Machine-Guided Discovery of a Real-World Rogue Wave Model

Dion Häfner[a,b,1], **Johannes Gemmrich**[c], **and Markus Jochum**[b]

[a]Pasteur Labs, Brooklyn, NY, USA; [b]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark; [c]University of Victoria, Victoria, British Columbia, Canada

This manuscript was compiled on September 25, 2023

Big data and large-scale machine learning have had a profound impact on science and engineering, particularly in fields focused on forecasting and prediction. Yet, it is still not clear how we can use the superior pattern matching abilities of machine learning models for scientific *discovery*. This is because the goals of machine learning and science are generally not aligned. In addition to being accurate, scientific theories must also be causally consistent with the underlying physical process and allow for human analysis, reasoning, and manipulation to advance the field. In this paper, we present a case study on discovering a new symbolic model for oceanic rogue waves from data using causal analysis, deep learning, parsimony-guided model selection, and symbolic regression. We train an artificial neural network on causal features from an extensive dataset of observations from wave buoys, while selecting for predictive performance and causal invariance. We apply symbolic regression to distill this black-box model into a mathematical equation that retains the neural network's predictive capabilities, while allowing for interpretation in the context of existing wave theory. The resulting model reproduces known behavior, generates well-calibrated probabilities, and achieves better predictive scores on unseen data than current theory. This showcases how machine learning can facilitate inductive scientific discovery, and paves the way for more accurate rogue wave forecasting.

ocean waves | rogue waves | machine learning | symbolic regression | causality

Rogue waves are extreme ocean waves that have caused countless accidents, often with fatal consequences (1). They are defined as waves whose crest-to-trough height H exceeds a threshold relative to the significant wave height H_s. The significant wave height is defined as four times the standard deviation of the sea surface elevation. Here, we use a rogue wave criterion with a threshold of 2.0:

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

A rogue wave is therefore by definition an unlikely sample from the tail of the wave height distribution, and can in principle occur by chance under any circumstance. This 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 (2). However, the availability of large observation arrays (3) makes them an ideal target for machine-learning based analysis (4, 5).

In this study, we present a neural network-based model that predicts rogue wave probabilities from the sea state, trained solely on observations from buoys (6). The resulting model respects the causal structure of rogue wave generation; therefore, it can generalize to unseen physical regimes, is robust to distributional shift, and can be used to infer the relative importance of rogue wave generation mechanisms.

While a causally consistent neural network is useful for prediction and qualitative insight into the physical dynamics, the ability for scientists to analyze, test, and manipulate a model is crucial to recognize its limitations and integrate it into the research canon. Despite advances in interpretable AI (7), this is still a major challenge for most machine learning models.

To address this, we transform our neural network into a concise equation using symbolic regression (8, 9). The resulting model combines several known wave dynamics, outperforms current theory in predicting rogue wave occurrences, and can be interpreted within the context of wave theory. We see this as an example of "data-mining inspired induction" (10), an extension to the scientific method in which machine learning guides the discovery of new scientific theories.

We achieve this through the following recipe (Fig. 1):

1. A-priori analysis of causal pathways that leads to a set of presumed causal parameters (Section 1).
2. Training an ensemble of regularized neural network predictors, and parsimony-guided model selection based on causal invariance (Section 2).
3. Distillation of the neural network into a concise mathematical expression via symbolic regression (Section 3).

Finally, we analyze both the neural network and symbolic model in the context of current wave theory (Section 4). Both models reproduce well-known behavior and point towards new insights regarding the relative importance of different mechanisms in the real ocean.

Significance Statement

Machine learning has had a transformative impact on predictive science and engineering. But due to their black-box nature, better machine learning models do not always lead to greater human understanding, the first goal of science. We show how this can be overcome by using machine learning to transform a vast database of wave observations into a new, human-readable equation for the occurrence probability of rogue waves — rare ocean waves that routinely damage ships and offshore structures. This equation can be analyzed and incorporated into the research canon. Our work demonstrates the potential of causal analysis, machine learning, and symbolic regression to drive scientific discovery in a real-world application.

All authors 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

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

[1]To whom correspondence should be addressed. E-mail: dion.haefner@simulation.science

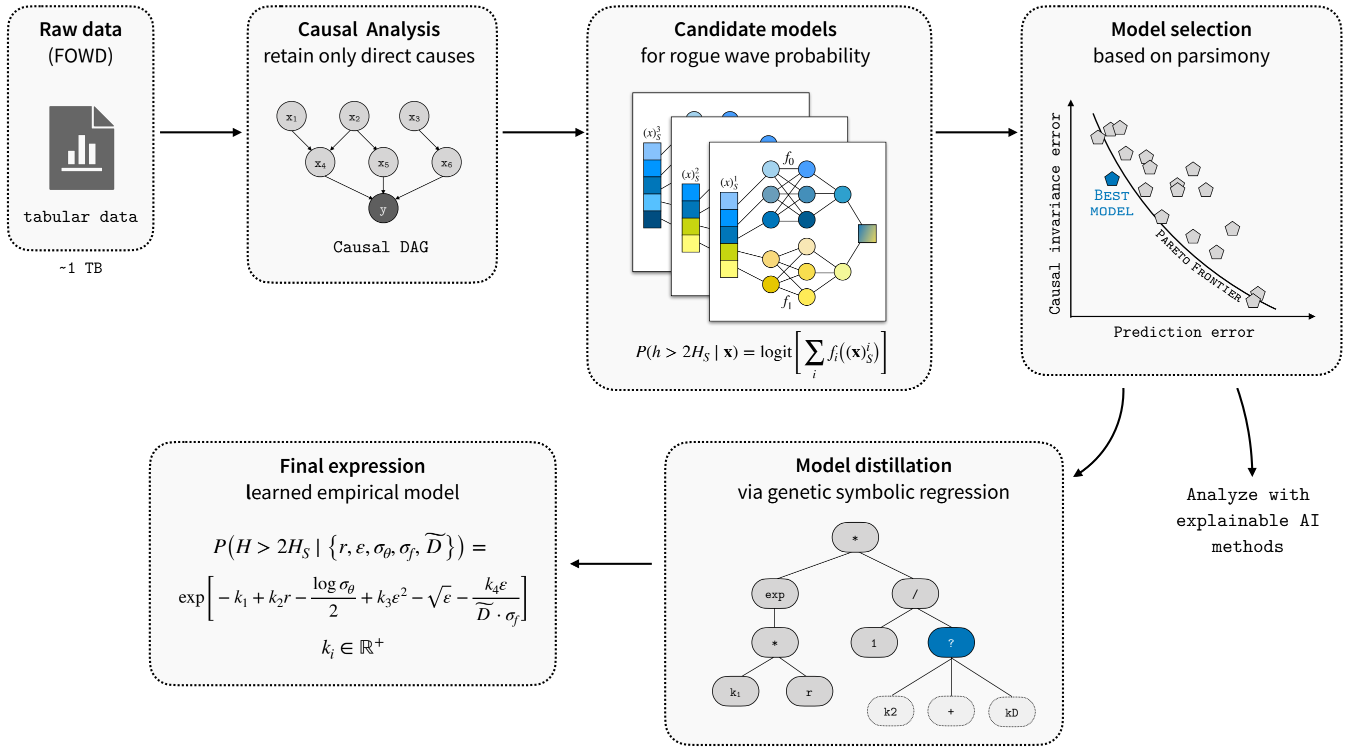

Fig. 1. Overview of our study. Starting out with large amounts of tabular data from wave buoys, we use a causal analysis to identify the most important features for predicting rogue waves. We then train an ensemble of neural networks on subsets of these features, and select the best one based on its predictive performance and causal invariance. Finally, we use symbolic regression to distill the model into a concise mathematical equation. We analyze the neural network and symbolic expression in terms of their performance on unseen data, and compare them to existing theory. This closes the arc between data, machine learning, and theory.

1. A causal graph for rogue wave generation

To create a causal machine learning model it is crucial to expose it only to parameters with causal relevance. Otherwise, the model may prefer to encode spurious associations over true causal relationships, simply because they can be easier to learn. This requires us to identify the causal structure of rogue wave generation.

