Reduced Order Modeling of Wave Energy Systems via Sequential Bayesian Experimental Design and Machine Learning

Eirini Katsidoniotaki[1,2,3], Stephen Guth[1], Malin Göteman[2,3], and Themistoklis P. Sapsis[1]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02138, USA
[2]Department of Electrical Engineering, Uppsala University, SE 75237, Sweden
[3]Centre of Natural Hazards and Disaster Science (CNDS), SE 75236, Sweden

October 3, 2023

Abstract

The establishment of wave energy technologies faces significant challenges regarding their survivability in extreme ocean conditions. A crucial aspect in designing reliable systems lies in quantifying the statistics of extreme forces exerted on structures. However, predicting these rare events poses a challenge due to limited available data and inherent uncertainty. While computational fluid dynamics (CFD) models can capture extreme forces in wave-structure interactions, obtaining statistical information using these models is computationally impractical. This study addresses the computational cost and challenges associated with quantifying extreme events by introducing a framework that combines novel schemes in Bayesian experimental design (BED) with machine learning techniques. The framework develops a surrogate model that maps extreme sea states and structural characteristics of a wave energy converter to the force in the mooring system. In the initial phase of the study, a hybrid surrogate model is constructed by combining two machine learning techniques: Gaussian process regression (GPR) and long-short term memory (LSTM) neural networks. This hybrid model is trained using data derived from CFD simulations, enabling it to learn the intricate input-output relationship. In the subsequent phase, an active learning approach is employed to construct an effective surrogate model that requires the most informative training data points. The active learning scheme in BED guides the sequential sample selection process by identifying parameter space regions that induce extreme forces while reducing model uncertainty. This study demonstrates the capability of machine learning to predict complex time series, including instantaneous peaks, at magnitudes much faster than classical modeling practices. Additionally, the carefully selected training samples expedite the convergence of the surrogate model, which ultimately provides accurate statistics of extreme forces. This framework offers a favorable and innovative approach for real-world applications where each sample evaluation demands significant time and resources.

Moreover, the resulting surrogate model can be integrated into the design stage to (1) enable cost reduction by determining appropriate strength margins and adjusting conservative safety factors and (2) enhance system reliability. Furthermore, this framework holds potential for optimizing designs in various other real-world applications.

Keywords– Extreme events; Gaussian process regression (GPR); LSTM neural networks; surrogate models; active sampling; sequential Bayesian experimental design; wave energy system; CFD simulations

1 Introduction

Ocean waves hold immense potential for global electricity production [6], yet the current capacity of active wave power remains limited. Although numerous wave energy conversion systems have been conceptualized, the majority have not yet reached commercial viability [12,16,35]. To become a competitive large-scale energy technology, wave energy converters (WECs) must overcome fundamental challenges. One of the foremost obstacles is ensuring their robustness and survivability in extreme ocean conditions. Under such conditions, critical components like mooring lines can experience loads that are up to 100 times higher than the average [8]. Therefore, quantifying the statistics of extreme force-on-structure is crucial for informing optimal design choices in terms of reliability and cost. This entails estimating the probability density function (PDF) that includes the tails representing the extreme events. By gaining insights into these extreme force statistics, researchers and designers can pave the way towards enhancing the resilience and performance of wave energy systems.

When investigating the dynamics of offshore structures in the presence of high and steep waves, nonlinear phenomena such as breaking waves, viscous effects, and wave overtopping play a crucial role as they give rise to extreme loads on the structure. While high-fidelity computational fluid dynamics (CFD) simulations enable accurate modeling of force-on-structure in extreme ocean conditions [18,19,22,40,53], their computational cost is prohibitively expensive, making it impractical to analyze the statistics of extreme forces using these models. On the other hand, frequency-domain models based on linear potential flow theory offer faster solutions but fail to capture the non-linear dynamics [5], rendering them unsuitable for studying the extreme responses of wave energy converters (WECs). Linear time-domain models are often employed in the preliminary design phase due to their computational efficiency, allowing for the study of WEC dynamics [17, 48, 54]. However, it should be noted that these models are susceptible to solution uncertainties stemming from nonlinear extreme wave phenomena [2,53]. Thus, there exists a need to strike a balance between computational efficiency and accuracy in modeling extreme WEC responses, considering both nonlinear effects and computational feasibility.

Surrogate models prove invaluable in cases that demand a balance between accuracy and reduced computational cost, such as statistics reconstruction. These models serve as statistical approximations of physical systems by learning the relationship between input variables and corresponding output variables [46]. In recent years, machine learning techniques have gained popularity for constructing surrogate models. Notably, several studies have implemented surrogate models in offshore engineering applications. For instance, GPR [14] and deep neural operators [38] have been utilized to capture ship load statistics in irregular seas. Artificial neural networks have been employed to develop surrogate models for estimating fundamental frequency [39], fatigue damage [29], and dimensioning [44] of offshore wind turbine structures. Other studies have utilized generalized polynomial chaos [33] and

LSTM neural networks [50] to predict the dynamic mooring force in floating structures. While limited in number, there are also studies that focus on surrogate modeling for wave energy converter (WEC) applications. In one case, an LSTM neural network-based surrogate model was developed to estimate the WEC output power [34]. Similarly, deep learning techniques were employed to control the energy absorption of WECs in other studies [1, 28]. Finally, digital twins for WEC systems were developed in [20, 30]. These examples demonstrate the growing application of surrogate models in various aspects of offshore engineering, including WEC systems.

The limited availability of data on extreme events presents a significant challenge in constructing surrogate models that accurately predict these rare occurrences. Existing methods for surrogate model construction often rely on random sampling from biased distributions in regions of the input space associated with extreme outcomes, without considering the associated costs and efficiency of obtaining the training samples [45]. However, active learning (AL) offers a promising solution by sequentially selecting the most informative training samples, thereby circumventing the need for a massive amount of data. AL has proven to be an effective scheme within the context of Bayesian experimental design (BED), particularly for uncertainty quantification of response probability density function (PDF) estimates in nonlinear dynamic systems, utilizing Bayesian compressive sampling techniques [23, 24]. Despite its potential, only a few studies have explored the application of AL to extreme event statistics, emphasizing the significance of acquisition functions [3, 4, 32, 38, 42, 43]. For instance, in [13, 38], AL was applied to investigate the structural response of ships under random waves. Similarly, leveraging AL in combination with machine learning techniques presents a promising and innovative approach for addressing the challenges faced by emerging wave energy technologies.

