probabilities, and achieves good predictive scores on unseen data. This paves the way towards a higher quality rogue wave forecast.

1. Introduction

Oceanic rogue waves (also called freak waves) are extreme ocean waves that are suspected to have caused countless accidents, often with fatal consequences (Didenkulova 2019). They are typically defined as any wave whose crest-to-trough height H exceeds a certain threshold relative to the significant wave height H_s. The significant wave height in turn is defined as 4 times the standard deviation of the sea surface elevation. Here, we use a rogue wave criterion with a threshold of 2.0 instead of the perhaps more common 2.2 in order to increase the number of studied rogue waves (but our results can be extended to the stricter threshold):

$$H/H_s > 2.0 \tag{1}$$

A rogue wave is therefore by definition an unlikely sample from the tail of the wave height distribution (with a probability of about 3×10^{-4} under linear theory, Longuet-Higgins 1952). This also implies that rogue waves are an extreme event that can in principle occur by chance under any circumstance, which makes them difficult to analyze, and requires massive amounts of data. Therefore, research has mostly focused on theory and idealized experiments in wave tanks, often considering only 1-dimensional wave propagation (see Dudley et al. 2019, for a review).

In previous work we assembled a database of over 1 billion wave observations from buoys (FOWD, Häfner et al. 2021a). FOWD is a catalogue that maps individual wave observations to about 80 characteristic sea state parameters describing the circumstances under which the wave was measured. This allowed us to analyze how rogue wave occurrence probabilities depend on the sea state, where we found that crest-trough correlation is by far the best univariate predictor for rogue wave occurrence (Häfner et al. 2021b). This parameter is not included in today's operational freak wave forecasts such as that of the European Centre for Medium-Range Weather Forecasts (ECMWF), which instead focuses on nonlinear effects governed by parameters like characteristic steepness, directionality index, and the Benjamin-Feir index (ECMWF 2021). This suggests that an improved rogue wave forecast that takes crest-trough correlation into account is within reach. However, all analysis of FOWD so far has been purely descriptive and cannot be used for forecasting.

In this study, we present a neural network-based machine learning model that predicts rogue wave probabilities from the sea state, trained solely on observations. The resulting model respects the causal structure of rogue wave generation — that means it is robust to distributional shift and can be used to infer the relative importance of rogue wave generation mechanisms. We achieve this by combining a careful a-priori analysis of causal pathways that leads to

*Corresponding author: Dion Häfner, dion.haefner@nbi.ku.dk a set of presumed causal parameters (Section 2), regularization constraints, and an a-posteriori model evaluation to identify the model architecture that shows the highest invariance to shifting environments (Section 3). Analyzing this model allows us to study how rogue wave probabilities depend on the sea state, even in higher-dimensional settings (Section 4).

2. The causes of rogue waves

To ensure that our machine learning model learns causal relationships instead of mere associations, it is essential to only include parameters that carry causal meaning (otherwise the model might prefer spurious associations that are easier to learn).

There are many suspected causes of rogue waves (see Adcock and Taylor 2014, for an overview). Typically, research focuses on linear bandwidth-limited seas (Tayfun and Fedele 2007), weakly nonlinear seas (Gemmrich and Garrett 2011; Fedele et al. 2016), or the highly nonlinear modulational instability (Onorato et al. 2006). Apart from these universal mechanisms, there are also countless possible interactions with localized features like topography such as (Trulsen et al. 2012) or underwater currents, interactions with currents like in the Agulhas (Mallory 1974) or in Drake passage (Didenkulova et al. 2021), or crossing sea states at high crossing angles (McAllister et al. 2019).

The go-to tool to analyze causal relationships is a causal DAG (directed acyclic graph), where nodes represent variables and edges $A \rightarrow B$ imply that A is a cause of B (usually in the probabilistic sense in that the probability distribution $P(B)$ depends on A). In the frame of this analysis, we would like to relate sea state parameters $\mathcal{P}$ to physical effects Φ, which in turn influence wave observations O. The resulting causal graph for rogue waves containing the previously discussed pathways is shown in Fig. 1.

Following this causal structure, we use the following set of sea state parameters as candidates for representing the various causal pathways (see Appendix A for more information on each parameter):

Crest-trough correlation r (a parameter related to spectral bandwidth) to account for linear effects. As we showed in Häfner et al. (2021b), the rogue wave probability p is conditionally independent of other bandwidth measures (such as narrowness and peakedness) when conditioning on r, but not vice-versa. This suggests that r is the dominant causal factor behind linear rogue wave formation.

Steepness ε governing weakly nonlinear effects (such as second-order and third-order bound waves) and wave breaking (Miche 1944; Goda 2010).

Relative high-frequency energy E_h (fraction of total energy contained in the spectral band 0.25 Hz to 1.5 Hz) as a proxy for the strength of local winds.

Relative depth $\widetilde{D}$ (based on peak wavenumber), which is central for nonlinear effects (Korteweg and De Vries 1895; Janssen 2018) and wave breaking (Miche 1944).

Benjamin-Feir index BFI which controls third-order nonlinear free waves (Janssen 2018) and the modulational instability (Janssen 2003).

Ursell number Ur which quantifies nonlinear effects in shallow water (Ursell 1953).

Dominant directional spread σ_θ which has an influence on third-order nonlinear waves (Janssen 2018) and wave breaking (McAllister et al. 2019).

Spectral bandwidth ν appearing in the expression for the influence of third-order nonlinear waves (Janssen 2018).

Directionality index R (the ratio of directional spread and spectral bandwidth) controlling third-order nonlinear free waves (often used in conjunction with the BFI, Janssen 2018).

There are some notable omissions from this list. Firstly, mean period and significant wave height, perhaps the most studied sea state parameters of all. While these parameters *do* play an important role in the rogue wave generation process, they appear higher up in the causal graph and are therefore not direct causes of rogue waves (instead, they *generate the conditions* that are causing rogue waves). Secondly, surface elevation skewness and kurtosis, which are also studied extensively in connection with rogue waves (Stansell 2004; Mori and Janssen 2006; Fedele and Tayfun 2009), but which we have shown to be too noisy to be of use for rogue wave prediction (Häfner et al. 2021b).

Unfortunately, there are still entirely unobserved causal paths, since we do not have access to data on local winds, topography, or currents. Additionally, all measurements are potentially biased estimates of the true sea state parameters. Therefore we cannot rely on the resulting model to be causally consistent by itself, and we will have to perform a-posteriori verification on the learned model to check for causal consistency across different environments (see Section 3c).

3. A causally consistent predictive model

a. Input data

The main data source for this analysis is the Free Ocean Wave Dataset (FOWD, Häfner et al. 2021a). The pre-filtered version of FOWD consists of 1.4 billion wave measurements, originally recorded by 158 CDIP wave buoys (Behrens et al. 2019) along the Pacific and Atlantic coasts of the US, Hawaii, and overseas US territories at water depths between 10 m to 4000 m and a minimum significant wave height of 1 m. Each buoy records the sea surface elevation at

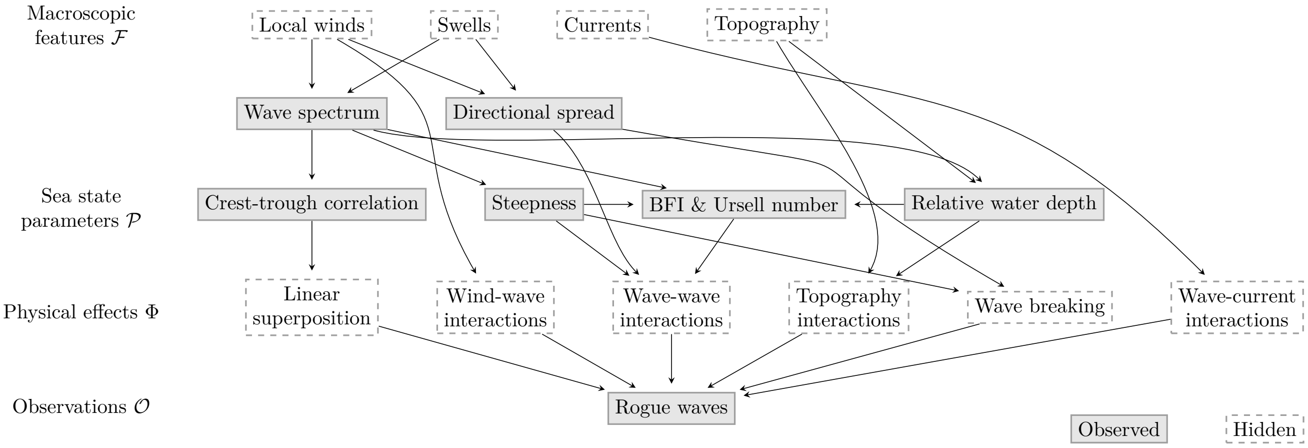


FIG. 1. The causes of rogue waves as a causal DAG (directed acyclic graph). Arrows $A \longrightarrow B$ imply that A causes B.

a sampling frequency of 1.28 Hz, which amounts to over 700 yr of time series in total. FOWD then extracts every zero-crossing wave from the CDIP data and computes about 20 characteristic sea state parameters from the history of the wave within a 10 min, 30 min, and dynamically sized window (as suggested in Boccotti 2000). Here, we only use parameters based on the dynamic window size.

Due to the massive data volume of the full FOWD catalogue (~ 1 TB), we use an aggregated version that maps each sea state to the maximum wave height of the following 100 waves (similar to what is used in Häfner et al. 2021b). This reduces the data volume by a factor of 100 and inflates all rogue wave probabilities, which can be corrected for via $p = 1 - (1 - \hat{p})^{1/100}$ where $\hat{p}$ is the inflated probability, assuming that rogue waves occur independently from each other.

This process leaves us with 12.9M data points containing just over 100,000 rogue waves exceeding $2H_s$. This input dataset is freely available for download at `https://erda.ku.dk/archives/7072a6eb3d181149deb56b7d8739805e/published-archive.html`.

b. Model architecture

Our core assumption is that the rogue wave probability can be modelled as:

$$\operatorname{logit} P\big(y = 1 \,\big|\, \mathbf{x}\big) \sim \sum_i f_i\big(\mathbf{x}^{(S_i)}\big) + b \tag{2}$$

where y is a binary label indicating whether the current wave is a rogue wave, $\mathbf{x}^{(S_i)}$ denotes the i-th subset of all causal sea state parameters $\mathbf{x}$ (see Section 2), $\operatorname{logit}(p) = \log(p) - \log(1-p)$ is the logit function, f_i are arbitrary nonlinear functions to be learned, and b is a constant bias term. By including only a subset $\mathbf{x}^{(S_i)}$ of all parameters $\mathbf{x}$ as input for each term we can limit which parameters may interact with each other as an additional regularizing constraint.

For example, to include the effects of linear superposition and nonlinear corrections for free and bound waves similar to ECMWF (2021) we can use:

$$\operatorname{logit} P\big(y = 1 \,\big|\, \mathbf{x}\big) \sim \underbrace{f_1(r)}_{\text{linear}} + \underbrace{f_2(\mathrm{BFI}, R)}_{\text{free waves}} + \underbrace{f_3(\varepsilon, \widetilde{D})}_{\text{bound waves}} \tag{3}$$

with Benjamin-Feir index BFI, directionality index R, steepness ε, and relative depth $\widetilde{D}$.

We parametrize the functions f_i via fully connected neural networks (FCNs), which have been shown to be universal function approximators (Hornik 1991), and that can be trained efficiently for large amounts of data. The set of functions f_i can be represented as a single multi-head FCN (one head for each input subset $\mathbf{x}^{(S_i)}$) with a linear output layer (Fig. 2). We use a simple feedforward architecture with 3 hidden layers and ReLU activation functions (rectified linear unit, Nair and Hinton 2010).