There are several hypothesized causes of rogue waves (see 11, for an overview). Typically, research focuses on linear superposition in finite-bandwidth seas (12), wave breaking (13), and wave-wave interactions in weakly nonlinear seas (14, 15) or through the modulational instability (16). Apart from these universal mechanisms, there are also countless possible interactions with localized features such as non-uniform topography (17), wave-current interactions like in the Agulhas (18) or the Antarctic Circumpolar Current (19), or crossing sea states at high crossing angles affecting wave breaking (20). We call this set of mechanisms the *physical effects* Φ.

Since ocean waves are generated by a complex dynamical system, their true cause is a set of extrinsic *environmental conditions* E that are high-dimensional and not feasible to capture in full detail. However, most physical effects are mediated by one or several *sea state parameters* $\mathcal{P}$, which are the characteristic aggregated parameters that appear in theoretical models of the respective wave dynamics, and that are included in operational wave forecasts. In this study, we would like obtain a model that relates relevant sea state parameters $\mathcal{P}$ to wave observations $\mathcal{O}$, which ideally also lets us infer the relative importance of physical effects Φ.

The go-to tool to analyze causal relationships is a causal DAG (Directed Acyclic Graph; 21). In a causal DAG, nodes represent variables and edges $A \to B$ imply that A is a cause of B (usually in the probabilistic sense in that the probability distribution $P(B)$ depends on A).

We create a causal graph for rogue wave formation based on the hypothesized causal mechanisms discussed above and their corresponding theoretical models and parameters (Fig. 2). Following this causal structure, we use the following set of sea state parameters as candidates for representing the various causal pathways (see Methods section for more information on each parameter):

- Crest-trough correlation r, to account for the linear effect of wave groups on crest-to-trough rogue waves (22). r is the dominant causal factor behind linear rogue wave formation (4).

- Steepness ε governing weakly nonlinear effects, such as second-order and third-order bound waves, and wave breaking (13, 23).

- 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 (24).

- Relative depth $\widetilde{D}$ (based on peak wavelength), which is central for nonlinear shallow-water effects (25, 26) and wave breaking (13).

- Dominant directional spread σ_θ, which has an influence

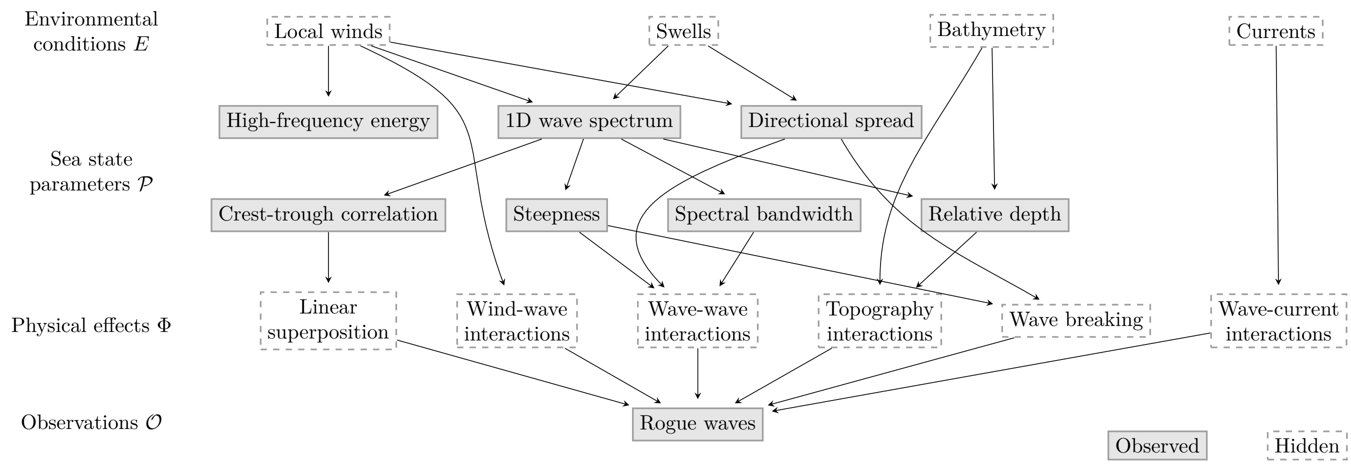

Fig. 2. The causes of rogue waves as a causal DAG (directed acyclic graph). Arrows $A \to B$ imply that A causes B.

on third-order nonlinear waves (26) and wave breaking (20).

- Spectral bandwidth ν_f (narrowness) and σ_f (peakedness), appearing for example in the expression for the influence of third-order nonlinear waves (26).

We also include a number of derived parameters that commonly appear in wave models and govern certain nonlinear (wave-wave) phenomena:

- Benjamin-Feir index BFI, which controls third-order nonlinear free waves (26) and the modulational instability (27).
- Ursell number Ur, which quantifies nonlinear effects in shallow water (28).
- Directionality index R (the ratio of directional spread and spectral bandwidth), which has an influence on third-order nonlinear free waves and is typically used in conjunction with the BFI (26).

These parameters cover most causal pathways towards rogue wave generation. Still, there are some at least partially unobserved causes, as we do not have access to data on local winds, topography, or currents. Additionally, our in-situ measurements are potentially biased estimates of the true sea state parameters, and there is no guarantee that any given training procedure will converge to the true causal model. This implies that we cannot rely on a model being causally consistent by design; instead, we perform a-posteriori verification on the learned models to find the perfect trade-off between causal consistency and predictive performance (see Section 2C).

2. An approximately causal neural network

A. Input data. We use the Free Ocean Wave Dataset (FOWD, 6), which contains 1.4 billion wave measurements recorded by the 158 CDIP wave buoys (3) along the Pacific and Atlantic coasts of the US, Hawaii, and overseas US territories. Water depths range between 10 m to 4000 m, and we require a significant wave height of at least 1 m. Each buoy records the sea surface elevation at a sampling frequency of 1.28 Hz, producing over 700 years of time series in total. FOWD extracts every zero-crossing wave from the surface elevation data and computes a number of characteristic sea state parameters from the history of the wave within a sliding window.

Due to the massive data volume of the full FOWD catalogue ($\sim$ 1 TB), we use an aggregated version that maps each sea state to the maximum wave height of the following 100 waves (as in 4). This reduces the data volume by a factor of 100 and inflates all rogue wave probabilities to a bigger value $\hat{p}$. We correct for this via $p = 1 - (1 - \hat{p})^{1/100}$, assuming that rogue waves occur independently from each other. This is a good approximation in most conditions, but may underestimate seas with a strong group structure (see Section 5B).

The final dataset has 12.9M data points containing over 100,000 rogue waves exceeding 2 times the significant wave height. Our dataset is freely available for download (see Data Availability section).

B. Neural network architecture. The probability to measure a rogue wave based on the sea state can be modelled as a sum of nonlinear functions, each of which only depends on a subset of the sea state parameters representing a different causal path (act via different *physical effects* in Fig. 2):

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

Here, y is a binary label indicating whether the current wave is a rogue wave, $\mathbf{x}^{(S_i)}$ is the i-th subset of all causal sea state parameters $\mathbf{x}$, $\operatorname{logit}(p) = \log(p) - \log(1-p)$ 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 f_i, we can restrict which parameters may interact non-additively with each other, which is an additional regularizing constraint that increases interpretability and prevents interactions between inputs from different causal pathways. For example, to include the effects of linear superposition and nonlinear corrections for free and bound waves (as in 29), Eq. (2) can be written as:

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

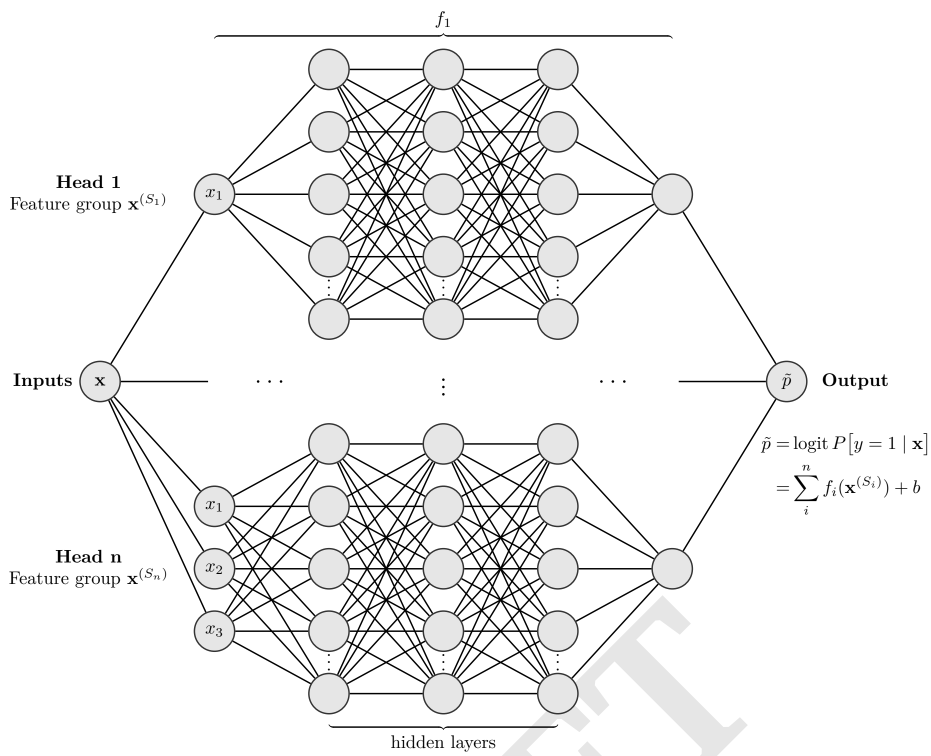

Fig. 3. Neural network architecture (multi-head FCN) used to predict rogue wave probabilities. Each input head receives a different subset of the full parameter set $\mathbf{x}$ to limit the amount of non-causal interactions between parameters.

We use a neural network with fully-connected layers (FCN) to model the functions f_i, which are universal function approximators (30), 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 with a linear output layer (Fig. 3). We use a small feed-forward architecture with 3 hidden layers and ReLU activation functions (rectified linear units, 31). To the best of our knowledge, this is the first time that a multi-head neural network has been used to restrict the interactions between input parameters to be consistent with a causal model.

The neural network outputs a scalar $\tilde{p} \in (-\infty, \infty)$, the log-odds of a rogue wave occurrence for the given sea state. For training, we use the Adam optimizer (32) and backpropagation to minimize a cross-entropy loss for binary classification with an added ℓ_2 regularization term for kernel parameters:

$$L(p, y, \theta) = y \cdot \log(p) + (1 - y) \cdot \log(1 - p) + \lambda \|\theta\|_2 \qquad [4]$$

with predicted probability $p = \text{logit}^{-1}(\tilde{p})$, 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, 33). For this, 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 used to construct a multivariate Gaussian approximation of the loss surface that we can sample from. This results in slightly better predictions, and gives us a way to quantify how confident the neural network is in its predictions.

C. Causal consistency and predictive accuracy. Although we include only input parameters that we assume to have a direct causal connection with rogue wave generation, there is no guarantee that the neural network will infer the correct causal model. In fact, the presence of measurement bias and unobserved causal paths makes it unlikely that the model will converge to the true causal structure. To search for an approximately causally consistent model we will have to quantify its causal performance.

We achieve this through the concept of invariant causal prediction (ICP; 34, 35). The key insight behind ICP is that the parameters of the true causal model will be invariant under distributional shift, that is, an intervention on an upstream "environment" node in the causal graph that controls which distribution the data is drawn from. Re-training the model on data with different spurious correlations *between* features should still lead to the same dependency of the target *on* the features (see also 36).

We split the dataset randomly into separate training and validation sets, in chunks of 1M waves. We train the model on the full training dataset and perform ICP on the validation dataset, which we partition into subsets representing different conditions in space, time, depth, spectral properties, and degrees of non-linearity (Table 1). This changes the dominant characteristics of the waves in each subset (representing e.g. storm and swell conditions), inducing distributional shift. Then, we re-train the model separately on each subset and compute the root-mean-square difference between predictions

Table 1. The subsets of the validation dataset used to evaluate model performance and invariance.

Subset name	Condition	# waves
southern-california	Longitude $\in (-123.5, -117)^{\circ}$, latitude $\in (32, 38)^{\circ}$	265M
deep-stations	Water depth $> 1000\,\mathrm{m}$	28M
shallow-stations	Water depth $< 100\,\mathrm{m}$	154M
summer	Day of year $\in (160, 220)$	51M
winter	Day of year $\in (0, 60)$	91M
Hs > 3m	$H_s > 3\,\mathrm{m}$	58M
high-frequency	Relative swell energy < 0.15	43M
low-frequency	Relative swell energy > 0.7	46M
long-period	Mean zero-crossing period $> 9\,\mathrm{s}$	100M
short-period	Mean zero-crossing period $< 6\,\mathrm{s}$	42M
cnoidal	Ursell number > 8	40M
weakly-nonlinear	Steepness > 0.04	83M
low-spread	Directional spread $< 20^{\circ}$	25M
high-spread	Directional spread $> 40^{\circ}$	25M
full	(all validation data)	472M

of the re-trained model P_k and the full model P_{tot} on the k-th data subset $\mathbf{x}_{(k)}$:

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

As the total consistency error we use the root-mean-square of Eq. (5) 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 always 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 of these assumptions are violated here, we merely search for an approximately causal model that minimizes $\mathcal{E}$.

However, we cannot use $\mathcal{E}$ as the only criterion when selecting a model. The invariance error can only account for change in the prediction (variance), but not for its overall closeness to the true solution (bias). Therefore, we select a model that is Pareto-optimal with respect to the invariance error $\mathcal{E}$ and a predictive score $\mathcal{L}$. This will not establish absolute causal consistency, but will allow us to select a model that is near-optimal given the constraints.

For $\mathcal{L}$ we use the log of the likelihood ratio between the predictions of our neural network and a baseline model that predicts the empirical base rate $\overline{y}_k = \frac{1}{n} \sum_i^n y_{k,i}$, averaged over all environments k:

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

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

To evaluate model calibration (the tendency to produce over- or under-confident probabilities), we compute a calibration curve by binning the predicted rogue wave probabilities. We then compare each bin to the observed rogue wave frequency, and compute the weighted root-mean-square residual between measured ($\overline{y}_i$) and predicted (p_i) log-odds:

$$\mathcal{C} = \sqrt{\sum_{i=1}^{n_b} w_i \left(\mathrm{logit}(p_i) - \mathrm{logit}(\overline{y}_i) \right)^2} \qquad [9]$$

To account for uncertainty in the observations (e.g. close to the extremes), the weights w_k are based on the 33 % credible interval of $\overline{y}_i \sim \mathrm{Beta}(n_i^+, n_i^-)$ with n_i^+ rogue and n_i^- non-rogue measurements. This is similar to the expected calibration error (37), but models data uncertainty directly. We use a uniform bin size (in logit space) of 0.1.

D. Model selection. We train a total of 24 candidate models on different subsets of the relevant causal parameters (as identified in Section 1) and varying number of input heads (between 1 and 3). We evaluate their performance in terms of calibration, predictive performance, and causal consistency (Table 3).

We observe a clear anti-correlation between model complexity and predictive score on one hand and causal consistency on the other hand (Fig. 4). This is evidence that more complex models are indeed less biased but exploit more non-causal connections. We perform model selection based on *parsimony*: A good model is one where a small increase in either predictive performance or causal consistency implies a large decrease in the other, i.e., where the Pareto front is convex. This is similar to the metric used by PySR (9) to select the best symbolic regression model (Section 3).

Based on this, we choose model 18 with parameter groups $S_1 = \{r\}, S_2 = \{\varepsilon, \sigma_\theta, \sigma_f, \widetilde{D}\}$ (i.e., a model with two input heads) as the reference model for further analysis. The chosen model produces well-calibrated probabilities (Fig. 5), and is among the 5 best models in terms of predictive performance on the test dataset (not used during training or selection), despite using only 5 features with at most 4-way interactions.