This work presents a framework that combines machine learning techniques with Bayesian experimental design (BED) to develop an effective surrogate model for wave energy converter (WEC) systems. The main contribution lies in accurately estimating the extreme force statistics by carefully selecting samples. In the first part of the study, two machine learning techniques, the GPR and LSTM neural networks, are employed to construct a hybrid surrogate model. This model captures the relationship between the inputs, such as wave episodes from 50-year extreme sea states and structural system parameters, and the output, which represents the force in the mooring system. The second part of the study focuses on guiding the selection of informative training samples, in this case wave episodes, for the surrogate model. By leveraging the active learning (AL) scheme within BED, the framework identifies the most relevant samples from regions in the input parameter space associated with extreme forces and higher uncertainty. This approach optimizes the use of resources and time in practical applications. The novelty of this study lies in its effective practices, which overcome limitations of previous approaches, in two key areas: (1) modeling the wave energy system and (2) accurately estimating the extreme force statistics. By enhancing the optimal design process of wave energy technologies, the presented method accelerates their development and establishment, by providing an efficient and accurate method for modeling and predicting extreme forces in wave energy systems, ultimately facilitating the advancement and implementation of these technologies.

The paper structure is as follows: Section 2 serves as an introduction by providing a comprehensive review of the methods utilized in this research. Section 3 delves into the specific wave energy application, offering a detailed description of the WEC system under investigation. The formulation of the hybrid surrogate model, combining two machine learning techniques, is presented in section 4. This section outlines how these techniques are coupled to create an accurate model of the wave energy system. Section 5 focuses on the active learning scheme in conjunction with machine learning techniques for the construction of an effective surrogate model. It highlights the methodology employed to select informative training samples and reduce prediction uncertainty. In section 6, the

results of the study are presented, showcasing the performance and capabilities of the developed surrogate model in predicting extreme force statistics. Finally, section 7 summarizes the key findings and draws meaningful conclusions based on the study's outcomes.

2 Background on Machine Learning Methods

2.1 Surrogate Modeling

From the modeling perspective, the physical system can be viewed as a 'black-box' that relates the input, $x \in \mathcal{X}$, which can be any stochastic variable, to the output, y, which represents the corresponding quantity of interest. The black-box constitutes the real function, $y = f(x)$, which characterizes the input-output relationship. The real function $f(x)$ can be approximated with a surrogate model $\hat{f}(x)$, which is a statistical model that learns the input-output relationship based on training data sets of known input-output pairs. Recently, machine learning techniques have been widely employed for the surrogate construction. The advantage of the surrogate modeling compared to classical modeling practices, e.g. CFD simulations, is that the output $\hat{y}_*$ for any new input x_* is provided in orders of magnitude faster. In sections 2.2 and 2.3, two popular machine learning techniques are presented while the developed active learning technique for training samples selection is described in section 2.5.

2.2 Gaussian Process Regression

GPR is a powerful class of supervised machine learning algorithms that is able to create surrogate models for predicting a quantity of interest, y, based on input independent variables, x. In many applications, the output is a noisy realization of the GPR defined function $\hat{f}(x)$,

$$y = \hat{f}(x) + \epsilon. \tag{1}$$

where the noise is Gaussian, $\epsilon \sim N(0, \sigma_n^2)$. For a random variable x, the function $\hat{f}(x)$ follows a Gaussian distribution

$$\hat{f}(x) \sim \mathcal{GP}\left(m(x), k(x, x')\right). \tag{2}$$

Unlike other machine learning methods that learn exact values for every parameter in a function, the GPR implements Bayesian approach to infer a probability distribution over functions by defining the mean function, $m(x)$, and the covariance function, $k(x, x')$, of the process $\hat{f}(x)$ at any point $x \in \mathbb{R}^d$ [51].

$$\begin{aligned} m(x) &= \mathbb{E}\left[\hat{f}(x)\right], \\ k(x, x') &= \mathbb{E}[(\hat{f}(x) - m(x))(\hat{f}(x') - m(x'))] \end{aligned} \tag{3}$$

The GPR based surrogate model learns the input-output behavior from available training datasets. Assume the training dataset D_{train} of n observations, and testing dataset D_{test} of n' observations

$$\begin{aligned} D_{train} &= (\mathbf{X}, \mathbf{y}) = \{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y \in \mathbb{R}, \\ D_{test} &= \mathbf{X}_* = \{x_{*,i}\}_{i=1}^{n'}, x_{*,i} \in \mathbb{R}^d. \end{aligned} \tag{4}$$

According to the prior distribution, the joint distribution of the training outputs, $\mathbf{y}$, and the test outputs, $\mathbf{y}_*$ is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_n^2 I & K(\mathbf{X},\mathbf{X}_*) \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right), \tag{5}$$

where $K(\boldsymbol{X},\boldsymbol{X})$, $K(\boldsymbol{X},\boldsymbol{X}_*)$, $K(\boldsymbol{X}_*,\boldsymbol{X}_*)$ denote the $n \times n$, $n \times n'$, $n' \times n'$ covariance matrices. However, the posterior distribution can provide more accurately information about the prediction, $\mathbf{y}_*$. The posterior distribution is defined by conditioning the joint Gaussian prior distribution on the training outputs $\mathbf{y}$, training inputs $\boldsymbol{X}$ and test inputs $\boldsymbol{X}_*$. Therefore the posterior distribution at $\boldsymbol{X}_*$ is defined with mean, $\bar{\mathbf{y}}_*$, and variance, σ^2

$$\bar{\mathbf{y}}_* = K(\mathbf{X}_*,\mathbf{X})[K(\mathbf{X},\mathbf{X}) + \sigma_n^2 I]^{-1}\mathbf{y}, \tag{6}$$

$$\sigma^2 = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})[K(\mathbf{X},\mathbf{X}) + \sigma_n^2 I]^{-1}K(\mathbf{X},\mathbf{X}_*). \tag{7}$$

An advantage of GPR is the ability to provide measurements about the model *epistemic uncertainty*, i.e., caused due to lack of data, while the term σ_n^2 presents the *aleatoric uncertainty*.

2.2.1 Covariance function

The covariance functions (also called kernels) express some form of distance or similarity, specifying how much two random variables change together. If the inputs x and x' are close to each other, it is expected that $f(x)$ and $f(x')$ will be close as well. For variables with inputs which are very close the covariance is almost unit, while the covariance decreases as the distance in the input space increases.

$$cov(f(x), f(x')) = k(x, x') \tag{8}$$

The covariance function can be defined by various kernel functions. In this work, the squared exponential covariance function is utilized which is given by:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right) \tag{9}$$

where σ_f^2 is the signal variance, that is a scale factor and determines the variation of the function from its mean values, $l > 0$ is the lengthscale and determines how reliable the function fits the training data. These are free parameters and together with the noise variance, σ_n^2, constitute the GPR hyperparameters; $\boldsymbol{\theta} = (l, \sigma_f, \sigma_n)$. In this case the kernel function utilizes separate length scale for each predictor, the Automatic Relevance Determination (ARD) squared exponential kernel is defined as

$$k(x, x') = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{n=1}^{d}\frac{(x_n - x'_n)^2}{l_n^2}\right] = \sigma_f^2 \exp\left(-\frac{1}{2}(x - x')^\top L^{-1}(x - x')\right) \tag{10}$$

where $L^{-1} = \text{diag}(l_n)^2$ and the dimension of the hyperparameters increases to $d + 1$; $\boldsymbol{\theta} = (l_1, l_2, ..., l_d, \sigma_f, \sigma_n)$.