The neural network outputs a scalar $\tilde{p} = \operatorname{logit} P\big(y = 1 \,|\, \mathbf{x}\big) \in (-\infty, \infty)$, the log-odds of a rogue wave occurrence for the given sea state. For training, we use the Adam optimizer and backpropagation to minimize a typical cross-entropy loss for binary classification with added ℓ_1 and ℓ_2 regularization terms for kernel parameters:

$$\begin{aligned} L(p, y, \theta) &= y \cdot \log(p) + (1-y) \cdot \log(1-p) \\ &\quad + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2 \end{aligned} \tag{4}$$

with predicted probability $p = \operatorname{logit}^{-1}(\tilde{p})$, observed labels $y \in \{0, 1\}$ (rogue wave or not), and neural network kernel parameters θ.

To estimate uncertainties in the neural network parameters and resulting predictions, we use Gaussian stochastic weight averaging (SWAG, Maddox et al. 2019). For this,

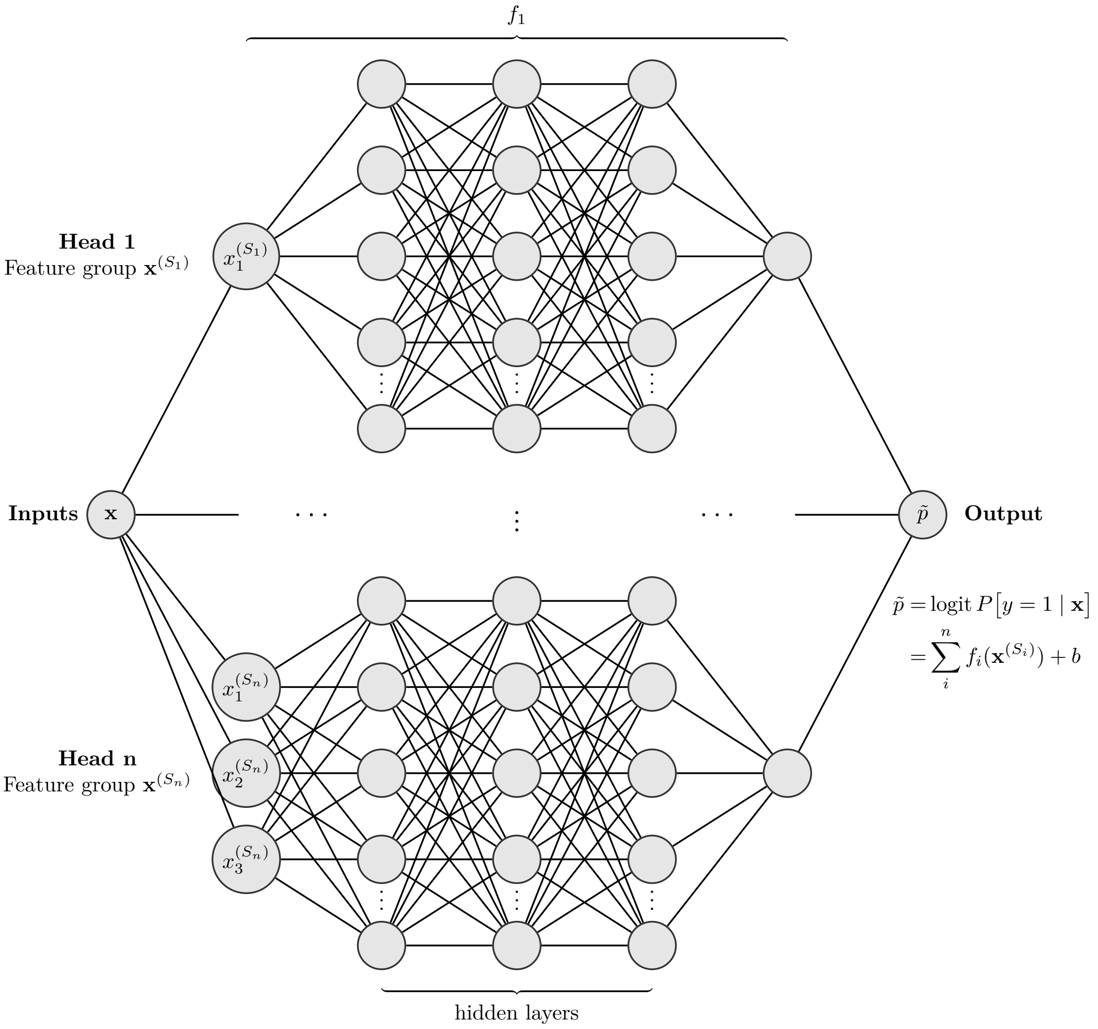

f_1
Head 1
Feature group $\mathbf{x}^{(S_1)}$
$x_1^{(S_1)}$
Inputs
$\mathbf{x}$
Output
$\tilde{p}$
$\tilde{p} = \text{logit}\, P\big[y = 1 \mid \mathbf{x}\big]$
$= \sum_i^n f_i(\mathbf{x}^{(S_i)}) + b$
Head n
Feature group $\mathbf{x}^{(S_n)}$
$x_1^{(S_n)}$
$x_2^{(S_n)}$
$x_3^{(S_n)}$
hidden layers

FIG. 2. Neural network architecture (multi-head FCN) used to predict rogue wave probabilities. Each input head receives a different subset of the full parameter set **x** to limit the amount of non-causal interactions between parameters.

we train the network for 50 epochs, then start recording the optimizer trajectory after each epoch for another 50 epochs. The observed covariance structure of the sampled parameters is then used to construct a multivariate Gaussian approximation of the loss surface close to the minimum that we can sample from. This results in slightly better predictions, but more importantly, also gives us a way to quantify how confident the neural network is in its predictions.

Appendix B lists the full set of model hyperparameters.

c. Causal consistency

Even though we only include input parameters that we assume to have a direct causal connection with rogue wave generation, there is no guarantee that the neural network will learn the correct causal connections. In fact, the presence of measurement bias and unobserved causal paths makes it unlikely that the model will converge to the true causal structure unless the right amount of regularization and architectural constraints are applied. To search for a maximally causally consistent model we will have to quantify its causal consistency.

We can achieve this through the concept of invariant causal prediction (ICP, Peters et al. 2016, 2017). The key insight behind ICP is that only the true causal model will be invariant under distributional shift, in the sense that re-training the model on data with different correlations *between* features should still lead to the same dependency of the target *on* the features. Ideally, the chosen environments lead to a significant distributional shift by changing the relative importance of different causal mechanisms, but not the causal structure itself.

To exploit ICP we split the full dataset randomly into separate training and validation sets, in chunks of 1M waves (to ensure that the validation data is truly unseen by the model, but covers roughly the same range of sea states). We train the model on the training dataset (66 % of all

TABLE 1. The subsets of the validation data set used to evaluate model invariance.

Subset name	Condition	# waves
southern-california	Longitude $\in (-123.5, -117)°$, Latitude $\in (32, 38)°$	233M
deep-stations	Water depth > 1000 m	33M
shallow-stations	Water depth < 100 m	138M
summer	Day of year $\in (160, 220)$	44M
winter	Day of year $\in (0, 60)$	88M
Hs > 3m	$H_s > 3$ m	55M
high-frequency	Relative swell energy < 0.15	40M
low-frequency	Relative swell energy > 0.7	42M
long-period	Mean zero-crossing period > 9 s	40M
short-period	Mean zero-crossing period < 6 s	90M
cnoidal	Ursell number > 8	34M
weakly-nonlinear	Steepness > 0.04	80M
spectral-narrow	Directionality index < 0.3	68M
spectral-wide	Directionality index > 1	37M
full	(all validation data)	438M

data) and perform ICP on the validation dataset (34 % of all data), which we partition into subsets representing different conditions in space, time, depth, mean period, and degrees of non-linearity (Table 1). Then, we re-train the model separately on each subset and compare the performance on the k-th data subset $\mathbf{x}_{(k)}$ between the re-trained model P_k and the full model P_{tot} by computing the root-mean-square error of predictions:

$$\mathcal{E}_k^2 = \frac{1}{n_k}\sum_i^{n_k}\left(\operatorname{logit} P_k\left(\mathbf{x}_i^{(k)}\right) - \operatorname{logit} P_{\text{tot}}\left(\mathbf{x}_i^{(k)}\right)\right)^2 \qquad (5)$$

where n_k is the number of data points in the subset $\mathbf{x}_{(k)}$. As the total consistency score we use the root-mean-square across all environments:

$$\mathcal{E} = \sqrt{\frac{1}{n_E}\sum_k^{n_E}\mathcal{E}_k^2} \qquad (6)$$

Under a noise-free, infinite dataset and an unbiased training process that identifies the true causal model we would find $\mathcal{E} = 0$, i.e., re-training the model on the unseen data subset would not contribute any new information and leave the model perfectly invariant. Since all 3 of these assumptions are violated here, we merely search for the model that minimizes $\mathcal{E}$, while also having a competitive predictive score and well-calibrated probabilities (since for example a trivial model predicting $P(\mathbf{x}) = c$ with constant c has perfect invariance of predictions).

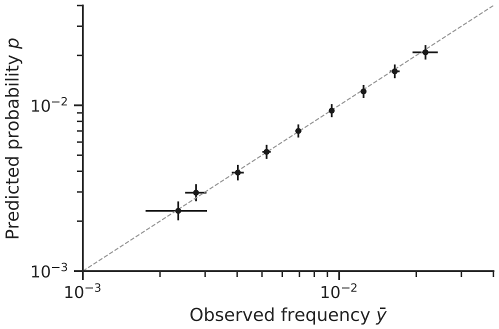


FIG. 3. Our model outputs well-calibrated probabilities. Shown is the binned predicted probability p vs. the observed rogue wave incidence $\bar{y}$. Error bars for p indicate 3 standard deviations estimated via SWAG sampling. Error bars for $\bar{y}$ indicate 99 % credible interval assuming $\bar{y}_i \sim \text{Beta}(n_i^+, n_i^-)$ with n_i^+ rogue and n_i^- non-rogue measurements in the i-th bin. Dashed line indicates perfect calibration.

d. Other performance metrics

We also need performance metrics that describe the predictive capabilities of the model and measure overfitting, specifically whether the model tends to produce overconfident or underconfident predictions.

To evaluate predictive performance we use the mean of a base rate-adjusted log-likelihood score across all environments (Table 1):

$$\mathcal{L}(p, \bar{y}) = \frac{1}{n_E}\sum_k^{n_E}\left(I(p_k) - I(\bar{y}_k)\right) \qquad (7)$$

$$I(x) = x \cdot \log(x) + (1 - x) \cdot \log(1 - x) \qquad (8)$$

with predicted probabilities p and base rate $\bar{y} = \frac{1}{n}\sum_i^n y$ for each environment. This gives the log-likelihood of observing the data given the model, relative to a model that just predicts its base rate $\bar{y}$. Using the mean over environments ensures that the model performs well in different physical regimes instead of focusing on average conditions. Since we do not have access to ground truth probabilities we cannot quantify predictive quality in an absolute sense, and the above score can only be used to compare different models on the same data.

To evaluate model calibration we compute a calibration curve by binning the predicted probabilities and comparing each bin to the observed rogue wave frequency (Fig. 3). As the total calibration score we use the root-mean-square error between measured and predicted log-odds:

$$C = \sqrt{\frac{1}{n_b}\sum_{k=1}^{n_b}\left(\operatorname{logit}(p_i) - \operatorname{logit}(\bar{y}_i)\right)^2} \qquad (9)$$

with number of bins n_b and corresponding mean prediction p_i and observed base rate $\bar{y}_i$.

4. Results

We train a total of 24 different candidate models using the procedure described in Section 3 and evaluate their performance in terms of calibration, data likelihood, and causal consistency (Table 2).

We observe a clear anti-correlation between model complexity and predictive score on the one hand and causal consistency on the other hand, as long as only causal parameters are included (as identified in Section 2). When including additional non-causal parameters (experiment 24), the predictive score decreases slightly while the causal consistency error increases drastically (as expected).