The relatively low number of input features allows us to analyze the model in detail using explainable AI methods (Section 4A).

3. Learning an empirical equation for rogue wave risk

To make our model fully interpretable, we transform the learned neural network into an equation via symbolic regression. Common approaches to symbolic regression include Eureqa (39), AI Feynman (40), SINDy (41), and QLattice (42). Here, we use PySR (8, 9), a symbolic regression package based on genetic programming (43). Genetic algorithms build a large ensemble of candidate models and select the best ones, before mutating and recombining them into the next generation. In the case of symbolic regression, mathematical expressions are represented as a tree of constants and elementary symbols. In principle, this allows PySR to discover expressions of unbounded complexity.

PySR's central metric to quantify the goodness of an equation is again based on *parsimony*, in the form of the derivative of predictive performance with respect to the model complexity — if the true model has been discovered, any additional complexity can at best lead to minor performance gains (by overfitting to noise in the data).

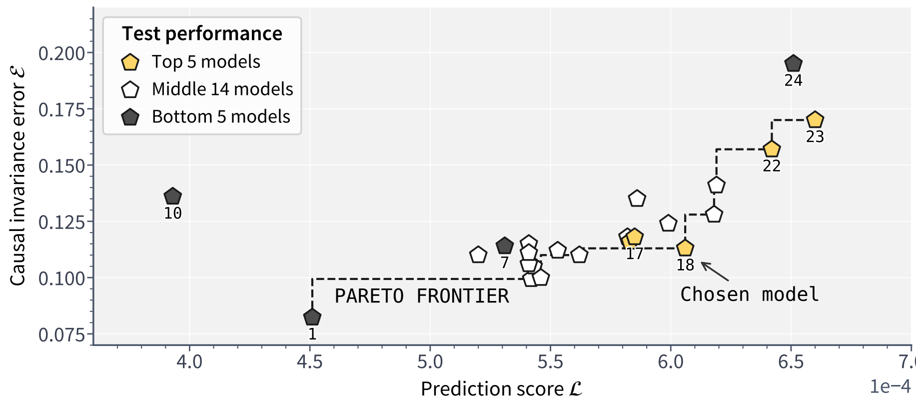

Fig. 4. There is a clear trade-off between causal invariance ($\mathcal{E}$) and predictive performance ($\mathcal{L}$) of our neural network predictors. We choose the model that lies in the most convex part of the Pareto frontier. Scores are evaluated on validation data. Test performance is based on prediction scores on held-out test data (from unseen stations).

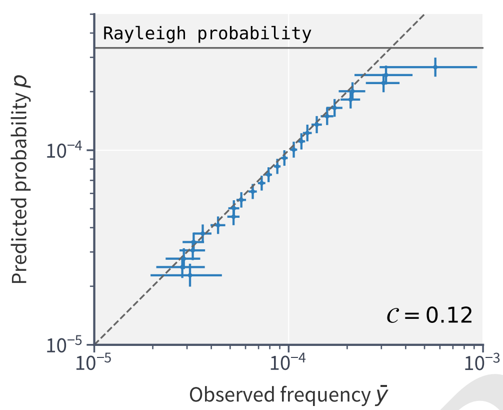

Fig. 5. Our model outputs well-calibrated probabilities, even for unseen stations. Shown is the binned predicted probability p vs. the observed rogue wave frequency $\overline{y}$ on the test data. Error bars for p indicate 3 standard deviations estimated via SWAG sampling. Error bars for $\overline{y}$ indicate 95 % credible interval assuming $\overline{y}_i \sim \mathrm{Beta}(n_i^+, n_i^-)$. Bins with less than 10 observed rogue waves are excluded. Dashed line indicates perfect calibration. Solid line indicates probability as predicted by linear theory in the narrow-bandwidth limit (Rayleigh distribution; 38).

In our case, we seek to find an expression f from the space of possible expression graphs $\mathcal{T}_O$ with allowed operators O that approximates the rogue wave log-probability as predicted by the neural network $\mathcal{N}$ over the dataset x:

$$\text{Find } f \in \mathcal{T}_O \text{ that minimizes } \sum_i \frac{1}{\mathrm{Var}(y_i)} \left[f(x_i)^2 - \sigma(\mathbb{E}[y_i])^2 \right]$$

where $\sigma(x) = -\log(1 + \exp(-x))$, and y_i is the set of SWAG samples from $\mathcal{N}(x_i)$. A sensible set of operators O is key to ensure interpretability of the resulting expression; we choose the symbols $O = \{+, -, \times, \div, \log, \cdot^{-1}, \sqrt{\cdot}, \cdot^2\}$ to facilitate expressions that are similar to current theoretical models of the form $P \sim A \exp(B)$. We normalize all input features to approximately unit scale by converting directional spread to radians.

PySR assembles a league of candidate expression and presents the Pareto-optimal solutions of increasing complexity to the user. We select the best solution by hand, picking the expression with the best parsimony score that contains all input features and at least two terms containing the steepness ε (to account for the various causal pathways in which steepness affects rogue waves). The final equation is shown in Fig. 8, and discussed in Section 4B.

4. Results

A. Neural network. We analyze the behavior of our neural network predictor, which reveals important insights about the physical dynamics of rogue waves and their prediction.

A.1. 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 in either metric. Especially the exclusion of crest-trough correlation leads to catastrophic results, even when including other bandwidth measures like σ_θ in its place (Table 3).

This suggests that the above set of parameters represents the dominant rogue wave generation processes in the form of linear superposition in finite-bandwidth seas with a directional contribution and weakly nonlinear corrections.

The crest-trough correlation r is still lacking mainstream adoption as a rogue wave indicator (for example, it is not part of ECMWF's operational forecast; 29), despite being a key parameter for crest-to-trough rogue waves (4, 22, 44). The other parameters are consistent with other empirical studies such as Fedele (45), which considers the same parameters in conjunction with rogue crests during storms. They are also similar to the ingredients to ECMWF's rogue wave forecast (29), which is based on the effects of 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.

Numerous previous studies have found the BFI to be a poor predictor of rogue wave risk in realistic sea states (4, 14, 15, 45–48) due to its strong underlying assumptions such as unidirectionality. This study extends this to the fully nonparametric and nonlinear case.

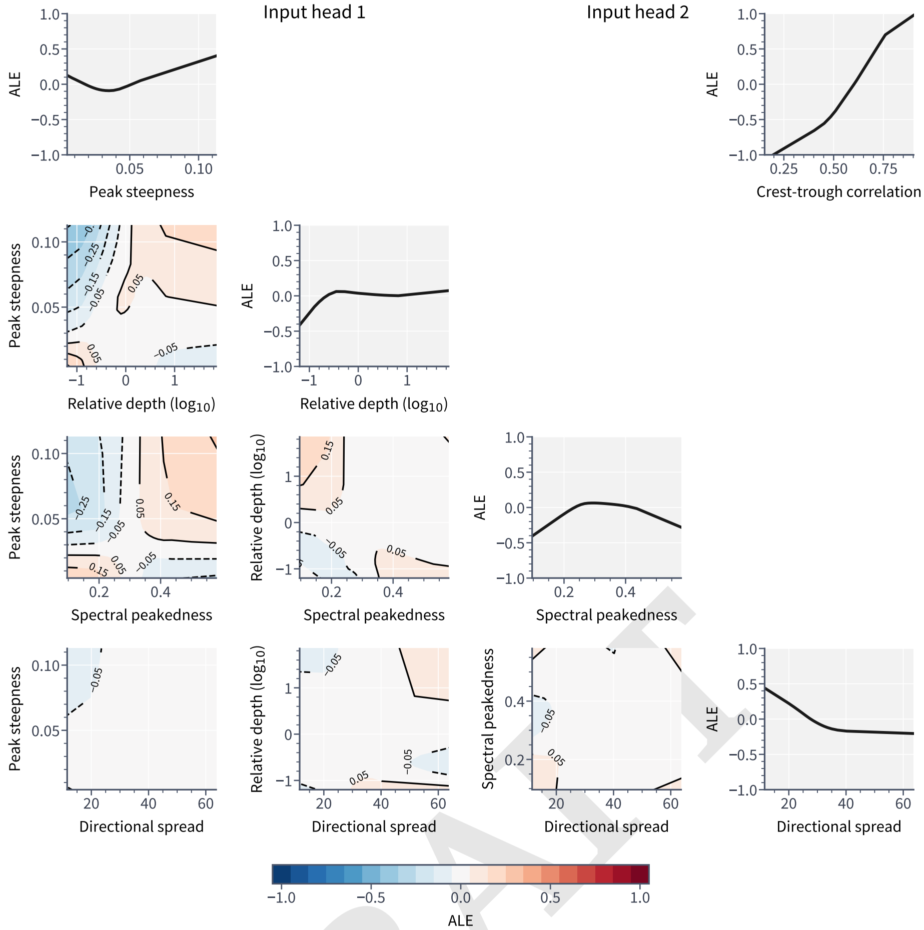

Fig. 6. ALE (accumulated local effects) plot matrix for experiment 18. Shown is the change in rogue wave risk (in logits) from the average as each parameter is varied. The total effect is the sum of all 1D, 2D, and higher-order contributions (not shown).