2.2.2 Hyperparameter optimization

The accuracy of the GPR-based model significantly depends on the hyperparameters, $\boldsymbol{\theta}$, value. The hyperparameters are defined via an optimization procedure by maximizing the log marginal

likelihood, which is given by

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top [K(\mathbf{X}, \mathbf{X}) + \sigma_n^2]^{-1}\mathbf{y} - \frac{1}{2}\log[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2] - \frac{N}{2}\log 2\pi. \tag{11}$$

The derivatives of the log marginal likelihood with respect to $\boldsymbol{\theta}$ are computed and the hyperparameters are updated using gradient descend method.

2.3 Long-Short Term Memory Neural Networks

Recurrent Neural Networks (RNNs) are networks with loops allowing information to persist; connect previous information to the present task. The LSTM neural network is a type of RNN that models temporal sequences and is capable to learn long-term dependencies. The LSTM is designed to avoid the exploding/vanishing gradient descent problem that usually appears in the RNNs. The LSTM uses two separate paths to make predictions, i.e.; one path is for long-term memories and one for short-term memories. Unlike to a typical RNN, the LSTM is based on a much more complicated unit, see Figure 1. In addition, the LSTM uses sigmoid and hyperbolic tangent activation functions. A reminder for the reader, the sigmoid activation function takes any value and convert it to a number between 0 and 1, while the hyperbolic tangent activation function turns the value to a number between -1 to 1.

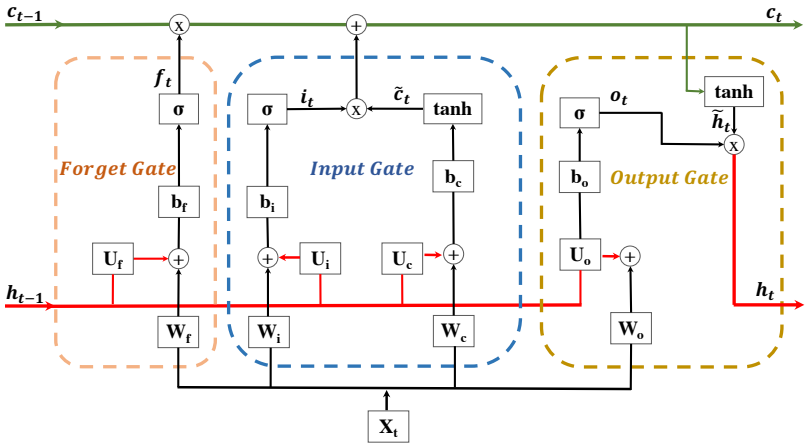


Figure 1: The structure inside an LSTM unit.

In Figure 1, the green line on the top of the LSTM unit is called the *cell state* and represents the *long-term memory*. Although the long-term memory can be modified by multiplication and addition, there are no weights and biases that can modify directly its value. This characteristic allows the long-term memory to flow through a series of units (Figure 2) without facing the issue of exploding/vanishing. The c_{t-1} is the long-term memory previous time step and the c_t is the new long-term memory for the current time step. The red line on the bottom of the LSTM unit is called the *hidden state* and represents the *short-term memory*. The h_{t-1} is the short-term memory from the previous time step and the h_t is the new short-term memory and it is the output of the LSTM cell. The short-term memory (or hidden state) is directly connected to weights that can modify its

value. The long- and short-term memories interact through the gates and result in predictions. First, in the forget gate, the short-term memory of the previous time step, h_{t-1} is combined with the input, x_t, using the weights and biases and the outcome is plugged in the sigmoid, σ, activation function. The sigmoid activation functions turns the value into a number between 0 and 1, f_t, which is then multiplied by the previous step long-term memory, c_{t-1}. In practice, the forget gate of an LSTM unit determines what percentage of the long-term memory is remembered. Second, the new long-term memory, c_t, is defined through the input gate. The right side of the input gate, that contains the hyperbolic tangent, *tanh*, activation function combines the short-term memory and the input using weights and biases to create a potential long-term memory, $\tilde{c}_t$. The right side of the input gate that contains the sigmoid, σ, activation function determines what percentage, i_t, of this potential long-term memory is added to the previous long-term memory, c_{t-1}. After this stage, the new long-term memory is defined, c_t. Third, the new short-term memory of the LSTM, h_t, is finally updated though the output gate. The new long-term memory, c_t, is used as input in the hyperbolic tangent, *tanh*, activation function which then provides the potential short-term memory, $\tilde{h}_t$. However, the left side of the output gate that contains the sigmoid, σ, activation function has to decide the percentage, o_t, of this potential short-term memory, $\tilde{h}_t$, to pass on. Similar logic has been previously used in the forget and input gates. The combination of the outcome of the two activation functions determine the new short-term memory, h_t, which is the output of the entire LSTM unit. Equations (12)-(18) show the mathematics at each gate and the definition of the new long- and short-memories.

$$\text{Forget gate:} \quad f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right) \tag{12}$$

$$\text{Input gate:} \quad i_t = \sigma\left(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i\right) \tag{13}$$

$$\tilde{c}_t = \tanh\left(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c\right) \tag{14}$$

$$\text{Cell state output:} \quad c_t = i_t * \tilde{c}_t + f_t * c_{t-1} \tag{15}$$

$$\text{Output gate:} \quad o_t = \sigma\left(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o\right) \tag{16}$$

$$\tilde{h}_t = \tanh\left(c_t\right) \tag{17}$$

$$\text{Hidden layer output:} \quad h_t = o_t * \tilde{h}_t \tag{18}$$

where W_f, W_i, W_o, W_c are the weight matrices of the input x_t, U_f, U_i, U_o, U_c are the weight matrices of the hidden state h_{t-1}, and b_f, b_i, b_o, b_c are the bias terms. The σ represents the sigmoid function and *tanh* represents the hyperbolic tangential function.

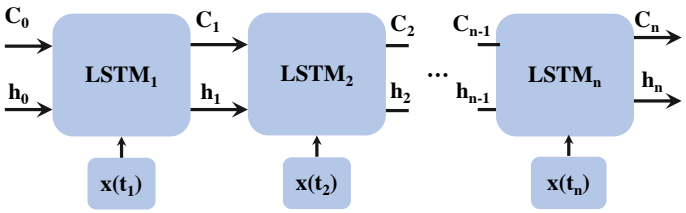


Figure 2: An LSTM layer that consists of n LSTM units.

The above description refers to a single LSTM unit that accepts as input the value for one time step. As shown in Figure 2, to make a prediction based on a sequence of inputs, $x(t_1, t_2, ..., t_n)$, a

sequence of LSTM units is created which constitutes the LSTM layer. In this case, each unit accepts the value assigned to a certain time step. In the current study, one LSTM layer is sufficient to predict the output. However, more layers can be used for more complex predictions but the model becomes difficult to train. The LSTM layer is then connected to a linear layer that takes the LSTM output - new short-term memory, h_t, - and calculates the final output, y_t, which is the actual mooring line force,

$$y_t = W \cdot h_t + b \tag{19}$$

where W is the weight matrix and b is the bias term. The LSTM model of this study is built using Keras which is a high-level, deep learning Application Programming Interface (API) for implementing neural networks written in Python. Keras is the high level API of TensorFlow platform which is used to build machine learning models.