Still, all experiments show a significant non-zero consistency error $\mathcal{E}$ of around 0.1 (mean SWAG uncertainties in predictions are about 0.03). Selecting among the remaining models is therefore a trade-off between bias (prediction score) and variance (consistency score). There are 2 models that represent a compromise with good performance in all metrics:

1. Model 16 with two parameter groups $S_1 = \{r, R\}$, $S_2 = \{\varepsilon, \widetilde{D}, R\}$.
2. Model 17 with a single parameter group $S_1 = \{r, \varepsilon, \widetilde{D}, R\}$.

We apply Occam's razor and choose model 16 as the reference model for further analysis — even though it has slightly lower predictive and calibration scores — due to the lower number of possible parameter interactions (at most 3-way interactions instead of 4-way).

Our analysis of this reference model and selected other experiments leads to the following 3 main results of this study.

a. *Rogue wave models should account for crest-trough correlation, steepness, relative depth, and directionality*

Only this parameter combination achieves good causal consistency and predictive scores at the same time, and experiments that exclude any of these parameters perform unconditionally worse (see e.g. experiments 1–6). Especially the exclusion of crest-trough correlation leads to catastrophic results, even when including other bandwidth measures in its place (experiments 3, 13).

Models that use the directionality index R over raw directional spread σ_θ are generally more causally consistent (experiment 17 vs. 19). Also, an interaction between crest-trough correlation and directionality index seems essential to achieve optimal performance (12 vs. 16). Higher predictive performance can be achieved by including spread σ_θ and spectral narrowness ν directly (instead of via $R = \sigma_\theta^2/2\nu^2$), at the expense of increased overfitting (experiment 21).

This suggests that this set of parameters represents the dominant rogue wave generation processes in the form of linear bandwidth-limited superposition with a directional correction (r, R) and weakly nonlinear corrections $(\varepsilon, \widetilde{D}, R)$. This is consistent with other empirical studies such as Fedele et al. (2019), which consider the same parameters in conjunction with rogue crests during storms (except crest-trough correlation, which is not meaningful for crest heights).

This set of parameters is also similar to the ingredients to ECMWF's freak wave forecast (ECMWF 2021), which is based on second and third-order bound and free waves and uses steepness, relative depth, directional spread, and spectral bandwidth. However, in our model these parameters are combined differently; a model enforcing the same interactions (steepness and relative depth for bound wave contribution, BFI and directionality index for free wave contribution) performs poorly, even in the deep-water regime where the BFI is most applicable (experiment 14).

Numerous previous studies have found the BFI to be a poor predictor of rogue wave risk in realistic sea states (Fedele et al. 2016, 2019; Gramstad and Trulsen 2007; Xiao et al. 2013; Häfner et al. 2021b) due to its strong underlying assumptions (such as unidirectionality). This study extends this to the fully nonparametric and nonlinear case, in which the predictive qualities of the BFI remain low, even when including interactions with directionality. This suggests that the contribution of the modulational instability and third-order free waves to rogue wave risk is negligible.

We study how our model uses different parameters by visualizing their impact on the prediction of the respective head of the neural network. For this, we make use of a functional decomposition called accumulated local effects (ALE, Apley and Zhu 2019), which measures the influence of infinitesimal changes in each parameter on the prediction outcome (see also Molnar 2020). This removes correlations between parameters, so we can plot the contribution of each parameter and 2-parameter interaction in isolation, where the total effect is the sum of all individual contributions. For example, a total value of ALE = 1 implies a predicted rogue wave probability that is about $\exp(1) = 2.71$ times higher than baseline.

From the ALE plot (Fig. 4), we find that crest-trough correlation has by far the biggest influence and explains about 1 order of magnitude in rogue wave risk variation (as already observed in Häfner et al. 2021b). To first order, higher crest-trough correlation, lower directionality index, higher relative depth, and higher steepness lead to higher rogue wave risk, but parameter interactions can lead to more complicated, non-monotonic relationships (for example in very shallow water, see Section 4c).

Perhaps the most surprising finding is the strong interaction between crest-trough correlation and directionality index. This could imply that directional corrections are necessary for the accurate modelling of linear interactions, and could suggest a promising line for further theoretical research.

TABLE 2. Full list of experiments. $\mathcal{L}$: Prediction score (higher is better). $\mathcal{E}$: Invariance error (lower is better). $\mathcal{C}$: Calibration error (lower is better). Color coding ranges between (median − IQR, median + IQR) with inter-quartile range IQR.

	Feature groups			Scores		
ID	1	2	3	$\mathcal{L} \times 10^4$	$\mathcal{E} \times 10^2$	$\mathcal{C} \times 10^2$
1	$\{r\}$			4.62	8.52	6.90
2	$\{r,\ R\}$			5.05	8.58	3.86
3	$\{\varepsilon,\ \widetilde{D},\ R\}$			0.03	22.59	6.21
4	$\{r,\ \widetilde{D},\ R\}$			5.56	7.95	4.34
5	$\{r,\ \varepsilon,\ R\}$			5.49	8.83	3.83
6	$\{r,\ \varepsilon,\ \widetilde{D}\}$			5.35	8.89	7.05
7	$\{r,\ R\}$	$\{\varepsilon,\ \widetilde{D}\}$		5.77	9.19	4.46
8	$\{r,\ R,\ \mathrm{Ur}\}$			5.70	7.99	3.94
9	$\{r,\ R\}$	$\{\mathrm{Ur},\ R\}$		5.64	7.49	4.31
10	$\{r,\ R,\ \mathrm{BFI}\}$			5.60	7.75	4.51
11	$\{r,\ R\}$	$\{\mathrm{BFI},\ R\}$		5.46	8.20	4.44
12	$\{r\}$	$\{\varepsilon,\ \widetilde{D},\ R\}$		5.67	9.24	4.67
13	$\{\sigma_f\}$	$\{\varepsilon,\ \widetilde{D},\ R\}$		4.11	12.16	6.30
14	$\{r\}$	$\{\varepsilon,\ \widetilde{D}\}$	$\{\mathrm{BFI},\ R\}$	5.64	9.77	6.02
15	$\{r,\ R\}$	$\{\varepsilon,\ \widetilde{D},\ \sigma_\theta\}$		6.22	10.63	5.20
16	$\{r,\ R\}$	$\{\varepsilon,\ \widetilde{D},\ R\}$		5.87	8.63	3.62
17	$\{r,\ \varepsilon,\ \widetilde{D},\ R\}$			5.98	8.60	2.96
18	$\{r\}$	$\{\varepsilon,\ \widetilde{D}\}$	$\{\mathrm{BFI},\ \sigma_f,\ \sigma_\theta\}$	6.01	11.10	8.43
19	$\{r,\ \varepsilon,\ \widetilde{D},\ \sigma_\theta\}$			5.97	9.71	6.45
20	$\{r,\ \varepsilon,\ \widetilde{D},\ R,\ E_h\}$			6.10	9.14	5.33
21	$\{r,\ \varepsilon,\ \widetilde{D},\ \sigma_\theta,\ \nu\}$			6.31	10.04	4.00
22	$\{r,\ \varepsilon,\ \widetilde{D},\ R,\ \mathrm{BFI}\}$			6.05	8.84	6.81
23	$\{r,\ \varepsilon,\ \widetilde{D},\ \sigma_\theta,\ \sigma_f,\ E_h,$ $\mathrm{BFI},\ R\}$			6.91	12.69	3.68
24	$\{r,\ \varepsilon,\ \widetilde{D},\ \sigma_\theta,\ \sigma_f,\ E_h,$ $H_s,\ \overline{T},\ \kappa,\ \mu,\ \lambda_p\}$			6.70	56.44	7.27

Symbols

r	Crest-trough correlation	ν	Spectral bandwidth (narrowness)
σ_f	Spectral bandwidth (peakedness)	σ_θ	Directional spread
ε	Peak steepness $H_s k_p$	R	Directionality index $\sigma_\theta^2/(2\nu^2)$
BFI	Benjamin-Feir index	$\widetilde{D}$	Relative peak water depth $Dk_p/(2\pi)$
E_h	Relative high-frequency energy	Ur	Ursell number
$\overline{T}$	Mean period	κ	Kurtosis
μ	Skewness	H_s	Significant wave height

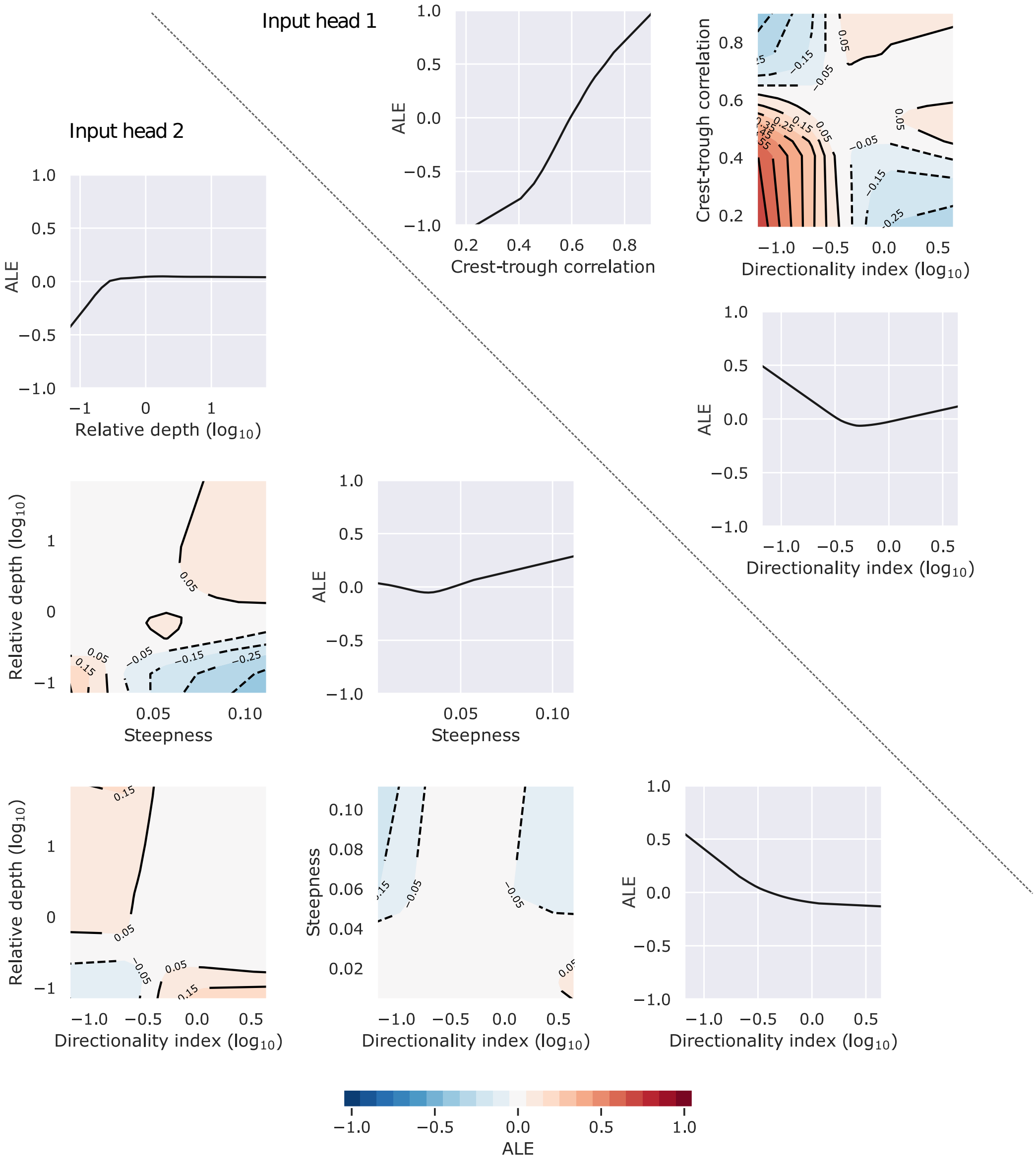


FIG. 4. ALE (accumulated local effects) plot matrix for experiment 16. Shown is the change in rogue wave risk (in logits) from the average as each parameter is varied. The total effect is the sum of all 1D, 2D, and higher-order contributions (not shown).

b. The Rayleigh distribution is an upper bound for real-world rogue wave risk

Despite the clear enhancing properties of weakly nonlinear corrections, the Rayleigh wave height distribution remains an upper bound for real-world (crest-to-trough) rogue waves. The Rayleigh distribution is the theoretical wave height distribution for linear narrowband waves (Longuet-Higgins 1952), i.e., the limit $r \longrightarrow 1$, $\varepsilon \longrightarrow 0$, $\widetilde{D} \longrightarrow \infty$, and $R \longrightarrow 0$, and reads:

$$P(H/H_s > k) = \exp(-2k^2) \qquad (10)$$

Only in the most extreme conditions does our model predict a similarly high probability, for example for $R = 0.05$, $\varepsilon = 0.01$, $r = 0.85$, and $\widetilde{D} = 0.2$, which gives $p = 3.2 \times 10^{-4}$ (compared to 3.3×10^{-4} for Rayleigh distributed waves).