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 the accumulated local effects decomposition (ALE, 49), which measures the influence of infinitesimal changes in each parameter on the prediction outcome (see also 7). From the ALE plot (Fig. 6), we find that crest-trough correlation has by far the biggest influence of all parameters and explains about 1 order of magnitude in rogue wave risk variation, which is consistent with earlier model-free approaches (4). To first order, higher crest-trough correlation, lower directional spread, larger relative depth (deep water), and higher steepness lead to larger rogue wave risk, but parameter interactions can lead to more complicated, non-monotonic relationships (for example in very shallow water, see Section 4A.3).

A.2. The Rayleigh distribution is an upper bound for real-world rogue wave risk. Despite the clear enhancement by 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 narrow-band waves (38), i.e., the limit $r \to 1$, $\varepsilon \to 0$, $\sigma_f \to 0$, $\widetilde{D} \to \infty$, and $\sigma_\theta \to 0$, and reads:

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

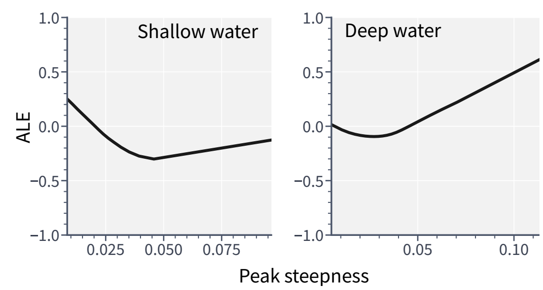

Fig. 7. Our model predicts a positive association between steepness and rogue waves in deep water, and a negative association in shallow water. Shown is the 1-dimensional ALE (accumulated local effects) plot in both cases. Here, deep water are sea states with $\widetilde{D} > 3$ and shallow water with $\widetilde{D} < 0.1$.

Only in the most extreme conditions does our model predict a similarly high probability, for example for $\sigma_\theta = 13^\circ$, $\varepsilon = 0.008$, $\sigma_f = 0.14$, $r = 0.88$, and $\widetilde{D} = 0.6$, which gives the same probability as the Rayleigh distribution, $p = 3.3 \times 10^{-4}$.

In the opposite extreme, rogue wave probabilities can fall to as little as 10^{-5} for low values of r and high values of σ_θ (such as 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.

$$P\big(H > 2H_S \mid r, \varepsilon, \sigma_\theta, \sigma_f, \widetilde{D}\big) = \exp\left[\underbrace{-\,12. + 3.8r}_{\text{I}} - \underbrace{\frac{\log \sigma_\theta}{2}}_{\text{II}} + \underbrace{66.\varepsilon^2}_{\text{III}} - \underbrace{\sqrt{\varepsilon}}_{\text{IV}} - \underbrace{\frac{0.23\varepsilon}{\widetilde{D}\cdot\sigma_f}}_{\text{V}}\right]$$

Fig. 8. Our empirical equation for rogue wave risk, as identified through the distillation of our neural network predictor via symbolic regression. This equation outperforms existing wave theory on unseen stations from our dataset, while being fully interpretable. Numbered terms are discussed in Section 4B. All floating point coefficients are rounded to two significant digits.

A.3. There is a clear separation between deep water and shallow water regimes. All models with high causal invariance scores include an interaction between steepness and relative water depth. Looking at this more closely, we find that a stratification on deep and shallow water sea states reveals 2 distinct regimes (Fig. 7).

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 (26). The opposite is true in shallow water ($\widetilde{D} < 0.1$), where we find a clear *negative* association with steepness. This is likely due to depth-induced wave breaking (23). 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.

B. Symbolic expression. The final expression for the rogue wave probability, as discovered via symbolic regression, is given in Fig. 8. It consists of an exponential containing five additive terms:

(I) $-12 + 3.8r$. The term with the largest coefficients is the one containing r, as expected. Comparison with the exponential term in the Tayfun distribution P_t, Eq. (28), reveals that this is approximately a linear expansion around $r \approx 1$:

$$\log P_t(H/H_s > h) \sim -\frac{4h^2}{1+r} \qquad [11]$$

$$= -12 + 4r + \mathcal{O}(r^2)\Big|_{r\approx 1}^{h=2} \qquad [12]$$

This is an important sanity check for the model, since it shows that it is able to re-discover existing theory purely from data.

(II) $-\log\sigma_\theta/2$. This encodes the observed enhancement for narrow sea states and has no direct relation to existing quantitative theory. Its functional form is somewhat problematic, since it causes the model to diverge for $\sigma_\theta \to 0$ (unidirectional seas). However, the model has only seen real-world seas with $\sigma_\theta \gtrsim 0.2$, so we may replace this term with one that yields similar predictions for the relevant range of σ_θ, and does not diverge for $\sigma_\theta \to 0$.

One possible candidate is

$$\frac{1-\sigma_\theta}{1+\sigma_\theta}, \qquad [13]$$

which has a relative RMS error of about 5 % over the range $\sigma_\theta \in (20, 90)^\circ$ compared to the original term.

(III) $66\varepsilon^2$. Encodes the influence of weakly nonlinear effects for large values of $\varepsilon \gtrsim 0.1$.

(IV) $-\sqrt{\varepsilon}$. This term encodes the observed negative association between steepness and rogue waves for low values of ε that could be due to wave breaking, or may be an artifact of our sensor.

(V) $0.23\varepsilon/(\widetilde{D}\cdot\sigma_f)$. Since $\widetilde{D} \sim k_p D$ and $\varepsilon \sim k_p H_s$, this term is proportional to the relative wave height $\eta = H_s/D$ and $1/\sigma_f$. η is the most important parameter in the theory of shallow-water waves, and appears for example in the Korteweg-de Vries equation (25). Accordingly, this term dominates the dynamics in very shallow water. Dependencies on $1/\sigma_f$ occur in current theory (26), but are usually paired with σ_θ to form the directionality index R. This suggest that term V may be incomplete, and missing physical dynamics that are not prevalent in the data.

Overall, the equation is able to reproduce the same qualitative behavior as observed from the neural network, with the same well-calibrated outputs ($\mathcal{C} = 0.14$) and predictive performance (Section 5A) on the test data.

5. Discussion

A. Validation against theory. We test our models (neural network and symbolic equation) against existing wave theory based on their mean predictive score $\mathcal{L}$ across the environments from Table 1 on the held-out test data (unseen stations). As theoretical baselines we use the models from Longuet-Higgins (Rayleigh, 38), Tayfun (22), Mori & Janssen (50), and a hybrid combining Tayfun and Mori & Janssen (see Methods section).

The results are shown in Fig. 9. Since the Rayleigh and Mori & Janssen models do not account for crest-trough correlation, their predictions vastly overestimate the occurrence rate of observed rogue waves. The Tayfun and hybrid models perform better, but are still outperformed by our models except in cnoidal seas. Our models are better predictors than the baseline (predicting the empirical per-environment rogue wave frequency) in all environments.

The neural network performs better than the symbolic equation in all environments, albeit only by a small margin. This shows that the symbolic equation is able to capture the main features of the full model, despite its compact representation.