The goal of the prediction is to identify the mooring line force time series based on the solution provided by GPR model. That is to say, the LSTM model accepts as input the mooring force which has been previously predicted from the GPR model and learns how to improve it in order to provide the actual mooring line force that matches the CFD results. The GPR-based mooring force is the only feature that the LSTM model considers for the prediction and thus the parameter *number of features* in Keras library is set equal to one.

2.3.1 Loss function

Deep learning neural networks (including LSTM) are usually trained using the *gradient descent* optimization algorithm. At the training process, the error of the model is repeatedly estimated through the loss function and the weights and biases of the model are constantly updated to reduce the loss function error in the next iteration.

At this stage, the LSTM model is trained to match the mooring force time series predicted by the GPR model to the real solution, e.g., from CFD simulations. The LSTM model is called to accurately predict the peaks (extremes) in the force time series, therefore, the *loss function* should be properly selected. The Mean Absolute Error (MAE) is an appropriate loss function as it is more robust to outliers. The MAE is calculated as the average of the absolute difference between the predicted and actual values. Specifically, in the LSTM training, the MAE loss function takes the form

$$\text{MAE} = \frac{1}{b}\frac{1}{t}\sum_{i=1}^{b}\sum_{j=1}^{t}|\hat{y}_j - y_j|_i \tag{20}$$

where $\hat{y}$ is the LSTM prediction and y is the CFD solution. As it is explained in section 4.2, for LSTM training purposes the datasets is split into b batches and each batch comprises of t time steps. The error between the predicted and the real output is computed for each time step, j, and the batch-average error is calculated. Finally, the *overall loss function* MAE is the average of the batch-average errors.

2.4 Dimension Reduction

Although supervised machine learning methods for time series forecast have been recently become very popular, e.g. LSTM, many of the current methods, e.g. GPR, model the relationships between the *finite vector* dependent variable (output) and one or more explanatory values (inputs). In this study, the output is a *time series*, not a vector quantity. In order to fit the output into the

mapping framework of the GPR method, the time series is represented as a low dimensional vector by employing a *dimension reduction* technique. Specifically, for the mooring force time series, $y(t)$, the following reduced order representation is employed

$$y(t|\boldsymbol{x}) = \sum_{i=1}^{n} q_i(\boldsymbol{x})\hat{\mu}_{i,T}(t), \quad t \in [0, T]. \tag{21}$$

where $q_i(\boldsymbol{x})$ are the reduced order coefficients which are functions of input vector $\boldsymbol{x}$, $\hat{\mu}_{i,T}(t)$ are the eigenvectors of the force matrix, and n is the number of output modes that describe the force. For this problem, the output coefficients are computed by applying Principle Component Analysis (PCA) to the available training data from CFD simulations, i.e., a set of mooring force time series associated with particular sea states and WEC structural characteristics. Further explanation about the choice of output modes, n, is provided in section 4.1.

As shown in Figure 5 (Right), a hybrid surrogate model is developed by coupling two machine learning techniques into a hybrid surrogate model; the GPR which provides the input-output map of scalar quantities and the LSTM neural network that is able to handle time series prediction. As it is proven later, the latter method complements the former for achieving better prediction accuracy. The GPR supervised machine learning method provides the map function for each coefficient, q_i, to the input vector, $\boldsymbol{x}$. Further explanation can be found in section 3.3 and section 4.

2.4.1 Principal component analysis

Many techniques address the problem of reducing the dimensionality of time series data, and differ based on the assumptions they make on the underlying data generating process. For this problem, where the mooring force time series is smooth over a finite interval, we use PCA, sometimes referred to as Karhunen-Loève (KL) decomposition or proper orthogonal decomposition (POD) [11].

The first pre-processing step is to project the numerical mooring force sequence data obtained from the CFD simulations, $y(t)$, onto a fixed grid with p equally spaced time instants via *interpolation*. In this case, the force sequence, $y(t)$, consists of p points where point $a_{l,j}$ is the mooring force for the l^{th} examined data series out of the m datasets at j^{th} time instant. This step is especially important because many CFD solvers internally use variable time steps following the Courant–Friedrichs–Lewy condition [7]. To avoid interpolation issues, the projection time step should be smaller than the typical smallest CFD step. However, too small a time step, will slightly complicate LSTM neural network training, but this is discussed in section 4.2.

Once every data series has been projected onto a common grid trough interpolation, each vector is assembled into a $m \times p$ data matrix A [49]. PCA is a statistical procedure that maps the p–dimensional features (which are probably correlated) into a set of values of linearly uncorrelated variables (principal components). These components are ordered by the fractions of the total signal variance they contain. By keeping only the most important components, and dropping the less significant terms, the dimension of the signal may be reduced without significant loss of information.

The covariance matrix $C = A^{\intercal}A$ may be expressed as a product of three matrices via singular value decomposition (SVD) as

$$C = USV^{\intercal} \tag{22}$$

where U is an $m \times m$ unitary matrix, S is an $m \times p$ diagonal matrix, and V is an $p \times p$ unitary matrix. The columns of U are the eigenvectors $\mathbf{u}_l$ of $AA^{\intercal}$ and the columns of $V^{\intercal}$ are the eigenvectors $\mathbf{v}_j$ of $A^{\intercal}A$. The diagonal matrix S has r elements equal to the root square of the eigenvalues of $AA^{\intercal}$

or $A^\intercal A$ (both have the same eigenvalues), s_i. Further, it is assumed that the paired eigenvectors and eigenvalues are ordered in descending value of s_i, which is always possible by inserting a unitary (rotation) matrix R with the expression $UR^\intercal RSR^\intercal RV^\intercal$. Finally, because the covariance matrix is positive semi-definite, the eigenvalues s_i are real and non-negative.

The PCA transform is the linear projection of the original mooring force data (interpolated onto the common grid) onto the orthonormal eigenbasis defined by the $\mathbf{v}_j$ vectors. It has the property that the first PCA mode contains the maximum fraction of energy (in the L_2 sense) among all possible linear projections, and the second mode has maximum energy among basis vectors orthogonal to the first mode, and so on. In particular, the energy component of each mode i is given by the eigenvalue s_i, and the total energy of the signal is given by $\sum_{i=1}^{r} s_i$.

For a data series that consists of points $a_{l,j}$, we project onto the PCA basis using the relation

$$\tilde{q}_i = \sum_{j=1}^{p} a_{l,j} v_{j,i}$$

where $v_{j,i}$ is the point at the column vector of matrix $V^\intercal$. However, in order to normalize our data with the eigenvalues s_i, we instead use the modified projection

$$q_i = \sum_{j=1}^{p} a_{l,j} \frac{1}{\sqrt{s_i}} v_{j,i}, \tag{23}$$

with with the q_i are unitless and order one. We emphasize that dimensionality reduction using PCA is not “physics aware" and is entirely data driven. If a different data set were considered, simulated with a different set of focused waves, the recovered PCA modes will vary somewhat. However, previous work has found that PCA works well for ML applications involving structural responses to wave forcing [14].