In the opposite extreme, rogue wave probabilities can fall to as little as 10^{-5} for low values of r and high values of R (as e.g. in a sea with a strong high-frequency component and high directional spread). This suggests that bandwidth effects can create sea states that efficiently suppress extremes — a fact that could lead to safer shipping routes should a forecast based on these parameters become available.

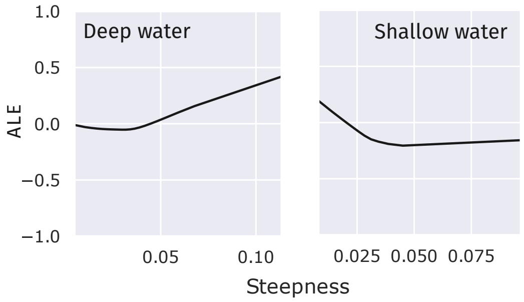


Fig. 5. Our model predicts a positive association between steepness and rogue waves in deep water, and a negative association in shallow water. Shown is the 1-dimensional ALE (accumulated local effects) plot in both cases. Here, deep water are sea states with $\widetilde{D} > 3$ and shallow water with $\widetilde{D} < 0.1$.

c. There is a clear separation between deep water and shallow water regimes

The inclusion of an interaction between steepness and relative water depth is essential for predictive invariance across several environments. Looking at this more closely, we find that a stratification on deep and shallow water sea states reveals 2 distinct regimes (Fig. 5).

In deep water, rogue wave risk is strongly positively associated with steepness, as expected from the contribution of second and third-order nonlinear bound waves (Janssen 2018).

The opposite is true in shallow water, where we find a clear *negative* association with steepness. This is less expected, since nonlinear effects are typically considered to *increase* rogue wave occurrence. But the theoretical expression for the contribution of wave-induced currents does contain a term proportional to $-\varepsilon\widetilde{D}^{-1}$ (Janssen 2018) and could therefore be responsible for the observed negative scaling with ε in shallow waters where $\widetilde{D}$ is small. Another possible explanation is wave breaking, which also depends inversely on relative depth (e.g. Goda 2010). In very shallow waters, more sea states have a steepness close to the breaking threshold, which removes taller waves that tend to have a higher steepness than average.

5. Limitations

This analysis has some notable limitations due to the fact that we are only using wave buoy observations:

1. We do not have observations for all causal pathways. This includes wind-wave interactions, currents, and local topography (see Fig. 1). Fortunately, all 3 of these unobserved causal pathways relate to localized phenomena that are unlikely to play a major roll in bulk analysis. Nevertheless, this implies that local rogue wave probabilities could be dramatically different in regimes where these causal pathways play a major role, e.g. over sloping topography (Trulsen et al. 2012) or in strong currents (Ying et al. 2011).
2. We only have 1-dimensional (time series) data. Because of that we cannot capture conditions where parameters are "imported" from elsewhere, e.g., a soliton generated in different conditions travelling into the observation area. While we expect most sea state parameters to be sufficiently stable in space for this to play a minor role, this is one mechanism through which we might underestimate the importance of nonlinear free waves.
3. Sensor bias. Systematic sensor bias is common in buoys, e.g. in the form of linearization of the sea state (McAllister and van den Bremer 2019) and can also lead to spurious causal relationships. For example, the fact that the best performing model uses both crest-trough correlation and the directionality index could also be because the directionality index can be used to correct noisy measurements of crest-trough correlation for this particular sensor.

   If the only goal is maximized predictive performance this adaptation to sensor characteristics is actually a good thing, since several noisy quantities can be synthesized into more robust ones. On the other hand, this may obscure the true causal structure and generalize poorly to other sensors.

All of these factors can potentially reduce the capabilities of our model to detect relevant causal pathways, and may lead to an underestimation of the true rogue wave risk, even though we are confident that we accurately capture the leading-order dynamics of rogue wave generation.

Since our analysis is agnostic to the data source at hand (all that is needed is a mapping from parameters to observed wave heights), it can easily be repeated on different data sources as they become available to validate our findings.

6. Next steps

a. Comparisons to theory

This study emphasizes the importance of bandwidth effects to predict rogue waves, while also including higher-order nonlinear effects that are governed by steepness and depend critically on relative water depth and directionality index. These parameters are frequently suggested by theory, but combined in a non-standard way to achieve greatly improved predictive performance and causal consistency. A logical next step would be to compare our predictions to existing theory, such as Fedele (2015); Janssen (2018), to study which parts of the prediction surface can be understood through it (and in turn quantify their predictive quality) and which parts are truly novel. We hope that this may ultimately lead to an improved theoretical understanding of real-world rogue waves.

b. An improved rogue wave forecast

Another, perhaps even more obvious next step is the comparison of ECMWF's operational freak wave forecast (ECMWF 2021). This operational forecast focuses on envelope wave heights and does not include crest-trough correlation or any other bandwidth parameter for linear wave interactions. Therefore we are confident that large improvements are within reach in terms of predicting crest-to-trough rogue waves.

However, the fact that our model is trained on observations may be problematic, since the model also learns to correct for systematic sensor bias (see Section 5). Also, having access to in-situ data on the sea state might lead to improved performance by itself. To correct for these effects, our model should be re-trained on *forecast* sea state parameters $\mathbf{x}$ and observed labels y. This way, both the current operational forecast and the empirical model would use the same input data, which allows for an apples-to-apples comparison.

c. Predicting super-rogue waves

Observed wave height distributions often show a flattening of the wave height distribution towards the extreme tail (e.g. Gemmrich and Garrett 2011; Casas-Prat and Holthuijsen 2010), which Adcock and Taylor (2014) call a Type 3 distribution. Therefore, we expect rogue wave probabilities to be more pronounced for even more extreme waves (e.g. with $H/H_s > 2.4$).

The lack of sufficient direct observations in these regimes calls for a different strategy. One approach could be to transform this classification problem (rogue wave or not) into a regression, where the predicted variables are the parameters of a candidate wave height probability distribution (e.g. shape and scale parameters of a Weibull distribution). Then, a similar analysis as in this study could be conducted for these parameters, which may reveal the main mechanisms influencing the risk for truly exceptional waves, and whether this flattening can be confirmed in our dataset.

Acknowledgments. Dion Häfner received funding from the Danish Hydrocarbon Research and Technology Centre (DHRTC).

Raw data were furnished by the Coastal Data Information Program (CDIP), Integrative Oceanography Division, operated by the Scripps Institution of Oceanography, under the sponsorship of the U.S. Army Corps of Engineers and the California Department of Parks and Recreation.

Computational resources were provided by DC^3, the Danish Center for Climate Computing.

The authors thank Jonas Peters for his input on the initial design of this study.

This publication was made possible by the following open-source software stack: JAX (Bradbury et al. 2018), flax (Heek et al. 2020), optax (Hessel et al. 2020), scikit-learn (Pedregosa et al. 2011), PyALE (Jomar 2020), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), matplotlib (Hunter 2007), Seaborn (Waskom 2021), pandas (Wes McKinney 2010), Jupyter (Kluyver et al. 2016).

Data availability statement. The preprocessed and aggregated version of the FOWD CDIP data used in this study is available for download at `https://erda.ku.dk/archives/7072a6eb3d181149deb56b7d8739805e/published-archive.html`.

APPENDIX A

Sea state parameters

Here, we give the definition of the sea state parameters central to this study. For a more thorough description of how parameters are computed from buoy displacement time series see Häfner et al. (2021a).

All parameters can be derived from the non-directional wave spectrum $\mathcal{S}(f)$ (wave spectral density depending on frequency f), with the exception of directional spread σ_θ, which is estimated from the horizontal motion of the buoy and taken from the raw CDIP data.

Many parameters are computed from moments of the wave spectrum, where the n-th moment m_n is defined as

$$m_n = \int_0^\infty f^n \mathcal{S}(f)\,\mathrm{d}f \tag{A1}$$

The expressions used for the relevant sea state parameters are:

Significant wave height:

$$H_s = 4\sqrt{m_0} \tag{A2}$$

Spectral bandwidth (narrowness):

$$\nu_f = \sqrt{m_2 m_0 / m_1^2 - 1} \tag{A3}$$

Peak wavenumber k_p, computed via the peak period (as in Young 1995):

$$\overline{T}_p = \frac{\int \mathcal{S}(f)^4\,\mathrm{d}f}{\int f\mathcal{S}(f)^4\,\mathrm{d}f} \tag{A4}$$

The peak frequency $f_p = 1/\overline{T}_p$ then leads to the peak wavenumber through the dispersion relation for linear waves in intermediate water:

$$f^2 = \frac{gk}{(2\pi)^2}\tanh(kD) \tag{A5}$$

with gravitational acceleration g and water depth D. An approximate inverse is given in Fenton (1988).

Relative depth:

$$\widetilde{D} = \frac{D}{\lambda} = \frac{1}{2\pi} k_p D \tag{A6}$$

with wave length λ.

Peak steepness:

$$\varepsilon = H_s k_p \tag{A7}$$

Benjamin-Feir index:

$$\mathrm{BFI} = \frac{\varepsilon \nu}{\sigma_f} \sqrt{\max\{\beta/\alpha, 0\}} \tag{A8}$$

where σ_f is spectral bandwidth estimated through peakedness and ν, α, β are coefficients depending only on $\widetilde{D}$ (full expression given in Serio et al. 2005).

Directionality index:

$$R = \frac{\sigma_\theta^2}{2\nu_f^2} \tag{A9}$$

Crest-trough correlation:

$$r = \frac{1}{m_0} \sqrt{\rho^2 + \lambda^2} \tag{A10}$$

$$\rho = \int_0^\infty \mathcal{S}(\omega) \cos\left(\omega \frac{\overline{T}}{2}\right) \mathrm{d}\omega \tag{A11}$$

$$\lambda = \int_0^\infty \mathcal{S}(\omega) \sin\left(\omega \frac{\overline{T}}{2}\right) \mathrm{d}\omega \tag{A12}$$

where ω is the angular frequency and $\overline{T} = m_0/m_1$ the spectral mean period (Tayfun and Fedele 2007).

APPENDIX B

Model implementation and hyperparameters

All performance critical model code is implemented in JAX (Bradbury et al. 2018), using neural network modules from flax (Heek et al. 2020) and optimizers from optax (Hessel et al. 2020). We run each experiment on a single Tesla P100 GPU in about 40 minutes, including SWAG sampling and re-training on every validation subset.

The hyperparameters for all experiments are shown in Table B3.

TABLE B3. Hyperparameters used in experiments.