B. Limitations. Using only wave buoy observations for our analysis, we acknowledge the following limitations:

- We did not have sufficient data on local winds, currents, or topography, which implies that some relevant causal pathways are unobserved (see Fig. 2). While we expect these effects to play a minor role in bulk analysis, they could dramatically affect local rogue wave probabilities in specific conditions, for example over sloping topography (17) or in strong currents (51).

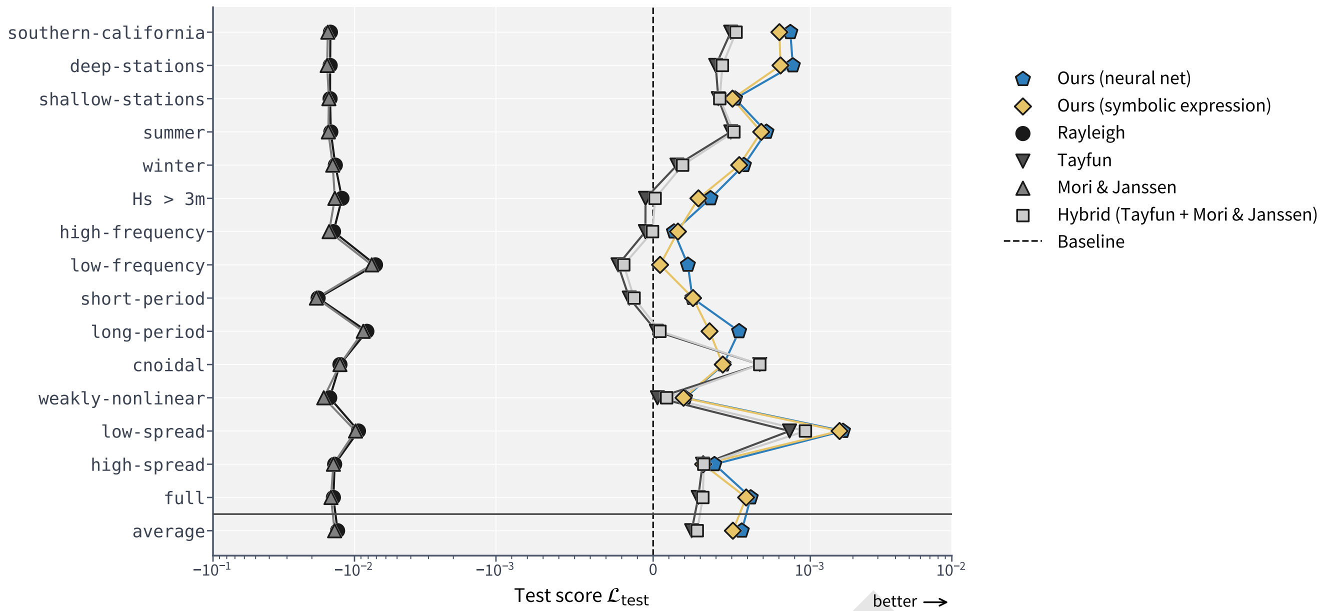

Fig. 9. Comparison between our models and existing theory on held-out test data. Our models perform similar to each other and outperform existing theory on this dataset in all but one data subset (cnoidal seas). x-scale is linear in $(-10^{-3}, 10^{-3})$, and logarithmic otherwise.

- We only have one-dimensional (time series) data and cannot capture imported parameters, such as solitons generated elsewhere that travel into the observation area. While we expect this to play a minor role, it could underestimate the importance of nonlinear free waves.

- Systematic sensor bias is common in buoys and can lead to spurious causal relationships. This may obscure the true causal structure and hurt model generalization to other sensors. However, this adaptation to sensor characteristics may be desirable in forecasting scenarios, where it allows the model to synthesize several noisy quantities into more robust ones.

- By aggregating individual waves into 100-wave chunks, we underestimate the per-wave rogue wave probability in sea states in which rogue waves do not occur independently of each other, such as seas with a strong group structure.

These limitations could potentially reduce our model's ability to detect relevant causal pathways and underestimate the true rogue wave risk. Our analysis is agnostic to the data source and can be repeated on different sources to validate our findings.

6. Next steps

A. An improved rogue wave forecast. Our empirical model can be compared directly to existing rogue wave risk indicators by evaluating them on forecast sea state parameters. ECMWF's operational rogue wave forecast (29) focuses on envelope wave heights which does not account for crest-trough correlation, and is conceptually similar to the Mori & Janssen model in Section 5A. Therefore, we are confident that substantial improvements are within reach in terms of predicting crest-to-trough rogue waves, even without using a black-box model.

B. Predicting super-rogue waves. Observed wave height distributions often show a flattening of the wave height distribution towards the extreme tail (11, 14, 52). Therefore, we expect rogue wave probabilities to be more pronounced for even more extreme waves (for example with $H/H_s > 2.4$, as recently observed in 53).

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 free parameters of a candidate wave height probability distribution (such as 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.

C. Commoditization of data-mining based induction. There is a pronounced lack of established methods for machine learning aimed at scientific discovery. We have shown that incorporating and enforcing causal structure can overcome many of the shortcomings of standard machine learning approaches, like poorly calibrated predictions, non-interpretability, and incompatibility with existing theory. However, the methods we leveraged are still in their infancy and rely on further community efforts to be end-to-end automated and adopted at scale. Particularly, parsimony-based model selection (as in Section 2D and Section 3) is still a manual process that requires a firm understanding of model intrinsics and the domain at hand. Nonetheless, we believe that the potential benefits of causal and parsimony-guided machine learning for real-world problems are too great to ignore, and we hope that this study will inspire further research in this direction.

Materials and Methods

Sea state parameters. Here, we give the definition of the sea state parameters used in this study. For a more thorough description of how parameters are computed from buoy displacement time series see Häfner et al. (6).

All parameters can be derived from the non-directional wave spectrum $\mathcal{S}(f)$, 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.

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

$$m_n = \int_0^\infty f^n \mathcal{S}(f)\,\mathrm{d}f \qquad [14]$$

The expressions for the relevant sea state parameters are:

- Significant wave height:
$$H_s = 4\sqrt{m_0} \qquad [15]$$
- Spectral bandwidth (narrowness):
$$\nu_f = \sqrt{m_2 m_0/m_1^2 - 1} \qquad [16]$$
- Spectral bandwidth (peakedness):
$$\sigma_f = \frac{m_0^2}{2\sqrt{\pi}}\left(\int_0^\infty f \cdot \mathcal{S}(f)\,\mathrm{d}f\right)^{-1} \qquad [17]$$
- Peak wavenumber k_p, computed via the peak period (as in 54):
$$\overline{T}_p = \frac{\int \mathcal{S}(f)^4\,\mathrm{d}f}{\int f\cdot\mathcal{S}(f)^4\,\mathrm{d}f} \qquad [18]$$
This leads to the peak wavenumber through the dispersion relation for linear waves in intermediate water of depth D:
$$f(k)^2 = \frac{gk}{(2\pi)^2}\tanh(kD) \qquad [19]$$
An approximate inverse is given in Fenton (55).
- Relative depth, based on the wave length λ:
$$\widetilde{D} = \frac{D}{\lambda} = \frac{1}{2\pi}k_p D \qquad [20]$$
- Peak steepness:
$$\varepsilon = H_s k_p \qquad [21]$$
- Benjamin-Feir index:
$$\mathrm{BFI} = \frac{\varepsilon\nu}{\sigma_f}\sqrt{\max\{\beta/\alpha, 0\}} \qquad [22]$$
where ν, α, β are coefficients depending only on $\widetilde{D}$ (full expression given in 56).
- Directionality index:
$$R = \frac{\sigma_\theta^2}{2\nu_f^2} \qquad [23]$$
- Crest-trough correlation:
$$r = \frac{1}{m_0}\sqrt{\rho^2+\lambda^2} \qquad [24]$$
$$\rho = \int_0^\infty \mathcal{S}(\omega)\cos\left(\omega\frac{\overline{T}}{2}\right)\mathrm{d}\omega \qquad [25]$$
$$\lambda = \int_0^\infty \mathcal{S}(\omega)\sin\left(\omega\frac{\overline{T}}{2}\right)\mathrm{d}\omega \qquad [26]$$
where ω is the angular frequency and $\overline{T} = m_0/m_1$ the spectral mean period (12).

