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 $\mathcal{S}$:

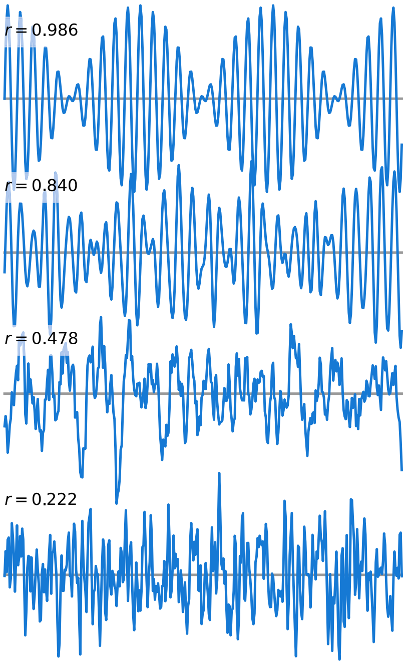


FIG. 1. The crest–trough correlation r is higher in ‘‘groupy,’’ low-bandwidth sea states. Shown are surface elevations generated from Ochi–Hubble spectra (Ochi and Hubble 1976) with increasing spectral bandwidth (from top to bottom) and the corresponding value of r.

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

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

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

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:

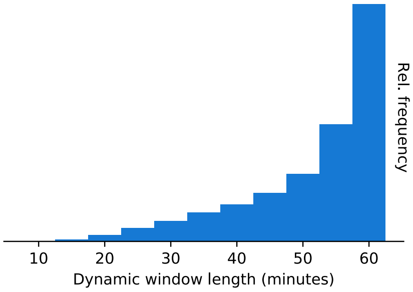


FIG. 2. Most dynamic windows are longer than 30 min. Shown is a histogram of the determined optimal window size across all Hawaiian CDIP stations.

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*

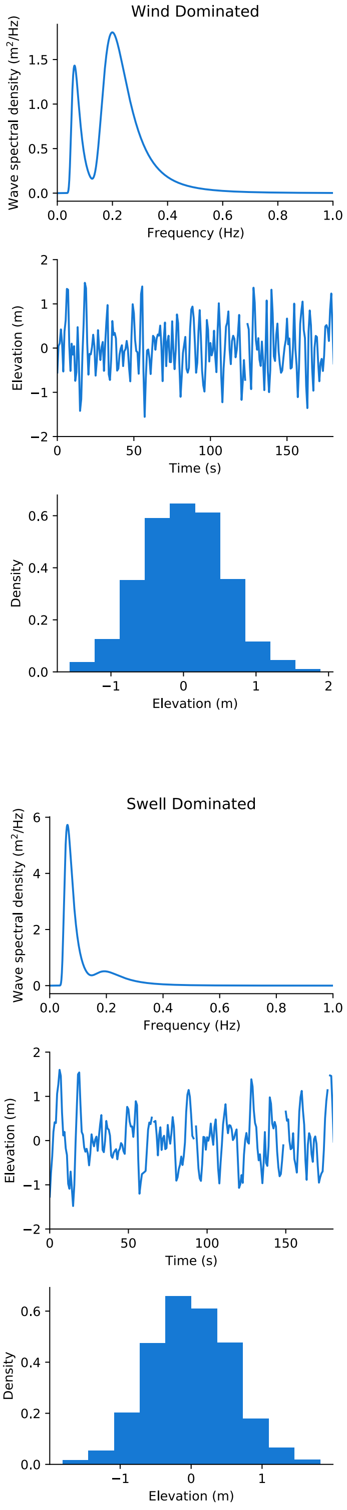
Because of FOWD's running-window approach (see section 2b), FOWD output datasets are about 10 times as big as the input surface elevation time series (since every wave results in about 80 output features). This demands that the processing implementation does not store entire output files in memory.