Hyperparameters	
Optimizer	Adam
Learning rate	10^{-4}
Number of hidden layers	3
Neurons in hidden layers	(32, 16, 8)
ℓ_1 penalty λ_1	0
ℓ_2 penalty λ_2	1×10^{-5}
Number of training epochs	50
Number of SWAG epochs	50
Number of SWAG posterior samples	100
Train-test split	66% – 34%

References

Adcock, T. A. A., and P. H. Taylor, 2014: The physics of anomalous ('rogue') ocean waves. *Reports on Progress in Physics*, **77 (10)**, 105 901, doi:10.1088/0034-4885/77/10/105901.

Apley, D. W., and J. Zhu, 2019: Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv:1612.08468 [stat]*, arXiv: 1612.08468.

Behrens, J., J. Thomas, E. Terrill, and R. Jensen, 2019: CDIP: Maintaining a Robust and Reliable Ocean Observing Buoy Network. *2019 IEEE/OES Twelfth Current, Waves and Turbulence Measurement (CWTM)*, 1–5, doi:10.1109/CWTM43797.2019.8955166.

Boccotti, P., 2000: *Wave Mechanics for Ocean Engineering*. Elsevier, google-Books-ID: 1319kgDa8GUC.

Bradbury, J., and Coauthors, 2018: JAX: composable transformations of Python+NumPy programs. URL http://github.com/google/jax.

Casas-Prat, M., and L. H. Holthuijsen, 2010: Short-term statistics of waves observed in deep water. *Journal of Geophysical Research: Oceans*, **115 (C9)**, doi:10.1029/2009JC005742.

Didenkulova, E., 2019: Catalogue of rogue waves occurred in the World Ocean from 2011 to 2018 reported by mass media sources. *Ocean & Coastal Management*, 105076, doi:10.1016/j.ocecoaman.2019.105076.

Didenkulova, E. G., T. G. Talipova, and E. N. Pelinovsky, 2021: Rogue Waves in the Drake Passage: Unpredictable Hazard. *Antarctic Peninsula Region of the Southern Ocean: Oceanography and Ecology*, E. G. Morozov, M. V. Flint, and V. A. Spiridonov, Eds., Advances in Polar Ecology, Springer International Publishing, Cham, 101–114, doi:10.1007/978-3-030-78927-5_7.

Dudley, J. M., G. Genty, A. Mussot, A. Chabchoub, and F. Dias, 2019: Rogue waves and analogies in optics and oceanography. *Nature Reviews Physics*, **1 (11)**, 675–689, doi:10.1038/s42254-019-0100-0, number: 11 Publisher: Nature Publishing Group.

ECMWF, 2021: Part VII: ECMWF Wave model. *IFS Documentation CY47R3*, IFS Documentation, ECMWF, URL https://www.ecmwf.int/node/20201.

Fedele, F., 2015: On the kurtosis of deep-water gravity waves. *Journal of Fluid Mechanics*, **782**, 25–36, doi:10.1017/jfm.2015.538, publisher: Cambridge University Press.

Fedele, F., J. Brennan, S. Ponce de León, J. Dudley, and F. Dias, 2016: Real world ocean rogue waves explained without the modulational instability. *Scientific Reports*, **6**, 27 715, doi:10.1038/srep27715.

Fedele, F., J. Herterich, A. Tayfun, and F. Dias, 2019: Large nearshore storm waves off the Irish coast. *Scientific Reports*, **9 (1)**, 15 406, doi:10.1038/s41598-019-51706-8, number: 1 Publisher: Nature Publishing Group.

Fedele, F., and M. A. Tayfun, 2009: On nonlinear wave groups and crest statistics. *Journal of Fluid Mechanics*, **620**, 221–239, doi:10.1017/S0022112008004424, URL https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/on-nonlinear-wave-groups-and-crest-statistics/CF946526383D92D4E12A0F950CE4FB1C, publisher: Cambridge University Press.

Fenton, J. D., 1988: The numerical solution of steady water wave problems. *Computers & Geosciences*, **14 (3)**, 357–368, doi:10.1016/0098-3004(88)90066-0.

Gemmrich, J., and C. Garrett, 2011: Dynamical and statistical explanations of observed occurrence rates of rogue waves. *Natural Hazards and Earth System Science*, **11 (5)**, 1437–1446, doi:10.5194/nhess-11-1437-2011.

Goda, Y., 2010: Reanalysis of Regular and Random Breaking Wave Statistics. *Coastal Engineering Journal*, **52 (1)**, 71–106, doi:10.1142/S0578563410002129, publisher: Taylor & Francis _eprint: https://doi.org/10.1142/S0578563410002129.

Gramstad, O., and K. Trulsen, 2007: Influence of crest and group length on the occurrence of freak waves. *Journal of Fluid Mechanics*, **582**, 463–472, doi:10.1017/S0022112007006507, publisher: Cambridge University Press.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585 (7825)**, 357–362, doi:10.1038/s41586-020-2649-2.

Heek, J., A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee, 2020: Flax: A neural network library and ecosystem for JAX. URL http://github.com/google/flax.

Hessel, M., D. Budden, F. Viola, M. Rosca, E. Sezener, and T. Hennigan, 2020: Optax: composable gradient transformation and optimisation, in JAX! URL http://github.com/deepmind/optax.

Hornik, K., 1991: Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4 (2)**, 251–257, doi:10.1016/0893-6080(91)90009-T.

Hunter, J. D., 2007: Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, **9 (3)**, 90–95, doi:10.1109/MCSE.2007.55.

Häfner, D., J. Gemmrich, and M. Jochum, 2021a: FOWD: A Free Ocean Wave Dataset for Data Mining and Machine Learning. *Journal of Atmospheric and Oceanic Technology*, **-1 (aop)**, doi:10.1175/JTECH-D-20-0185.1, publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology.

Häfner, D., J. Gemmrich, and M. Jochum, 2021b: Real-world rogue wave probabilities. *Scientific Reports*, **11 (1)**, 10 084, doi:10.1038/s41598-021-89359-1, number: 1 Publisher: Nature Publishing Group.

Janssen, P., 2018: Shallow-water version of the Freak Wave Warning System. Technical memorandum 813, ECMWF. URL https://www.ecmwf.int/en/elibrary/18063-shallow-water-version-freak-wave-warning-system.

Janssen, P. A. E. M., 2003: Nonlinear Four-Wave Interactions and Freak Waves. *Journal of Physical Oceanography*, **33 (4)**, 863–884, doi:10.1175/1520-0485(2003)33<863:NFIAFW>2.0.CO;2, publisher: American Meteorological Society.

Jomar, D., 2020: PyALE: A Python implementation of accumulated local effect plots. URL https://github.com/DanaJomar/PyALE.

Kluyver, T., and Coauthors, 2016: Jupyter notebooks - a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides, and B. Scmidt, Eds., IOS Press, Netherlands, 87–90.

Korteweg, D. J., and G. De Vries, 1895: On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **39 (240)**, 422–443.

Longuet-Higgins, M. S., 1952: On the statistical distribution of the height of sea waves. *JMR*, **11**, 245–266.

Maddox, W., T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson, 2019: A Simple Baseline for Bayesian Uncertainty in Deep Learning. *arXiv:1902.02476 [cs, stat]*, arXiv: 1902.02476.

Mallory, J. K., 1974: Abnormal Waves on the South East Coast of South Africa. *The International Hydrographic Review*, URL https://journals.lib.unb.ca/index.php/ihr/article/view/23802.

McAllister, M. L., S. Draycott, T. a. A. Adcock, P. H. Taylor, and T. S. v. d. Bremer, 2019: Laboratory recreation of the Draupner wave and the role of breaking in crossing seas. *Journal of Fluid Mechanics*, **860**, 767–786, doi:10.1017/jfm.2018.886.

McAllister, M. L., and T. S. van den Bremer, 2019: Lagrangian Measurement of Steep Directionally Spread Ocean Waves: Second-Order Motion of a Wave-Following Measurement Buoy. *Journal of Physical Oceanography*, **49 (12)**, 3087–3108, doi:10.1175/JPO-D-19-0170.1, publisher: American Meteorological Society.

Miche, M., 1944: Mouvements ondulatoires de la mer en profondeur constante ou décroissante. *Annales de Ponts et Chaussées, 1944, pp(1) 26-78, (2)270-292, (3) 369-406*, publisher: École nationale des ponts et chaussées.

Molnar, C., 2020: *Interpretable Machine Learning*. URL https://christophm.github.io/interpretable-ml-book/.

Mori, N., and P. A. E. M. Janssen, 2006: On Kurtosis and Occurrence Probability of Freak Waves. *Journal of Physical Oceanography*, **36 (7)**, 1471–1483, doi:10.1175/JPO2922.1, URL https://journals.ametsoc.org/jpo/article/36/7/1471/10720/On-Kurtosis-and-Occurrence-Probability-of-Freak, publisher: American Meteorological Society.

Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress, Madison, WI, USA, 807–814, ICML'10.

Onorato, M., A. R. Osborne, M. Serio, L. Cavaleri, C. Brandini, and C. T. Stansberg, 2006: Extreme waves, modulational instability and second order theory: wave flume experiments on irregular waves. *European Journal of Mechanics - B/Fluids*, **25 (5)**, 586–601, doi:10.1016/j.euromechflu.2006.01.002.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Peters, J., P. Bühlmann, and N. Meinshausen, 2016: Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78 (5)**, 947–1012, doi:10.1111/rssb.12167, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12167.

Peters, J., D. Janzing, and B. Schölkopf, 2017: *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series, MIT Press, Cambridge, MA, USA.

Serio, M., M. Onorato, A. R a Osborne, and P. Janssen, 2005: On the computation of the Benjamin-Feir Index. *Nuovo Cimento della Societa Italiana di Fisica C*, **28**, 893–903, doi:10.1393/ncc/i2005-10134-1.

Stansell, P., 2004: Distributions of freak wave heights measured in the North Sea. *Applied Ocean Research*, **26 (1)**, 35–48, doi:10.1016/j.apor.2004.01.004, URL http://www.sciencedirect.com/science/article/pii/S0141118704000379.

Tayfun, M. A., and F. Fedele, 2007: Wave-height distributions and nonlinear effects. *Ocean Engineering*, **34 (11)**, 1631–1649, doi:10.1016/j.oceaneng.2006.11.006.

Trulsen, K., H. Zeng, and O. Gramstad, 2012: Laboratory evidence of freak waves provoked by non-uniform bathymetry. *Physics of Fluids*, **24 (9)**, 097 101, doi:10.1063/1.4748346, publisher: American Institute of Physics.

Ursell, F., 1953: The long-wave paradox in the theory of gravity waves. *Mathematical Proceedings of the Cambridge Philosophical Society*, **49 (4)**, 685–694, doi:10.1017/S0305004100028887, URL https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/longwave-paradox-in-the-theory-of-gravity-waves/5A178FB13BD7B9C3314A49A495A860DB, publisher: Cambridge University Press.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, 261–272, doi:10.1038/s41592-019-0686-2.

Waskom, M. L., 2021: seaborn: statistical data visualization. *Journal of Open Source Software*, **6 (60)**, 3021, doi:10.21105/joss.03021.

Wes McKinney, 2010: Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt, and Jarrod Millman, Eds., 56 – 61, doi:10.25080/Majora-92bf1922-00a.

Xiao, W., Y. Liu, G. Wu, and D. K. P. Yue, 2013: Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution. *Journal of Fluid Mechanics*, **720**, 357–392, doi:10.1017/jfm.2013.37, publisher: Cambridge University Press.

Ying, L. H., Z. Zhuang, E. J. Heller, and L. Kaplan, 2011: Linear and nonlinear rogue wave statistics in the presence of random currents. *Nonlinearity*, **24 (11)**, R67–R87, doi:10.1088/0951-7715/24/11/R01, URL https://doi.org/10.1088/0951-7715/24/11/r01, publisher: IOP Publishing.

Young, I. R., 1995: The determination of confidence limits associated with estimates of the spectral peak frequency. *Ocean Engineering*, **22 (7)**, 669–686, doi:10.1016/0029-8018(95)00002-3.