Model implementation and hyperparameters. All performance critical model code is implemented in JAX (57), using neural network modules from flax (58) and optimizers from optax (59). 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 whole training process can also be executed on CPU in about 2 hours. The hyperparameters for all experiments are shown in Table 2.

Table 2. Hyperparameters used in experiments.

Hyperparameters	
Optimizer	Adam
Learning rate	10^{-4}
Number of hidden layers	3
Neurons in hidden layers	$(32/\sqrt{n_h}, 16/\sqrt{n_h}, 8/\sqrt{n_h})$
ℓ_2 penalty λ_2	10^{-5}
Number of training epochs	50
Number of SWAG epochs	50
Number of SWAG posterior samples	100
Train-validation split	60% – 40%

n_h: number of input heads.

Full list of experiments. See Table 3.

Reference wave height distributions. We use the following theoretical wave height exceedance distributions for comparison (with rogue wave threshold κ, here $\kappa = 2$):

- Rayleigh (38):
$$P_{\mathrm{R}}(\kappa) = \exp\left(-2\kappa^2\right) \qquad [27]$$
- Tayfun (12, 22):
$$P_{\mathrm{T}}(\kappa) = \exp\left(\frac{-4}{1+r}\kappa^2\right) \qquad [28]$$
- Mori & Janssen (50, 60):
$$P_{\mathrm{MJ}}(\kappa) = \left(1+\frac{2\pi}{3\sqrt{3}}\frac{\mathrm{BFI}^2}{1+7.1R}\kappa^2(\kappa^2-1)\right)\exp\left(-2\kappa^2\right) \qquad [29]$$
- Hybrid:
$$P_{\mathrm{H}}(\kappa) = \left(1+\frac{2\pi}{3\sqrt{3}}\frac{\mathrm{BFI}^2}{1+7.1R}\kappa^2(\kappa^2-1)\right)\exp\left(\frac{-4}{1+r}\kappa^2\right) \qquad [30]$$

Data Availability. The preprocessed and aggregated version of the FOWD CDIP data used in this study is available for download at https://erda.ku.dk/archives/ee6b452c1907fbd48271b071c3cee10e/published-archive.html. All model code is openly available at https://github.com/dionhaefner/rogue-wave-discovery.

ACKNOWLEDGMENTS. Dion Häfner received funding from the Danish Offshore Technology Centre (DOTC). 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. Portions of this work were developed from the doctoral thesis of Dion Häfner (61). The authors thank Jonas Peters for helpful discussions in the early stages of this work. The authors thank two anonymous reviewers for their helpful comments. This publication was made possible by the following open-source software stack: JAX (57), flax (58), optax (59), PySR (9), scikit-learn (62), PyALE (63), NumPy (64), SciPy (65), matplotlib (66), Seaborn (67), pandas (68), Jupyter (69).

Table 3. 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	$\{r\}$			4.51	8.23	3.35
2	$\{r, R\}$	$\{\mathrm{Ur}\}$		5.42	9.94	5.54
3	$\{r, R, \mathrm{BFI}\}$			5.43	10.50	5.60
4	$\{r, R\}$	$\{\mathrm{Ur}, R\}$		5.46	9.99	4.57
5	$\{r, R\}$	$\{\varepsilon, \widetilde{D}\}$		5.53	11.20	5.79
6	$\{r, \varepsilon, \widetilde{D}\}$			5.20	11.00	3.60
7	$\{r, \varepsilon, R\}$			5.31	11.40	6.97
8	$\{r, \widetilde{D}, R\}$			5.41	11.50	7.31
9	$\{\varepsilon, \widetilde{D}, R\}$			-0.13	24.80	7.60
10	$\{\sigma_f\}$	$\{\varepsilon, \widetilde{D}, R\}$		3.93	13.60	9.02
11	$\{r\}$	$\{\varepsilon, \widetilde{D}, R\}$		5.41	10.60	7.18
12	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, R\}$	5.41	11.10	6.02
13	$\{r, R\}$	$\{\widetilde{D}, \varepsilon, \sigma_\theta\}$		5.99	12.40	4.06
14	$\{r, R\}$	$\{\widetilde{D}, \varepsilon, \sigma_f\}$		5.82	11.80	6.37
15	$\{r, R\}$	$\{\widetilde{D}, \varepsilon, R\}$		5.62	11.00	5.45
16	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta\}$			5.83	11.60	5.94
17	$\{r\}$	$\{\varepsilon, \widetilde{D}\}$	$\{\mathrm{BFI}, \sigma_f, \sigma_\theta\}$	5.85	11.80	6.40
18	$\boldsymbol{\{r\}}$	$\boldsymbol{\{\varepsilon, \widetilde{D}, \sigma_f, \sigma_\theta\}}$		6.06	11.30	4.43
19	$\{r, \varepsilon, \widetilde{D}, R, \lambda_p\}$			5.86	13.50	7.13
20	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \nu\}$			6.18	12.80	6.78
21	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \nu, E_h\}$			6.19	14.10	6.71
22	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, \nu, E_h\}$			6.42	15.70	4.97
23	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, \mathrm{BFI}, R\}$			6.60	17.00	5.75
24	$\{r, \varepsilon, \widetilde{D}, \sigma_\theta, \sigma_f, E_h, H_s, \overline{T}, \kappa, \mu, \lambda_p\}$			6.51	19.50	4.95