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 semi-automated *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

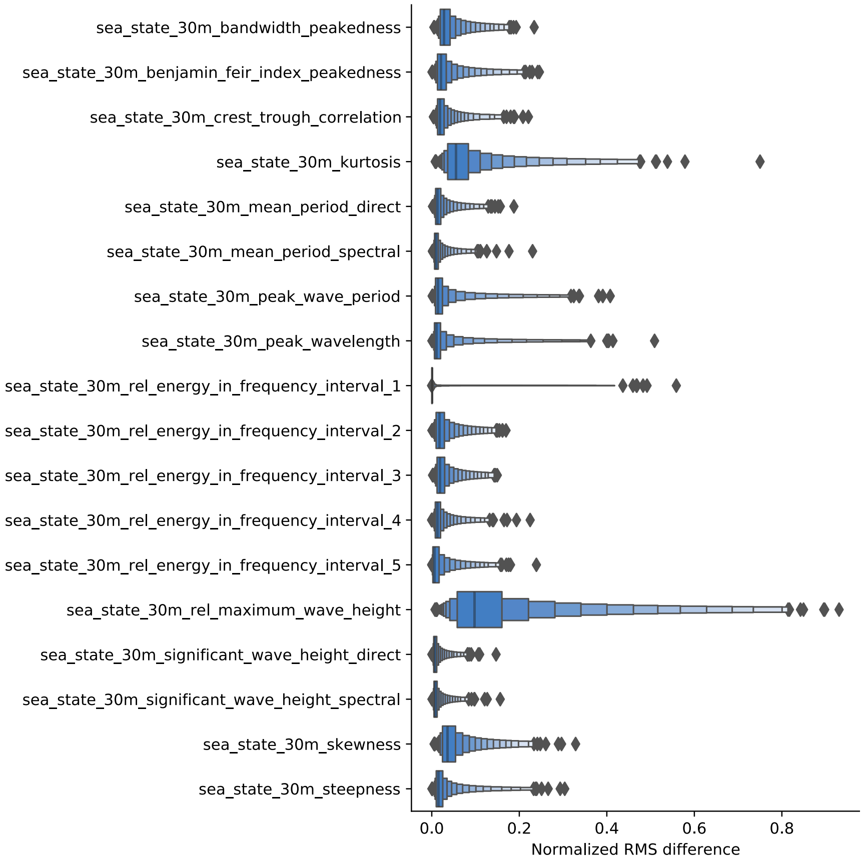


FIG. 4. In extreme cases, using running windows (instead of fixed chunks) leads to RMS differences of up to 50% of the characteristic scale of a parameter (Table 4). Shown is the distribution of normalized RMS difference between processing based on running windows and fixed chunks for some parameters.

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

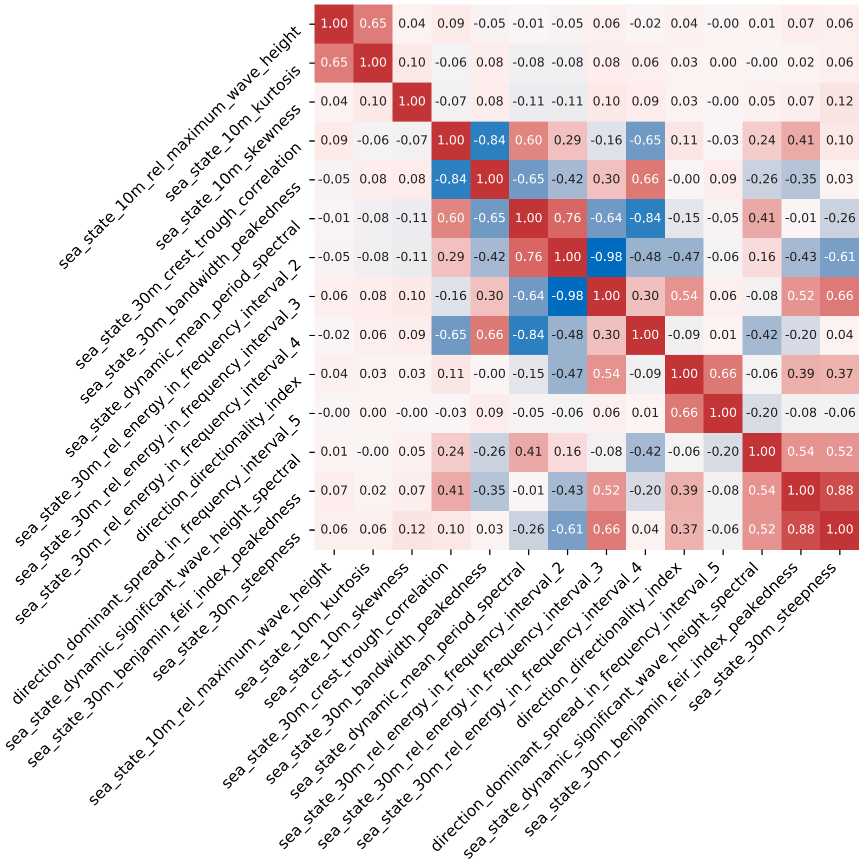


FIG. 5. Linear (Pearson) correlation matrix of selected parameters. Almost all parameters are strongly correlated with at least one other parameter, but exceptions exist (e.g., skewness, kurtosis/maximum wave height, and wind sea directional spread).