2.5 Active Learning

The question that arises at this point is *how to select the samples* for training the surrogate model in order to (1) provide accurate predictions – especially for extreme events – (2) while considering the difficulty to obtain the training samples. In conservative sampling approaches, the physical system (i.e. black-box function) is queried at a large number of input points, however, this approach is time and resources prohibitive as it translates to massive high-fidelity simulations and/or physical experiments.

This section presents a methodological framework that combines the *active learning* scheme in BED with machine learning techniques to effectively construct a surrogate model. As shown in Figure 3, the surrogate is trained on sequentially selected datasets of input–output pairs, $\mathcal{D}_t = \{D_{t-1} \cup (x_t, y_t)\}$, where t is the number of sampling iterations. That is to say, the next-best sample is determined through a procedure that actively discovers and sequentially selects the most informative training samples from the input parameter space, $\mathcal{X}$. These samples maximally reduce the surrogate model's uncertainty while simultaneously maximizing the information about extremes. This preference, that each data point is maximally informative, directly addresses the main practical limitation of obtaining large datasets from CFD simulations.

Algorithm 1 formalizes the iterative steps for efficiently training the surrogate model following the active learning technique. The constructed surrogate is a computational cheaper model and allows for massive samples evaluation, therefore, it can be finally used to forecast and quantify the extreme event statistics in the dynamic system.

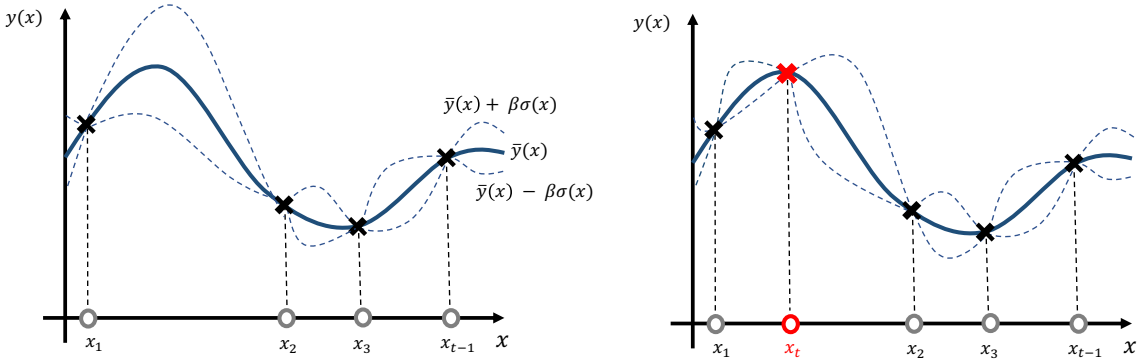


Figure 3: Estimation of the output mean value, $\bar{y}(x)$, for any input, x, from the developed surrogate model. **Left**: The surrogate model is constructed based on the dataset $\mathcal{D}_{t-1} = \{(x_1, y_1), (x_2, y_2), ..., (x_{t-1}, y_{t-1})\}$. **Right:** The active learning sampling chooses a new optimal sample, x_t, that reduces the surrogate model's uncertainty and provides accurate prediction about extremes.

Algorithm 1 Sequential search for active training of surrogate model.

Input: Number of iterations, t_{iter}
Initialize: Train the surrogate model on initial dataset of input-output pairs, $\mathcal{D}_0 = \{x_i, y_i\}_{i=1}^{t_{init}}$
for $t = 1$ **to** t_{iter} **do**
    Select next sample x_t by maximizing acquisition function $\alpha(x)$:

$$x_t = \arg\max_{x \in \mathcal{X}} \left(\alpha(x; y),\, \mathcal{D}_{t-1}\right)$$

    Evaluate black-box function (real system) at x_t and record y_t
    Augment dataset: $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{x_t,\, y_t\}$
    Retrain surrogate model
end for
return Final surrogate model

2.5.1 Acquisition functions

The acquisition function is the fundamental component of the active learning algorithm as it guides the exploration and exploitation of the input parameters space, $\mathcal{X}$, and determines what is the 'next-best' sample to query the black-box function. The choice of acquisition function is critical as different acquisition functions are appropriate for different features of the quantity of interest. In this study, two acquisition functions are examined; the uncertainty sampling (US) and the likelihood weighted uncertainty sampling (LW-US).

Uncertainty sampling: The US is one of the most common acquisition functions and identifies the 'next-best' sample for which the GPR model presents the highest predictive variance,

$$\alpha_{US}(x) = \sigma^2(x). \tag{24}$$

The US ensures that the uncertainty of the model is evenly distributed over the input space and its popularity lays on its straightforward implementation, inexpensive evaluation, analytic gradients that allow the use of gradient-based optimizers. However, the US acquisition function has usually the tendency to select points on the boundaries of the input space because the variance is often greater in regions where less data have been previously collected [3]. This does not necessarily mean that the close-to-boundaries region provides the most informative samples.

Likelihood-Weighted Uncertainty Sampling: The LW-US acquisition function drives the sample selection towards regions where the model prediction has high uncertainty yet extreme events have high likelihood to occur [3,42]. In practice, the LW-US is the US acquisition function multiplied by the likelihood-ratio, $w(x)$,

$$\alpha_{US-LW}(x) = \sigma^2(x)w(x). \tag{25}$$

The likelihood-ratio, $w(x)$, quantifies the importance of the output relative to the input,

$$w(x) = \frac{p_x(x)}{p_{\bar{y}_\mathcal{G}}(\bar{y}_\mathcal{G}(x))}, \tag{26}$$

where $p_{\bar{y}_\mathcal{G}}$ denotes the PDF of the GPR posterior mean, $\bar{y}_\mathcal{G}$, and p_x is the PDF of the input parameters, x. The p_x is usually referred to as *nominal distribution* and $p_{\bar{y}_\mathcal{G}}$ *importance distribution*. The likelihood ratio acts as a sampling weight that balances events that are probable to occur and those that are extreme. For instance, for data in the input space that have similar probability of being observed (same p_x), the likelihood ratio assigns more weight to those that are likely to relate with rare/extreme events (small $p_{\bar{y}_\mathcal{G}}$). Conversely, for samples that have similar probability for extreme/rare events (same $p_{\bar{y}_\mathcal{G}}$), it promotes those with higher probability of occurrence (larger p_x).

The importance distribution, $p_{\bar{y}_\mathcal{G}}$, is estimated at every iteration in Algorithm 1. A large number of data, $\{x_k\}_{k=1}^M$, are chosen from the input space via Latin hypercube sampling (LHS). These samples are evaluated through the GPR model and the posterior mean, $\bar{y}_\mathcal{G}$, is estimated for every input. The PDF $p_{\bar{y}_\mathcal{G}}$ is constructed via the kernel density estimator (KDE). Algorithm 2 describes precisely the steps for the definition of the importance distribution $p_{\bar{y}_\mathcal{G}}$.