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

1. E Didenkulova, Catalogue of rogue waves occurred in the World Ocean from 2011 to 2018 reported by mass media sources. *Ocean. & Coast. Manag.* p. 105076 (2019).
2. JM Dudley, G Genty, A Mussot, A Chabchoub, F Dias, Rogue waves and analogies in optics and oceanography. *Nat. Rev. Phys.* **1**, 675–689 (2019) Number: 11 Publisher: Nature Publishing Group.
3. J Behrens, J Thomas, E Terrill, R Jensen, CDIP: Maintaining a Robust and Reliable Ocean Observing Buoy Network in *2019 IEEE/OES Twelfth Current, Waves and Turbulence Measurement (CWTM).* pp. 1–5 (2019).
4. D Häfner, J Gemmrich, M Jochum, Real-world rogue wave probabilities. *Sci. Reports* **11**, 10084 (2021) Number: 1 Publisher: Nature Publishing Group.
5. A Cattrell, M Srokosz, B Moat, R Marsh, Can rogue waves be predicted using characteristic wave parameters? *J. Geophys. Res. Ocean.* **123**, 5624–5636 (2018).
6. D Häfner, J Gemmrich, M Jochum, FOWD: A Free Ocean Wave Dataset for Data Mining and Machine Learning. *J. Atmospheric Ocean. Technol.* **-1** (2021) Publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology.
7. C Molnar, *Interpretable Machine Learning.* (2020).
8. M Cranmer, et al., Discovering symbolic models from deep learning with inductive biases. *NeurIPS 2020* (2020).
9. M Cranmer, Interpretable machine learning for science with pysr and symbolicregression.jl (2023).
10. EO Voit, Perspective: Dimensions of the scientific method. *PLOS Comput. Biol.* **15**, e1007279 (2019).
11. TAA Adcock, PH Taylor, The physics of anomalous ('rogue') ocean waves. *Reports on Prog. Phys.* **77**, 105901 (2014).
12. MA Tayfun, F Fedele, Wave-height distributions and nonlinear effects. *Ocean. Eng.* **34**, 1631–1649 (2007).
13. M Miche, Mouvements ondulatoires de la mer en profondeur constante ou decroissante. *Annales de Ponts et Chaussees, 1944, pp(1) 26-78, (2)270-292, (3) 369-406* (1944).
14. J Gemmrich, C Garrett, Dynamical and statistical explanations of observed occurrence rates of rogue waves. *Nat. Hazards Earth Syst. Sci.* **11**, 1437–1446 (2011).
15. F Fedele, J Brennan, S Ponce de León, J Dudley, F Dias, Real world ocean rogue waves explained without the modulational instability. *Sci. Reports* **6**, 27715 (2016).
16. M Onorato, et al., Extreme waves, modulational instability and second order theory: wave flume experiments on irregular waves. *Eur. J. Mech. - B/Fluids* **25**, 586–601 (2006).
17. K Trulsen, H Zeng, O Gramstad, Laboratory evidence of freak waves provoked by non-uniform bathymetry. *Phys. Fluids* **24**, 097101 (2012) Publisher: American Institute of Physics.
18. JK Mallory, Abnormal Waves on the South East Coast of South Africa. *The Int. Hydrogr. Rev.* (1974).
19. EG Didenkulova, TG Talipova, EN Pelinovsky, Rogue Waves in the Drake Passage: Unpredictable Hazard in *Antarctic Peninsula Region of the Southern Ocean: Oceanography and Ecology*, Advances in Polar Ecology, eds. EG Morozov, MV Flint, VA Spiridonov. (Springer International Publishing, Cham), pp. 101–114 (2021).
20. ML McAllister, S Draycott, TaA Adcock, PH Taylor, TSvd Bremer, Laboratory recreation of the Draupner wave and the role of breaking in crossing seas. *J. Fluid Mech.* **860**, 767–786 (2019).
21. J Pearl, *Causality.* (Cambridge university press), (2009).
22. MA Tayfun, Distribution of large wave heights. *J. waterway, port, coastal, ocean engineering* **116**, 686–707 (1990).
23. Y Goda, Reanalysis of Regular and Random Breaking Wave Statistics. *Coast. Eng. J.* **52**, 71–106 (2010).
24. T Tang, D Barratt, HB Bingham, TS van den Bremer, TA Adcock, The impact of removing the high-frequency spectral tail on rogue wave statistics. *J. Fluid Mech.* **953**, A9 (2022).
25. DJ Korteweg, G De Vries, 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, Dublin Philos. Mag. J. Sci.* **39**, 422–443 (1895).
26. P Janssen, Shallow-water version of the Freak Wave Warning System, (ECMWF), Technical memorandum 813 (2018).
27. PAEM Janssen, Nonlinear Four-Wave Interactions and Freak Waves. *J. Phys. Oceanogr.* **33**, 863–884 (2003) Publisher: American Meteorological Society.
28. F Ursell, The long-wave paradox in the theory of gravity waves. *Math. Proc. Camb. Philos. Soc.* **49**, 685–694 (1953) Publisher: Cambridge University Press.
29. ECMWF, Part VII: ECMWF Wave model in *IFS Documentation CY47R3*, IFS Documentation. (ECMWF), (2021).
30. K Hornik, Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251–257 (1991).
31. V Nair, GE Hinton, Rectified linear units improve restricted boltzmann machines in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10. (Omnipress, Madison, WI, USA), pp. 807–814 (2010).
32. DP Kingma, J Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
33. W Maddox, T Garipov, P Izmailov, D Vetrov, AG Wilson, A Simple Baseline for Bayesian Uncertainty in Deep Learning. *arXiv:1902.02476 [cs, stat]* (2019) arXiv: 1902.02476.
34. J Peters, P Bühlmann, N Meinshausen, Causal inference by using invariant prediction: identification and confidence intervals. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **78**, 947–1012 (2016).
35. J Peters, D Janzing, B Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, Adaptive Computation and Machine Learning series ed. F Bach. (MIT Press, Cambridge, MA, USA), (2017).
36. C Heinze-Deml, J Peters, N Meinshausen, Invariant causal prediction for nonlinear models. *J. Causal Inference* **6**, 20170016 (2018).
37. P Xenopoulos, J Rulff, LG Nonato, B Barr, C Silva, Calibrate: Interactive analysis of probabilistic model output. *IEEE Transactions on Vis. Comput. Graph.* **29**, 853–863 (2022).
38. MS Longuet-Higgins, On the statistical distribution of the height of sea waves. *JMR* **11**, 245–266 (1952).
39. M Schmidt, H Lipson, Distilling free-form natural laws from experimental data. *science* **324**, 81–85 (2009).
40. SM Udrescu, et al., Ai feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Adv. Neural Inf. Process. Syst.* **33**, 4860–4871 (2020).
41. SL Brunton, JL Proctor, JN Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. national academy sciences* **113**, 3932–3937 (2016).
42. KR Broløs, et al., An approach to symbolic regression using Feyn. *arXiv preprint arXiv:2104.05417* (2021).
43. JH Holland, Genetic algorithms. *Sci. american* **267**, 66–73 (1992).
44. F Fedele, MA Tayfun, On nonlinear wave groups and crest statistics. *J. Fluid Mech.* **620**, 221–239 (2009) Publisher: Cambridge University Press.
45. F Fedele, J Herterich, A Tayfun, F Dias, Large nearshore storm waves off the Irish coast. *Sci. Reports* **9**, 15406 (2019) Number: 1 Publisher: Nature Publishing Group.
46. O Gramstad, K Trulsen, Influence of crest and group length on the occurrence of freak waves. *J. Fluid Mech.* **582**, 463–472 (2007) Publisher: Cambridge University Press.
47. W Xiao, Y Liu, G Wu, DKP Yue, Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution. *J. Fluid Mech.* **720**, 357–392 (2013) Publisher: Cambridge University Press.
48. J Gemmrich, J Thomson, Observations of the shape and group dynamics of rogue waves. *Geophys. Res. Lett.* **44**, 1823–1830 (2017).
49. DW Apley, J Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv:1612.08468 [stat]* (2019) arXiv: 1612.08468.
50. N Mori, M Onorato, PA Janssen, On the estimation of the kurtosis in directional sea states for freak wave forecasting. *J. Phys. Oceanogr.* **41**, 1484–1497 (2011).
51. LH Ying, Z Zhuang, EJ Heller, L Kaplan, Linear and nonlinear rogue wave statistics in the presence of random currents. *Nonlinearity* **24**, R67–R87 (2011) Publisher: IOP Publishing.
52. M Casas-Prat, LH Holthuijsen, Short-term statistics of waves observed in deep water. *J. Geophys. Res. Ocean.* **115** (2010).
53. J Gemmrich, L Cicon, Generation mechanism and prediction of an observed extreme rogue wave. *Sci. Reports* **12**, 1–10 (2022).
54. IR Young, The determination of confidence limits associated with estimates of the spectral peak frequency. *Ocean. Eng.* **22**, 669–686 (1995).
55. JD Fenton, The numerical solution of steady water wave problems. *Comput. & Geosci.* **14**, 357–368 (1988).
56. M Serio, M Onorato, A R a Osborne, P Janssen, On the computation of the Benjamin-Feir Index. *Nuovo Cimento della Soc. Italiana di Fisica C* **28**, 893–903 (2005).
57. J Bradbury, et al., JAX: composable transformations of Python+NumPy programs (2018).
58. J Heek, et al., Flax: A neural network library and ecosystem for JAX (2020).
59. M Hessel, et al., Optax: composable gradient transformation and optimisation, in JAX! (2020).
60. N Mori, PAEM Janssen, On Kurtosis and Occurrence Probability of Freak Waves. *J. Phys. Oceanogr.* **36**, 1471–1483 (2006) Publisher: American Meteorological Society.
61. D Häfner, Ph.D. thesis (2022).
62. F Pedregosa, et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
63. D Jomar, PyALE: A Python implementation of accumulated local effect plots (2020).
64. CR Harris, et al., Array programming with NumPy. *Nature* **585**, 357–362 (2020).
65. P Virtanen, et al., SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).
66. JD Hunter, Matplotlib: A 2d graphics environment. *Comput. Sci. & Eng.* **9**, 90–95 (2007).
67. ML Waskom, seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
68. Wes McKinney, Data Structures for Statistical Computing in Python in *Proceedings of the 9th Python in Science Conference*, eds. Stéfan van der Walt, Jarrod Millman. pp. 56 – 61 (2010).
69. T Kluyver, et al., Jupyter notebooks - a publishing format for reproducible computational workflows in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, eds. F Loizides, B Scmidt. (IOS Press, Netherlands), pp. 87–90 (2016).