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

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

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.

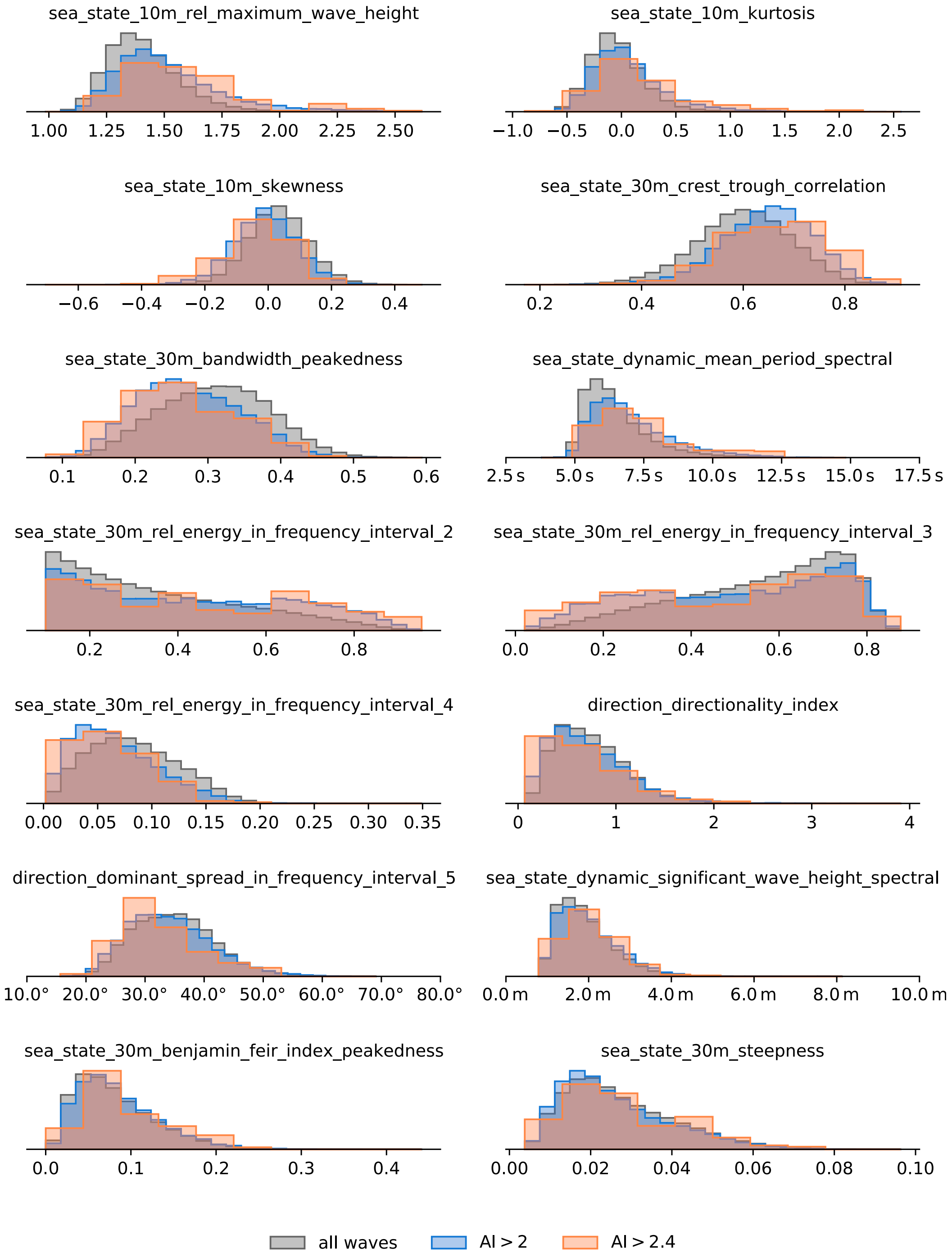


FIG. 6. Most parameters show a clear difference between the probability distributions of all sea states and those containing an extreme wave, but some just show a weak dependence (e.g., directional spread, significant wave height, and steepness). Shown are the probability density functions (PDFs) of 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.