3 Wave Energy System in Extreme Sea States

This study focuses on developing a hybrid surrogate model that estimates the force exerted to the mooring line of a wave energy converter. This estimation takes into account the varying (i)

Algorithm 2 Computation of the importance distribution, $p_{\bar{y}_{\mathcal{G}}}$.

1: **Input:** GPR model and nominal distribution, $p_x(x)$
2: **for do**
3: Draw a large number of samples, $\{x_k\}_{k=1}^{M}$ via LHS
4: Evaluate the samples through the GPR model to obtain the posterior mean values $\{\bar{y}_{\mathcal{G}_k}\}_{k=1}^{M}$
5: Perform KDE on the outputs, $\{\bar{y}_{\mathcal{G}_k}\}_{k=1}^{M}$, and obtain the PDF $p_{\bar{y}_{\mathcal{G}}}$
6: For a particular input-output pair $(x, \bar{y}_{\mathcal{G}}(x))$, the corresponding density value, $p_{\bar{y}_{\mathcal{G}}}(\bar{y}_{\mathcal{G}}(x))$, is defined by the PDF constructed at the previous step.
7: **end for**
8: **return** Importance distribution $p_{\bar{y}_{\mathcal{G}}}(\bar{y}_{\mathcal{G}}(x))$

site-specific extreme wave conditions and (ii) parameters related to the Power Take-Off (PTO) system. To provide context, as referenced in prior studies [19, 20, 22, 52], the WEC is designed to operate at the Dowsing site in the North Sea. The wave conditions selected for analysis are drawn from the 50-year return period environmental contour plot corresponding to this specific location. To construct the hybrid surrogate model, the GPR and LSTM machine learning techniques are implemented. The GPR component of the model maps the 50-year return period waves and PTO parameters to the corresponding mooring force, following a similar procedure as described in related works [14, 25]. Section 4 provides a detailed explanation of this process. The LSTM component complements the GPR model, enhancing the accuracy of predicting the mooring force time series.

In this study, two approaches are employed to train the surrogate model. First, existing data from Computational Fluid Dynamics (CFD) simulations are used, representing a traditional statistical learning technique without active learning. Second, the active learning scheme in Bayesian experimental design is applied to construct the hybrid surrogate model. In the latter approach, the training samples are not predetermined but are sequentially selected during the active learning iterations. The samples obtained from previous iterations inform the decision for selecting the most informative sample to quantify extreme forces. Finally, in both training approaches, the surrogate model is utilized to reconstruct the statistics of extreme mooring forces. This enables a comprehensive understanding of the extreme force characteristics in the system.

3.1 Extreme wave representation

As previously mentioned, the waves are carefully chosen from the 50-year environmental contour, and each wave is characterized by its significant wave height (H_s) and peak wave period (T_p). The numerical wave realization, typically represented as a 30-minute to 3-hour irregular wave episode, demands substantial computational resources in the context of CFD. Instead, as detailed in [21], an alternative approach involving equivalent design-wave methods is typically employed.

In this study, each 50-year wave is recreated as a *focused wave group* defined using the NewWave theory [47]. A focused wave is essentially the combination of sinusoidal wave components, all converging at a predetermined location and time, maximizing the resulting wave's amplitude (Figure 4). One significant advantage of the NewWave theory is its ability to quickly identify extreme wave episodes within a short time frame compared to the potential occurrence of such events throughout the entire sea state. This makes the focused wave groups exceptionally suitable for the computationally intensive CFD simulations.

The process of transforming the (H_s, T_p) pair into the focused wave realization unfolds as follows:

By employing the NewWave theory and considering the JONSWAP spectrum, the maximum focused wave amplitude and individual wave components amplitude are determined. These components are designed to converge at a specific preselected location and time, thereby defining the phase of each component. In conclusion, starting with the (H_s, T_p) pair, a focused wave profile is constructed. This profile is then simulated in its interaction with the WEC system using CFD, ultimately allowing us to record the mooring line force.

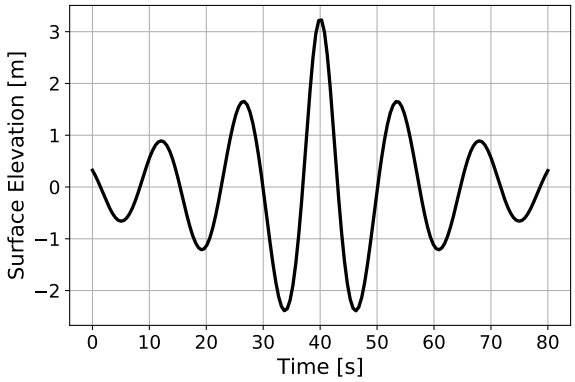


Figure 4: "Focused Wave Profile: Wave components focusing at t = 40 s with their amplitudes determined by NewWave Theory."

3.2 Wave energy converter

Figure 5 (Left) illustrates the schematic representation of the point-absorber wave energy system [27] investigated in this study. The system comprises two main components: a buoy and a direct-driven linear generator housed in a capsule on the seabed, referred to as the power take-off (PTO). These components are connected by a mooring line made of steel wire rope. The translator within the generator undergoes reciprocal motion, following the heaving motion of the buoy, which induces a varying magnetic flux in the stationary stator windings. To prevent excessive displacement, the translator has a limited stroke length that is reached when the buoy experiences large wave heights. The translator's motion is damped by internal upper and lower end-stop springs, ensuring that it does not collide with the top and bottom of the generator hull. During high waves, the end-stop spring acts as a restoring force and its magnitude depends on the stiffness of the spring, denoted as K_{es}. Another crucial parameter in the PTO is the damping coefficient, D_{PTO}, which governs the motion of the wave energy converter (WEC) and the conversion of wave power. WEC control strategies aim to maximize wave power extraction by adjusting the PTO damping coefficient in response to stochastic environmental conditions.

3.3 Experimental Design

The wave energy system operates under extreme sea states that are specifically chosen from the 50-year environmental contour for Dowsing location in the North Sea [52]. The primary objective of the surrogate model is to accurately predict the force exerted on the mooring line, which is a critical component of the system. Several factors contribute to the mooring force, including the

significant wave height (H_s) and peak wave period (T_p) characterizing the wave, the PTO end-stop spring stiffness (K_{es}), and damping coefficient (D_{PTO}). Each of these input variables, along with their corresponding units and ranges, is summarized in Table 1.

Table 1: Input variables for wave energy system surrogate model.

Symbol	Description	Units	Bounds
H_s	Significant wave height	m	[5.0, 7.8]
T_p	Peak wave period	s	[8.3, 14.0]
K_{es}	PTO end-stop spring stiffness	kN/m	[585, 960]
D_{PTO}	PTO damping coefficient	kNs/m	[48, 85.5]

3.4 Computational Fluid Dynamics

High-fidelity computational fluid dynamics (CFD) simulations are essential to capture relevant hydrodynamic nonlinearities when simulating the interaction of the WEC with the extreme wave episodes. The open-source software OpenFOAM is employed in this study.