CHAPTER CONTENTS

Extrapolation

3

3.1 NEXT STEPS IN EXTREME WAVE RESEARCH

The extreme wave research community is divided. Are rogue waves rare realizations in typical sea states or typical realizations in rare sea states? To make things worse, the term "rogue wave" is also used for extreme waves in other media such as optical fibers (see e. g. Dudley et al., 2019), with some overlapping creation mechanisms (like the modulational instability), but in general completely different characteristics. Discussions about rogue waves are therefore full of confusion and misunderstandings.

Figure 3.1: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"a scientist looking at the ocean swell during sunset"*.

Our work has started to answer some of these questions: It looks like the vast majority of rogue waves in our data are rare realizations of typical sea states. This does not necessarily mean that nonlinear phenomena like solitons or breathers do not exist, just that they seem to not play a major role in "everyday" rogue wave generation. So what are the implications for the field? The following sections outline some — in my opinion — much needed further work.

3.1.1 *A More Meaningful Rogue Wave Definition*

When taking the rogue wave definition (1.1) at face value, *every* wave larger than the chosen threshold is a rogue wave, regardless of the underlying creation mechanism (and regardless of the absolute wave height). Yet, the rogue wave definition is only meaningful if these waves do not behave like the rest of the wave population[1], so many authors choose to focus on examples where this is the case (whether or not they are represented in the real ocean).

We have shown that when we apply the rogue wave definition strictly, real-world rogue waves are well explained by bandwidth and directionality effects, plus some minor correction from weakly nonlinear dynamics. This may inform further research and lead to an improved understanding of the tail of the wave height distribution, which is valuable in itself, but it does not make a direct statement about the huge *freaks* (like the Draupner wave) that people usually have in mind when talking about rogue waves[2].

A good definition is one that is helpful in the context where it is applied. If the goal is to study *dangerous waves*, then other criteria that make waves dangerous should enter the definition, such as absolute height, steepness, or unexpectedness (Gemmrich and Garrett, 2008). If the goal is to study *huge*

[1] As is implied by "rogue" or "freak", as opposed to "unlikely wave".

[2] Only if large rogue waves share the same generation mechanisms as small rogue waves, which is not obvious in the presence of nonlinearities.

waves or *big storm waves* or *highly nonlinear waves*, the definition should reflect that. A more sensible definition could help to unify the different view points on extreme waves, and bring engineers and physicists closer together.

3.1.2 Firmer Anchoring in Observations

In wave research, new theories are typically validated against simulations and wave plume experiments. Now that we have the tools for it (via datasets like FOWD and the analysis we developed), in-situ observations should play a bigger role in this process. This could help to shift the focus from conditional probabilities (if conditions X are met, Y will happen) to joint probabilities (conditions X might never be met due to physical constraints, so Y is impossible; see also Mendes, Scotti, and Stansell, 2021).

There is much left to discover in FOWD and similar datasets, and we have only scratched the surface in terms of available data from many more providers. The field could profit tremendously from more rigorous empirical work, e. g. by examining the role of currents on rogue wave formation.

3.1.3 An Improved Operational Forecast

Finally, an improved operational forecast for maximum expected wave heights is now within reach, for example one that includes the effects of crest-trough correlation and directionality on rogue wave probabilities. This should lead to clear improvements in terms of forecasting skill for typical rogue waves (something we plan to address in follow-up work).

Many accidents in the ocean involve rogue waves in sea states similar to the ones we have studied, and an improved forecast will allow for better planning of shipping routes to avoid the most dangerous conditions, or exploit conditions that are known to be safe.

3.2 NEXT STEPS IN MACHINE LEARNING FOR SCIENCE

In the study of physical systems, the combination of computation, data, machine learning, and causal reasoning is an extremely powerful one (Lavin et al., 2021; Reichstein et al., 2019). Still, there is a long way to go before this approach becomes mainstream. The following sections outline some concrete steps towards this.

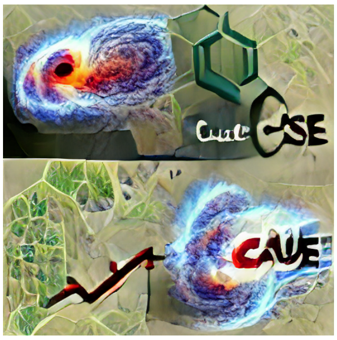


Figure 3.2: AI art generated by VQ-GAN + CLIP (Esser, Rombach, and Ommer, 2021; Radford et al., 2021). Prompt: *"cause and effect"*.

3.2.1 *More Methods Tailored to Physical Data*

Machine learning algorithms are usually not developed and evaluated with physical data in mind. One example are classification problems on tabular data (like in this study), where tree-based methods like boosted trees (e. g. via XGBoost, Chen and Guestrin, 2016) are typically found to perform best (Shwartz-Ziv and Armon, 2022). However, most machine learning applications on tabular data are on *human-centric* problems that have very different characteristics than physical problems, such as a vastly more complicated causal structure and discontinuous behavior.

As a consequence, methods other than the industry standard may be most appropriate on physical data. To address this, physical datasets should be accounted for during model evaluation by machine learning researchers to lead to methods that exploit *"The unreasonable effectiveness of mathematics in the natural sciences"* (Wigner, 1960), while making them more approachable for domain scientists.

3.2.2 *Tools for Uncertainty Estimation*

Reasoning under uncertainty is a staple in science, where answers that may be "good enough" in other domains (such as recommender systems) don't qualify, and where "I know that I don't know" is valuable information.

Most approaches to machine learning with uncertainties have long been prohibitively costly for large datasets like ours (i. e., have non-linear scaling with the dataset size), such as Gaussian process regression or Bayesian methods based on Monte Carlo sampling. Modern techniques like sparse Gaussian processes (Leibfried et al., 2020), variational inference (Blei, Kucukelbir, and McAuliffe, 2017), deep ensembles (Lakshminarayanan, Pritzel, and Blundell, 2017), and stochastic weight averaging (Maddox et al., 2019; Wilson and Izmailov, 2020) alleviate this, and represent a promising way towards integrating uncertainty information and machine learning. Still, more development is necessary for the widespread adoption of those or similar methods.

3.2.3 *Off-the-Shelf Causal Inference*

At its core, every scientific theory aims to describe a causal connection, so methods to identify and validate causality must be central to data-driven science. But causality in arbitrary physical systems is often intractable, and the borders between association and causation are blurred in the presence of deterministic chaos and multi-scale interactions.

This calls for a formalization of causality in these systems on the one hand (e. g. as in Peters, Bauer, and Pfister, 2020), and methods for causal inference that are tailored to physical systems on the other hand (in a similar way as it has happened in medicine or econometrics). This may include techniques like invariant causal prediction (Peters, Bühlmann, and Meinshausen, 2016) in connection with symbolic regression (Cranmer et al., 2020a), which enables the identification of causally consistent machine learning models that can then be distilled into simple mathematical expressions for fully data-driven discovery of scientific theories.

References

Adcock, Thomas A. A. and Paul H. Taylor (2014). "The physics of anomalous ('rogue') ocean waves." en. In: *Reports on Progress in Physics* 77.10, p. 105901. ISSN: 0034-4885. DOI: 10.1088/0034-4885/77/10/105901 (Cited on p. 11).

Adcock, Thomas A.A. and Paul H. Taylor (Mar. 2011). "Energy Input Amplifies Nonlinear Dynamics of Deep Water Wave Groups." In: *International Journal of Offshore and Polar Engineering* 21.01. ISSN: 1053-5381 (Cited on p. 11).

Alber, I. E. and Keith Stewartson (Nov. 1978). "The effects of randomness on the stability of two-dimensional surface wavetrains." In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 363.1715. Publisher: Royal Society, pp. 525–546. DOI: 10.1098/rspa.1978.0181 (Cited on p. 10).

Bakarji, Joseph et al. (Jan. 2022). "Discovering Governing Equations from Partial Measurements with Deep Delay Autoencoders." In: *arXiv:2201.05136 [cs, math]*. arXiv: 2201.05136 (Cited on p. 13).

Bar-Sinai, Yohai et al. (Aug. 2018). "Data-driven discretization: machine learning for coarse graining of partial differential equations." In: *arXiv:1808.04930 [cond-mat, physics:physics]*. arXiv: 1808.04930 (Cited on p. 12).

Behrens, James et al. (Mar. 2019). "CDIP: Maintaining a Robust and Reliable Ocean Observing Buoy Network." In: *2019 IEEE/OES Twelfth Current, Waves and Turbulence Measurement (CWTM)*, pp. 1–5. DOI: 10.1109/CWTM43797.2019.8955166 (Cited on p. 17).

Benjamin, T. Brooke and J. E. Feir (Feb. 1967). "The disintegration of wave trains on deep water Part 1. Theory." en. In: *Journal of Fluid Mechanics* 27.3. Publisher: Cambridge University Press, pp. 417–430. ISSN: 1469-7645, 0022-1120. DOI: 10.1017/S002211206700045X (Cited on p. 10).

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (Apr. 2017). "Variational Inference: A Review for Statisticians." In: *Journal of the American Statistical Association* 112.518. arXiv: 1601.00670, pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1285773 (Cited on p. 68).

Boccotti, P. (1989). *On Mechanics of Irregular Gravity Waves*. 3: Atti della Accademia ... / Memorie della. Accademia Nazionale dei Lincei (Cited on p. 7).

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners." In: *CoRR* abs/2005.14165. arXiv: 2005.14165 (Cited on p. 1).

Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz (Apr. 2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems." en. In: *Proceedings of the National Academy of Sciences* 113.15. Publisher: National Academy of Sciences Section: Physical Sciences, pp. 3932–3937. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1517384113 (Cited on p. 13).

Casas-Prat, Mercè and Leo H. Holthuijsen (2010). "Short-term statistics of waves observed in deep water." en. In: *Journal of Geophysical Research: Oceans* 115.C9. ISSN: 2156-2202. DOI: 10.1029/2009JC005742 (Cited on p. 17).

Chabchoub, A., N. P. Hoffmann, and N. Akhmediev (May 2011). "Rogue Wave Observation in a Water Wave Tank." In: *Physical Review Letters* 106.20, p. 204502. DOI: 10.1103/PhysRevLett.106.204502 (Cited on p. 9).

Champion, Kathleen et al. (Nov. 2019). "Data-driven discovery of coordinates and governing equations." en. In: *Proceedings of the National Academy of Sciences* 116.45. Publisher: National Academy of Sciences Section: Physical Sciences, pp. 22445–22451. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1906995116 (Cited on p. 13).

Chen, Ricky T. Q. et al. (Dec. 2019). "Neural Ordinary Differential Equations." In: *arXiv:1806.07366 [cs, stat]*. arXiv: 1806.07366 (Cited on p. 13).

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785 (Cited on p. 68).

Christou, Marios and Kevin Ewans (May 2014). "Field Measurements of Rogue Water Waves." In: *Journal of Physical Oceanography* 44.9, pp. 2317–2335. ISSN: 0022-3670. DOI: 10.1175/JPO-D-13-0199.1 (Cited on p. 17).

Clamond, Didier and John Grue (Jan. 2002). "Interaction between envelope solitons as a model for freak wave formations. Part I: Long time interaction." en. In: *Comptes Rendus Mécanique* 330.8, pp. 575–580. ISSN: 1631-0721. DOI: 10.1016/S1631-0721(02)01496-1 (Cited on p. 9).

Cranmer, Miles et al. (Nov. 2020a). "Discovering Symbolic Models from Deep Learning with Inductive Biases." In: *arXiv:2006.11287 [astro-ph, physics:physics, stat]*. arXiv: 2006.11287 (Cited on pp. 13, 69).

Cranmer, Miles et al. (July 2020b). "Lagrangian Neural Networks." en. In: *arXiv:2003.04630 [physics, stat]*. arXiv: 2003.04630 (Cited on p. 12).

Cranmer, Miles et al. (Oct. 2021). "A Bayesian neural network predicts the dissolution of compact planetary systems." en. In: *Proceedings of the National Academy of Sciences* 118.40. Publisher: National Academy of Sciences Section: Physical Sciences. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2026053118 (Cited on p. 12).