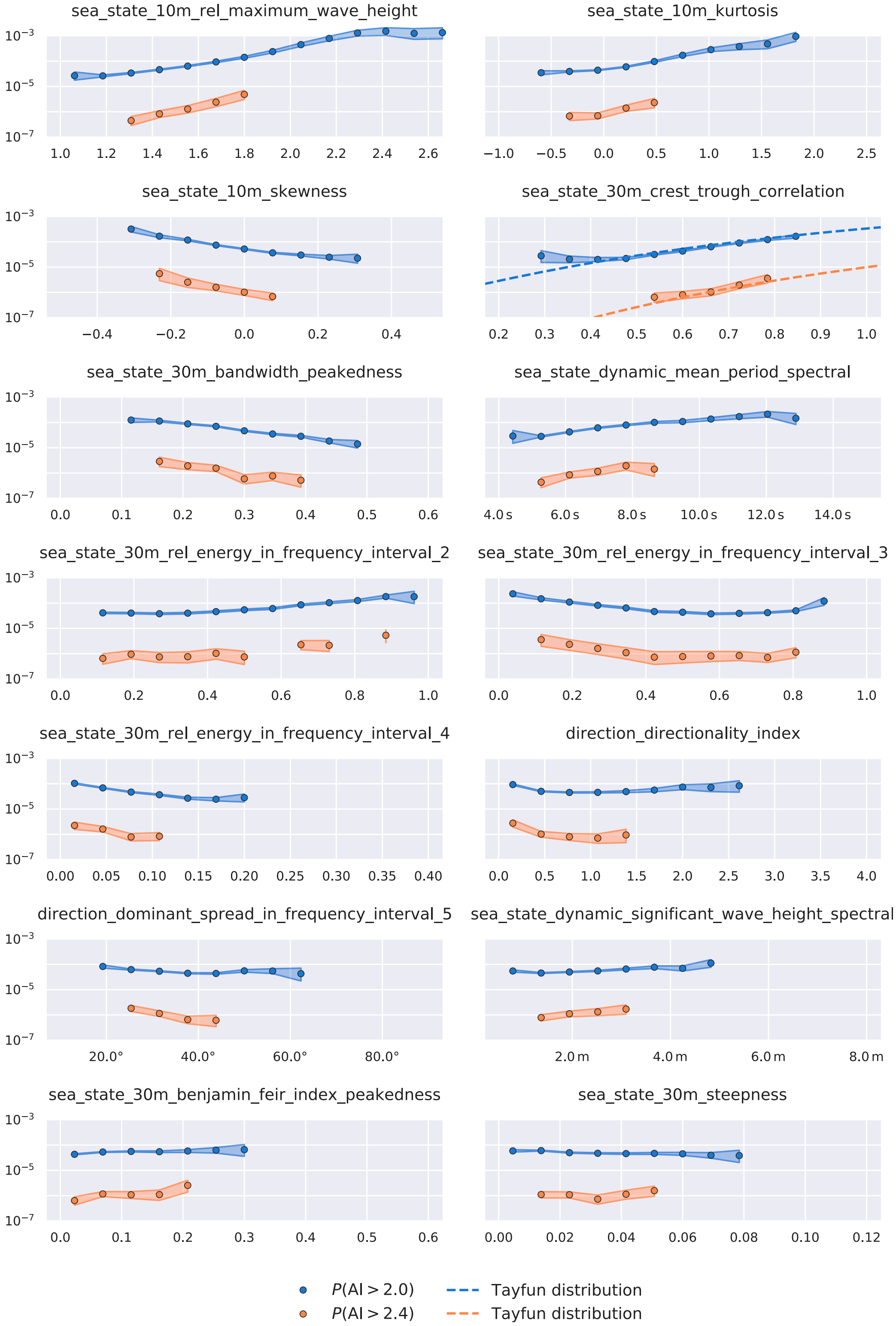


FIG. 7. Some sea state parameters are much more informative for rogue wave activity than others. Shown is the dependence of the rogue wave probability on several sea state parameters for AI > 2 and AI > 2.4. Symbols represent rogue wave probability posterior mean; shading represents the 95% minimum credible 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	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.