The incompressible Reynolds Averaged Navier-Stokes (RANS) equations are numerically solved using the cell centered finite volume method [10]. The turbulence is included using the $k-\omega$ SST turbulence model that utilizes the wall functions for capturing the turbulence effects on the WEC boundaries. The model uses the `overInterDyMFoam` solver to iteratively solve the RANS equations using the PIMPLE algorithm. The Volume of fluid (VoF) method is employed to capture the free water surface. The `overInterDyMFoam` solver provides the overset mesh functionality which allows complex mesh motions and interactions without the penalties associated with deforming meshes [18]. The wave excitation and the external forces define the WEC motion which is calculated through the library `sixDoFRigidBodyMotion`. The wave generation and absorption is implemented via the IHFOAM toolbox. Detailed description of the CFD model can be found in [18, 19, 22]. The setup of the numerical domain has been validated by comparing the WEC motion and loads under extreme wave conditions with measurements from physical experiments [22].

4 Hybrid Surrogate Model

The first part of this study evaluates the ability of machine learning technology to predict complex outputs in the form of time series. As illustrated by Figure 5 (Right), the GPR and LSTM methods are combined for building a hybrid surrogate model for the wave energy system approximating the relationship between the input vector $\boldsymbol{x} = [H_s, T_p, D_{PTO}, K_{es}]$ and the corresponding mooring force $y(t)$. The surrogate model is trained on CFD data – existing for the purposes of a previous study. Specifically, 73 datasets are totally available, with the 65 of them used for training the model and the rest kept for its validation.

The CFD simulations produce the mooring force as a *time series* record. However, as explained in section 2.2, the GPR method provides the mapping between scalar quantities, therefore, it does not predict time series directly. For that purpose, the *dimension reduction* presented in section 2.4 is employed to express the mooring force as a finite vector of coefficients, q_i. The dimensionality of this vector is defined by the PCA truncation order n, which is an important parameter balancing the

accurate force reconstruction against computational cost. The effects of varying number of retained PCA coefficients is discussed in section 4.1. For this work, $n = 12$.

To summarize, OpenFOAM generates the mooring force time series corresponding to a set of input variables. The dimension reduction step follows in order to approximate the force time series as a set of 12 PCA coefficients. In total, 12 GPR surrogate models are trained to provide the mapping between the input vector $\boldsymbol{x}$ and each coefficient q_i. For a particular $\boldsymbol{x}$, the GPR posterior mean $\bar{q}_i = \mathbb{E}[q_i|\boldsymbol{x}]$ is estimated for each of the 12 coefficients which are then combined using Equation (21) and reconstruct the force time series, $\bar{y}_{\mathcal{G}}(t)$. This is the first prediction step making use only the GPR component of the hybrid model.

Next, the LSTM component - which is able to handle directly time series - receives as input the GPR reconstructed mooring force $\bar{y}_{\mathcal{G}}(t)$ and provides as output the corrected force $y_{\mathcal{L}}(t)$. The LSTM model implements further correction providing a better approximation of the mooring force by capturing the peaks with higher accuracy. Given that this study focuses on the extremes, it is very important to properly estimate the peak forces. Figure 5 (Right) illustrates the mapping stages of the hybrid surrogate model as they have been described in this section.

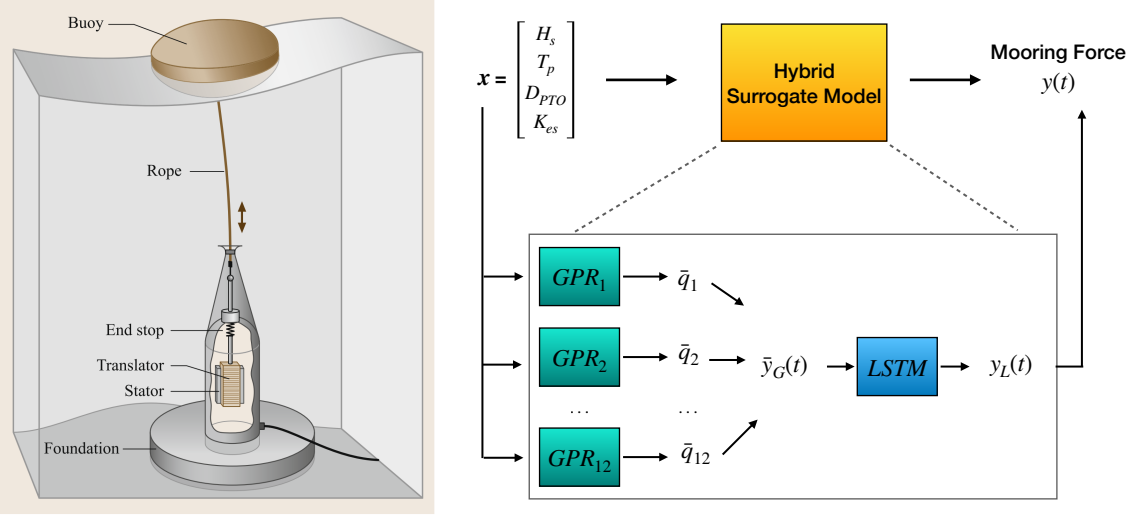


Figure 5: (**Left**): A schematic depiction of the point-absorber wave energy system [9]. (**Right**): The hybrid surrogate model maps the input–output relationship, with the input $\boldsymbol{x}$ representing the sea state characteristics, and structural variables of the system and the output $y(t)$ is the force in the mooring line. The model is built using the GPR and LSTM machine learning techniques. Due to the dimension reduction of the force time series, twelve GPR surrogate models are created–each for a PCA coefficient, $q_i(\boldsymbol{x})$. The GPR model maps the input, $\boldsymbol{x}$, and the output, $\{q_i(\boldsymbol{x})\}_{i=1}^{12}$. The GPR posterior mean, $\{\bar{q}_i(\boldsymbol{x})\}_{i=1}^{12}$, are used for the force reconstruction, $\bar{y}_{\mathcal{G}}(t)$, which constitutes the input for the LSTM model. The latter's role is to provide a more accurate prediction, $y_{\mathcal{L}}(t)$.

4.1 Number of retained PCA coefficients

The number of most informative PCA coefficients is identified using the ratio:

$$\text{Cumulative variance} = \frac{\sum_{i=1}^{n} s_i}{\sum_{j=1}^{r} s_j} \geq 0.99, \tag{27}$$

where the r is the number of non-zero eigenvalues of the force matrix and n is the number of retained PCA modes - see section 2.4. Figure 6 shows the cumulative variance with the number of PCA modes which starts to converge for $n > 10$. In this work, $n = 12$ modes are retained. This step is important because the number of retained PCA coefficients defines the computational cost by determing how many GPR models will be trained later.