Davey, A. and Keith Stewartson (June 1974). "On three-dimensional packets of surface waves." In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 338.1613. Publisher: Royal Society, pp. 101–110. DOI: 10.1098/rspa.1974.0076 (Cited on p. 10).

Dean, Robert G. and Robert A. Dalrymple (Jan. 1991). *Water Wave Mechanics For Engineers And Scientists.* en. Google-Books-ID: 1SM8DQAAQBAJ. World Scientific Publishing Company. ISBN: 978-981-4365-69-7 (Cited on pp. 8, 9).

Dematteis, Giovanni et al. (Dec. 2019). "Experimental Evidence of Hydrodynamic Instantons: The Universal Route to Rogue Waves." In: *Physical Review X* 9.4, p. 041057. DOI: 10.1103/PhysRevX.9.041057 (Cited on p. 10).

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *CoRR* abs/1810.04805. arXiv: 1810.04805 (Cited on p. 1).

Didenkulova, Ekaterina (Dec. 2019). "Catalogue of rogue waves occurred in the World Ocean from 2011 to 2018 reported by mass media sources." en. In: *Ocean & Coastal Management*, p. 105076. ISSN: 0964-5691. DOI: 10.1016/j.ocecoaman.2019.105076 (Cited on p. 4).

Didenkulova, Ekaterina G., Tatiana G. Talipova, and Efim N. Pelinovsky (2021). "Rogue Waves in the Drake Passage: Unpredictable Hazard." en. In: *Antarctic Peninsula Region of the Southern Ocean: Oceanography and Ecology.* Ed. by Eugene G. Morozov, Mikhail V. Flint, and Vassily A. Spiridonov. Advances in Polar Ecology. Cham: Springer International Publishing, pp. 101–114. ISBN: 978-3-030-78927-5. DOI: 10.1007/978-3-030-78927-5_7 (Cited on p. 11).

Dudley, John M. et al. (Nov. 2019). "Rogue waves and analogies in optics and oceanography." en. In: *Nature Reviews Physics* 1.11. Number: 11 Publisher: Nature Publishing Group, pp. 675–689. ISSN: 2522-5820. DOI: 10.1038/s42254-019-0100-0 (Cited on pp. 11, 66).

Dysthe, K. B. and Michael Selwyn Longuet-Higgins (Dec. 1979). "Note on a modification to the nonlinear Schrödinger equation for application to deep water waves." In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 369.1736. Publisher: Royal Society, pp. 105–114. DOI: 10.1098/rspa.1979.0154 (Cited on p. 10).

Esser, Patrick, Robin Rombach, and Bjorn Ommer (2021). "Taming Transformers for High-Resolution Image Synthesis." en. In: pp. 12873–12883 (Cited on pp. 1, 4, 12, 15, 17, 38, 51, 66, 68).

Fedele, Francesco (Nov. 2015). "On the kurtosis of deep-water gravity waves." en. In: *Journal of Fluid Mechanics* 782. Publisher: Cambridge University Press, pp. 25–36. ISSN: 0022-1120, 1469-7645. DOI: 10.1017/jfm.2015.538 (Cited on p. 10).

Fedele, Francesco et al. (June 2016). “Real world ocean rogue waves explained without the modulational instability.” en. In: *Scientific Reports* 6, p. 27715. ISSN: 2045-2322. DOI: 10.1038/srep27715 (Cited on p. 8).

Fedele, Francesco et al. (Oct. 2019). “Large nearshore storm waves off the Irish coast.” en. In: *Scientific Reports* 9.1. Number: 1 Publisher: Nature Publishing Group, p. 15406. ISSN: 2045-2322. DOI: 10.1038/s41598-019-51706-8 (Cited on p. 8).

Gemmrich, J. and C. Garrett (May 2011). “Dynamical and statistical explanations of observed occurrence rates of rogue waves.” en. In: *Natural Hazards and Earth System Science* 11.5, pp. 1437–1446. ISSN: 1684-9981. DOI: 10.5194/nhess-11-1437-2011 (Cited on p. 8).

Gemmrich, Johannes and Chris Garrett (Oct. 2008). “Unexpected Waves.” EN. In: *Journal of Physical Oceanography* 38.10. Publisher: American Meteorological Society Section: Journal of Physical Oceanography, pp. 2330–2336. ISSN: 0022-3670, 1520-0485. DOI: 10.1175/2008JPO3960.1 (Cited on p. 66).

Gramstad, Odin and Karsten Trulsen (July 2007). “Influence of crest and group length on the occurrence of freak waves.” en. In: *Journal of Fluid Mechanics* 582. Publisher: Cambridge University Press, pp. 463–472. ISSN: 1469-7645, 0022-1120. DOI: 10.1017/S0022112007006507 (Cited on p. 10).

Hannart, A. et al. (Jan. 2016). “Causal Counterfactual Theory for the Attribution of Weather and Climate-Related Events.” EN. In: *Bulletin of the American Meteorological Society* 97.1. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, pp. 99–110. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-14-00034.1 (Cited on p. 12).

Hasselmann, K. (1966). “Feynman diagrams and interaction rules of wave-wave scattering processes.” en. In: *Reviews of Geophysics* 4.1, pp. 1–32. ISSN: 1944-9208. DOI: 10.1029/RG004i001p00001 (Cited on p. 11).

Haver, Sverre (2004). “A possible freak wave event measured at the Draupner Jacket January 1 1995.” In: *Rogue waves*. Vol. 2004, pp. 1–8 (Cited on p. 5).

Hayer, Sverre and Odd Jan Andersen (2000). “Freak waves: rare realizations of a typical population or typical realizations of a rare population?” In: *The Tenth International Offshore and Polar Engineering Conference*. OnePetro (Cited on p. 5).

Holthuijsen, Leo H. (Feb. 2010). *Waves in Oceanic and Coastal Waters*. en. Cambridge University Press. ISBN: 978-1-139-46252-5 (Cited on pp. 4–7, 9).

Häfner, Dion, Johannes Gemmrich, and Markus Jochum (May 2021a). “FOWD: A Free Ocean Wave Dataset for Data Mining and Machine Learning.” EN. In: *Journal of Atmospheric and Oceanic Technology* -1.aop. Publisher: American Meteorological Society Section: Journal of Atmospheric and

Oceanic Technology. ISSN: 0739-0572, 1520-0426. DOI: 10.1175/JTECH-D-20-0185.1 (Cited on p. 17).

Häfner, Dion, Johannes Gemmrich, and Markus Jochum (May 2021b). "Real-world rogue wave probabilities." en. In: *Scientific Reports* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 10084. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89359-1 (Cited on p. 38).

IPCC (2007). "Climate change 2007: the physical science basis." In: *Agenda* 6.07, p. 333 (Cited on p. 13).

Janssen, Peter A. E. M. (Apr. 2003). "Nonlinear Four-Wave Interactions and Freak Waves." en. In: *Journal of Physical Oceanography* 33.4. Publisher: American Meteorological Society, pp. 863–884. ISSN: 0022-3670. DOI: 10.1175/1520-0485(2003)33<863:NFIAFW>2.0.CO;2 (Cited on p. 11).

Johnson, R. S. (Oct. 1997). *A Modern Introduction to the Mathematical Theory of Water Waves*. en. Google-Books-ID: oQ2Cw4Rnve8C. Cambridge University Press. ISBN: 978-0-521-59832-3 (Cited on p. 10).

Kharif, C. et al. (Feb. 2001). "Focusing of nonlinear wave groups in deep water." en. In: *Journal of Experimental and Theoretical Physics Letters* 73.4, pp. 170–175. ISSN: 1090-6487. DOI: 10.1134/1.1368708 (Cited on p. 10).

Kharif, Christian and Efim Pelinovsky (Nov. 2003). "Physical mechanisms of the rogue wave phenomenon." In: *European Journal of Mechanics - B/Fluids* 22.6, pp. 603–634. ISSN: 0997-7546. DOI: 10.1016/j.euromechflu.2003.09.002 (Cited on pp. 9, 10).

Kidger, Patrick (Feb. 2022). "On Neural Differential Equations." In: *arXiv:2202.02435 [cs, math, stat]*. arXiv: 2202.02435 (Cited on p. 13).

Kochkov, Dmitrii et al. (May 2021). "Machine learning–accelerated computational fluid dynamics." en. In: *Proceedings of the National Academy of Sciences* 118.21. Publisher: National Academy of Sciences Section: Physical Sciences. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2101784118 (Cited on p. 12).

Korteweg, Diederik Johannes and Gustav De Vries (1895). "On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves." In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 39.240, pp. 422–443 (Cited on p. 9).

Kouki, Pigi et al. (2020). "From the Lab to Production: A Case Study of Session-Based Recommendations in the Home-Improvement Domain." In: *Fourteenth ACM Conference on Recommender Systems*. New York, NY, USA: Association for Computing Machinery, 140–149. ISBN: 9781450375832 (Cited on p. 1).

Kretschmer, Marlene et al. (June 2016). "Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation." EN.

In: *Journal of Climate* 29.11. Publisher: American Meteorological Society Section: Journal of Climate, pp. 4069–4081. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-15-0654.1 (Cited on p. 12).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2017). “ImageNet Classification with Deep Convolutional Neural Networks.” In: *Commun. ACM* 60.6, 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386 (Cited on p. 1).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (Nov. 2017). “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.” In: *arXiv:1612.01474 [cs, stat]*. arXiv: 1612.01474 (Cited on p. 68).

Lavin, Alexander et al. (Dec. 2021). “Simulation Intelligence: Towards a New Generation of Scientific Methods.” en. In: *arXiv:2112.03235 [cs]*. arXiv: 2112.03235 (Cited on p. 68).

Le Mehaute, Bernard (Dec. 1969). *An Introduction to Hydrodynamics and Water Waves.* en. Springer Science & Business Media. ISBN: 978-3-642-85567-2 (Cited on p. 9).

Leibfried, Felix et al. (2020). “A Tutorial on Sparse Gaussian Processes and Variational Inference.” In: DOI: 10.48550/ARXIV.2012.13962 (Cited on p. 68).

Lemos, Pablo et al. (Feb. 2022). “Rediscovering orbital mechanics with machine learning.” In: *arXiv:2202.02306 [astro-ph]*. arXiv: 2202.02306 (Cited on p. 13).

Li, Li et al. (Sept. 2020). “Kohn-Sham equations as regularizer: building prior knowledge into machine-learned physics.” In: *arXiv:2009.08551 [physics]*. arXiv: 2009.08551 (Cited on p. 12).

Long, Zichao, Yiping Lu, and Bin Dong (Dec. 2019). “PDE-Net 2.0: Learning PDEs from Data with A Numeric-Symbolic Hybrid Deep Network.” In: *Journal of Computational Physics* 399. arXiv: 1812.04426, p. 108925. ISSN: 00219991. DOI: 10.1016/j.jcp.2019.108925 (Cited on p. 13).

Longuet-Higgins, Michael S (1952). “On the statistical distribution of the height of sea waves.” In: *JMR* 11, pp. 245–266 (Cited on pp. 5, 6).

Ma, Yifei et al. (2020). “Temporal-contextual recommendation in real-time.” In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2291–2299 (Cited on p. 1).

Maddox, Wesley et al. (Dec. 2019). “A Simple Baseline for Bayesian Uncertainty in Deep Learning.” In: *arXiv:1902.02476 [cs, stat]*. arXiv: 1902.02476 (Cited on p. 68).

Mallory, J. K. (1974). “Abnormal Waves on the South East Coast of South Africa.” en. In: *The International Hydrographic Review*. ISSN: 0020-6946 (Cited on p. 11).

Marcus, Gary (Mar. 2022). *Deep Learning Is Hitting a Wall* (Cited on p. 1).

McAllister, M. L. et al. (Feb. 2019). “Laboratory recreation of the Draupner wave and the role of breaking in crossing seas.” en. In: *Journal of Fluid Mechanics* 860, pp. 767–786. ISSN: 0022-1120, 1469-7645. DOI: 10.1017/jfm.2018.886 (Cited on p. 11).

Mendes, S., A. Scotti, and P. Stansell (Mar. 2021). “On the physical constraints for the exceeding probability of deep water rogue waves.” en. In: *Applied Ocean Research* 108, p. 102402. ISSN: 0141-1187. DOI: 10.1016/j.apor.2020.102402 (Cited on p. 67).

Molnar, Christoph (2020). *Interpretable Machine Learning* (Cited on p. 13).

Naess, Arvid (Jan. 1985). “On the distribution of crest to trough wave heights.” en. In: *Ocean Engineering* 12.3, pp. 221–234. ISSN: 0029-8018. DOI: 10.1016/0029-8018(85)90014-9 (Cited on p. 7).

Ochi, Michel K. and E. Nadine Hubble (1976). “Six-Parameter Wave Spectra.” In: *Coastal Engineering 1976*. Proceedings, pp. 301–328. ISSN: 9780872620834. DOI: 10.1061/9780872620834.018 (Cited on p. 6).

Onorato, M. et al. (Sept. 2006). “Extreme waves, modulational instability and second order theory: wave flume experiments on irregular waves.” en. In: *European Journal of Mechanics - B/Fluids*. Rogue waves 25.5, pp. 586–601. ISSN: 0997-7546. DOI: 10.1016/j.euromechflu.2006.01.002 (Cited on p. 10).

Onorato, Miguel and Davide Proment (Oct. 2012). “Approximate rogue wave solutions of the forced and damped nonlinear Schrödinger equation for water waves.” en. In: *Physics Letters A* 376.45, pp. 3057–3059. ISSN: 0375-9601. DOI: 10.1016/j.physleta.2012.05.063 (Cited on p. 10).

Onorato, Miguel, Davide Proment, and Alessandro Toffoli (Oct. 2011). “Triggering Rogue Waves in Opposing Currents.” In: *Physical Review Letters* 107.18. Publisher: American Physical Society, p. 184502. DOI: 10.1103/PhysRevLett.107.184502 (Cited on p. 11).

Peregrine, D. H. (July 1983). “Water waves, nonlinear Schrödinger equations and their solutions.” en. In: *The ANZIAM Journal* 25.1, pp. 16–43. ISSN: 1839-4078, 0334-2700. DOI: 10.1017/S0334270000003891 (Cited on p. 10).

Pestourie, Raphaël et al. (Nov. 2021). “Physics-enhanced deep surrogates for PDEs.” en. In: *arXiv:2111.05841 [physics]*. arXiv: 2111.05841 (Cited on p. 12).

Peters, Jonas, Stefan Bauer, and Niklas Pfister (Jan. 2020). “Causal models for dynamical systems.” In: *arXiv:2001.06208 [math, stat]*. arXiv: 2001.06208 (Cited on p. 69).

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2016). “Causal inference by using invariant prediction: identification and confidence intervals.” en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 947–1012. ISSN: 1467-9868. DOI: 10.1111/rssb.12167 (Cited on pp. 51, 69).

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (Nov. 2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. en. Ed. by Francis Bach. Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press. ISBN: 978-0-262-03731-0 (Cited on pp. 12, 51).

Rackauckas, Christopher et al. (Nov. 2021). "Universal Differential Equations for Scientific Machine Learning." en. In: *arXiv:2001.04385 [cs, math, q-bio, stat]*. arXiv: 2001.04385 (Cited on p. 13).

Radford, Alec et al. (July 2021). "Learning Transferable Visual Models From Natural Language Supervision." en. In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, pp. 8748–8763 (Cited on pp. 1, 4, 12, 15, 17, 38, 51, 66, 68).

Reichstein, Markus et al. (Feb. 2019). "Deep learning and process understanding for data-driven Earth system science." En. In: *Nature* 566.7743, p. 195. ISSN: 1476-4687. DOI: 10.1038/s41586-019-0912-1 (Cited on pp. 12, 68).

Reinbold, Patrick A. K. et al. (May 2021). "Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression." en. In: *Nature Communications* 12.1. Number: 1 Publisher: Nature Publishing Group, p. 3219. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23479-0 (Cited on p. 13).

Runge, Jakob et al. (June 2019). "Inferring causation from time series in Earth system sciences." en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 2553. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10105-3 (Cited on pp. 1, 12).

Schmidt, Michael and Hod Lipson (Apr. 2009). "Distilling Free-Form Natural Laws from Experimental Data." In: *Science* 324.5923. Publisher: American Association for the Advancement of Science, pp. 81–85. DOI: 10.1126/science.1165893 (Cited on p. 13).

Serio, Marina et al. (Nov. 2005). "On the computation of the Benjamin-Feir Index." In: *Nuovo Cimento della Societa Italiana di Fisica C* 28, pp. 893–903. DOI: 10.1393/ncc/i2005-10134-1 (Cited on p. 10).

Shukla, P. K. et al. (Aug. 2006). "Instability and Evolution of Nonlinearly Interacting Water Waves." In: *Physical Review Letters* 97.9. arXiv: nlin/0608012. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.97.094501 (Cited on p. 10).

Shwartz-Ziv, Ravid and Amitai Armon (May 2022). "Tabular data: Deep learning is not all you need." en. In: *Information Fusion* 81, pp. 84–90. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.11.011 (Cited on p. 68).

Silver, David et al. (Jan. 2016). "Mastering the game of Go with deep neural networks and tree search." en. In: *Nature* 529.7587. Number: 7587 Publisher: Nature Publishing Group, pp. 484–489. ISSN: 1476-4687. DOI: 10.1038/nature16961 (Cited on p. 1).

Silver, David et al. (Dec. 2017). "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." en. In: *arXiv:1712.01815 [cs]*. arXiv: 1712.01815 (Cited on p. 1).

Slunyaev, Alexey, Ira Didenkulova, and Efim Pelinovsky (Nov. 2011). "Rogue waters." In: *Contemporary Physics* 52.6, pp. 571–590. ISSN: 0010-7514. DOI: 10.1080/00107514.2011.613256 (Cited on pp. 10, 11).

Succi, Sauro and Peter V. Coveney (Apr. 2019). "Big data: the end of the scientific method?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 377.2142. Publisher: Royal Society, p. 20180145. DOI: 10.1098/rsta.2018.0145 (Cited on p. 1).

Sunde, Alv (1995). "Kjempebølger I Nordsjøen (Extreme waves in the North Sea)." In: *Vær & Klima* 1 (Cited on p. 5).

Tayfun, M. Aziz (Nov. 1990). "Distribution of Large Wave Heights." In: *Journal of Waterway, Port, Coastal, and Ocean Engineering* 116.6, pp. 686–707. DOI: 10.1061/(ASCE)0733-950X(1990)116:6(686) (Cited on p. 7).

Tayfun, M. Aziz and Francesco Fedele (Aug. 2007). "Wave-height distributions and nonlinear effects." In: *Ocean Engineering* 34.11, pp. 1631–1649. ISSN: 0029-8018. DOI: 10.1016/j.oceaneng.2006.11.006 (Cited on p. 7).

Toffoli, A. et al. (Dec. 2010). "Evolution of weakly nonlinear random directional waves: laboratory experiments and numerical simulations." en. In: *Journal of Fluid Mechanics* 664. Publisher: Cambridge University Press, pp. 313–336. ISSN: 1469-7645, 0022-1120. DOI: 10.1017/S002211201000385X (Cited on p. 10).

Trulsen, K., H. Zeng, and O. Gramstad (Sept. 2012). "Laboratory evidence of freak waves provoked by non-uniform bathymetry." In: *Physics of Fluids* 24.9. Publisher: American Institute of Physics, p. 097101. ISSN: 1070-6631. DOI: 10.1063/1.4748346 (Cited on p. 11).

Trulsen, Karsten (2018). "Rogue Waves in the Ocean, the Role of Modulational Instability, and Abrupt Changes of Environmental Conditions that Can Provoke Non Equilibrium Wave Dynamics." en. In: *The Ocean in Motion: Circulation, Waves, Polar Oceanography*. Ed. by Manuel G. Velarde, Roman Yu. Tarakanov, and Alexey V. Marchenko. Springer Oceanography. Cham: Springer International Publishing, pp. 239–247. ISBN: 978-3-319-71934-4. DOI: 10.1007/978-3-319-71934-4_17 (Cited on p. 11).

Udrescu, Silviu-Marian and Max Tegmark (Apr. 2020). "AI Feynman: A physics-inspired method for symbolic regression." In: *Science Advances* 6.16, eaay2631. ISSN: 2375-2548. DOI: 10.1126/sciadv.aay2631 (Cited on p. 13).

Ursell, F. (Oct. 1953). "The long-wave paradox in the theory of gravity waves." en. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 49.4. Publisher: Cambridge University Press, pp. 685–694. ISSN: 1469-8064, 0305-0041. DOI: 10.1017/S0305004100028887 (Cited on p. 9).

Varoquaux, Gaël and Veronika Cheplygina (2021). "How I failed machine learning in medical imaging–shortcomings and recommendations." In: *arXiv preprint arXiv:2103.10292* (Cited on p. 1).

Vinyals, Oriol et al. (Nov. 2019). "Grandmaster level in StarCraft II using multi-agent reinforcement learning." en. In: *Nature* 575.7782. Number: 7782 Publisher: Nature Publishing Group, pp. 350–354. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1724-z (Cited on p. 1).

Voit, Eberhard O. (Sept. 2019). "Perspective: Dimensions of the scientific method." en. In: *PLOS Computational Biology* 15.9. Publisher: Public Library of Science, e1007279. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007279 (Cited on pp. 2, 13).

Wigner, Eugene P. (1960). "The unreasonable effectiveness of mathematics in the natural sciences." en. In: *Communications on Pure and Applied Mathematics* 13.1, pp. 1–14. ISSN: 1097-0312. DOI: 10.1002/cpa.3160130102 (Cited on pp. 12, 68).

Wilson, Andrew Gordon and Pavel Izmailov (Feb. 2020). "Bayesian Deep Learning and a Probabilistic Perspective of Generalization." In: *arXiv:2002.08791 [cs, stat]*. arXiv: 2002.08791 (Cited on p. 68).

Xiao, Wenting et al. (Apr. 2013). "Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution." en. In: *Journal of Fluid Mechanics* 720. Publisher: Cambridge University Press, pp. 357–392. ISSN: 0022-1120, 1469-7645. DOI: 10.1017/jfm.2013.37 (Cited on p. 10).

Xu, Da et al. (Nov. 2019). "Self-attention with Functional Time Representation Learning." In: *arXiv:1911.12864 [cs, stat]*. arXiv: 1911.12864 (Cited on p. 1).

Zakharov, V. E. (Mar. 1968). "Stability of periodic waves of finite amplitude on the surface of a deep fluid." en. In: *Journal of Applied Mechanics and Technical Physics* 9.2, pp. 190–194. ISSN: 1573-8620. DOI: 10.1007/BF00913182 (Cited on p. 11).

Zanna, Laure and Thomas Bolton (2020). "Data-Driven Equation Discovery of Ocean Mesoscale Closures." en. In: *Geophysical Research Letters* 47.17, e2020GL088376. ISSN: 1944-8007. DOI: 10.1029/2020GL088376 (Cited on p. 13).

This document was typeset using the $\LaTeX\ 2_\varepsilon$ document class `dionsthesis`, based on `uiothesis` developed by Eivind Uggedal. It uses Linux Libertine, developed by the Libertine Open Fonts Project, and Fira Sans, developed by the Mozilla Foundation, as body fonts. `dionsthesis` is available at:

https://github.com/dionhaefner/dionsthesis/

The style of `uiothesis` was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. Typographic, structural and graphical decisions in this document follow the ideas presented in Jean-Luc Doumont's book *"Trees, Maps, and Theorems"*.