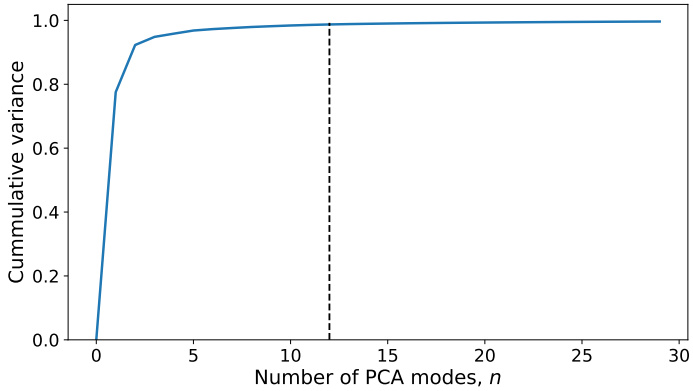


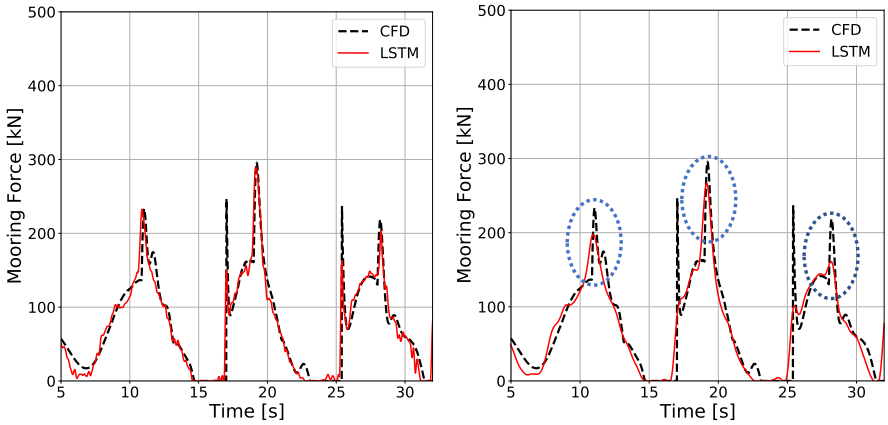
Figure 6: Cumulative retained energy of PCA truncation order n, based on the eigenspectrum of the mooring force dataset. In this work, $n = 12$ is chosen to retain $> 99\%$ of the signal energy.

4.2 LSTM hyper-parameters selection

To create the optimal LSTM architecture, it is necessary to select the values for several hyper-parameters, i.e., time series resolution, time-window, number of hidden units, batch size, number of epochs. In this study, a procedure similar to [50] is followed for the optimal hyper-parameter definition based on two criteria; (1) prediction accuracy and (2) time for training the LSTM model.

The *time series resolution* is initially defined and this step is followed by the definition of the *time-window* length. The predicted time series for several time-window lengths are compared to the CFD solution using the Mean Average Percentage Error (MAPE) metric, and based on that metric the time-window length is determined. Next, the number of the *hidden-units* is defined again through MAPE evaluation while the hyper-parameter *batch size* is selected by evaluating the loss function. The procedure for hyper-parameters selection is further described below:

1. Time resolution: The raw data from the CFD simulations are sampled in 0.01 s time intervals - following an interpolation procedure (see section 2.4.1). To accelerate the LSTM training process, it is examined the option to provide coarser time series. However, a critical point is that LSTM model should be able to capture the extreme forces, which are the instantaneous peaks



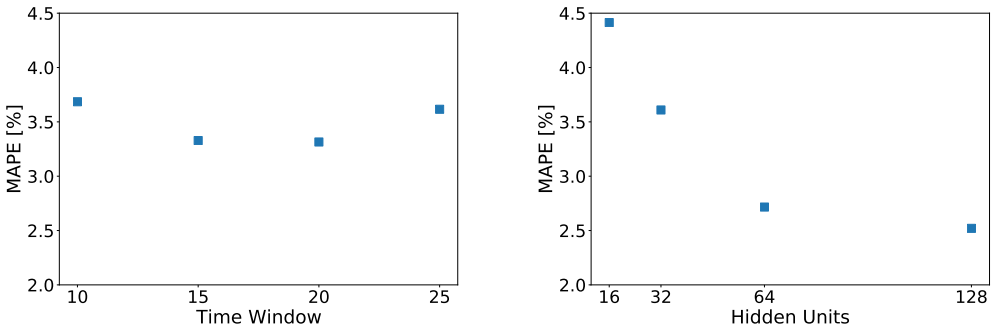
(a) Resolution 0.02 s (b) Resolution 0.04 s

Figure 7: The LSTM hyper-parameter time resolution is evaluated; (a) 0.02 s and (b) 0.04 s. The LSTM solution (red) is compared to the CFD solution (dashed black). For lower resolution (0.04 s) the prediction accuracy in the peaks drops - as indicated by the blue circles - leading to the selection of 0.02 s for the LSTM model training.

in the time series. Therefore, the time resolution is a sensitive parameter and should satisfy that the instantaneous peak forces are not vanished in the prediction. In this study, two resolutions are examined, i.e., dt = 0.02 s, 0.04 s. In Figure 7, the LSTM prediction is compared to the CFD solution for both resolutions. It is observed that the coarser resolution (dt = 0.04 s) underestimates the peaks while the finer (dt = 0.02 s) provides high accuracy, and thus it is prefered.

2. Time-window: The total training data is divided into shorter sequences which are called time-windows. Each sub-sequence consists of the number of time-steps that the LSTM model considers for predicting the next time-step value. In this stage, it is important to characterise the time-window length, i.e., how many time-steps per sub-sequence, as it affects the training time and prediction accuracy of the LSTM model. As shown in Figure 8(a), the time-window is selected after examining the LSTM prediction accuracy for several lengths (i.e., 10, 15, 20, 25 time steps). It is observed that the MAPE (< 4 %) does not vary significantly with the time-window length. As previously mentioned, one of the criteria to choose the hyper-parameters is the time for training the LSTM model. Therefore, the time-window of 20 time steps is selected because the training time is accelerated while the prediction accuracy is preserved.

3. Hidden units: The number of the hidden units is the size of the hidden nodes in a LSTM layer and has a great impact on the LSTM prediction accuracy. The hidden units determine the dimensions of the weight matrices and bias vectors discussed in section 2.3. For example, for an input vector $\boldsymbol{x}$ with dimensions $d\times 1$, and hidden units, h, the weight matrices of the input, W, have dimensions $d \times d$, the weight matrices of the hidden state, U, have dimensions, $h \times h$, and the bias vectors have dimension $h\times 1$. The influence of the hidden units is studied and the prediction error under different numbers of hidden units (32, 64, 128, 256) are compared in Figure 8(b). Finally, 64 hidden units are selected for the LSTM model because less weights and bias coefficients need to be trained–compared to the case with 128 units–and thus the computational cost for training the



(a) Time Window (b) Hidden Units

Figure 8: Selecting hyper-parameters for the LSTM model.

LSTM is reduced. In addition, more hidden units could result in overfitting.

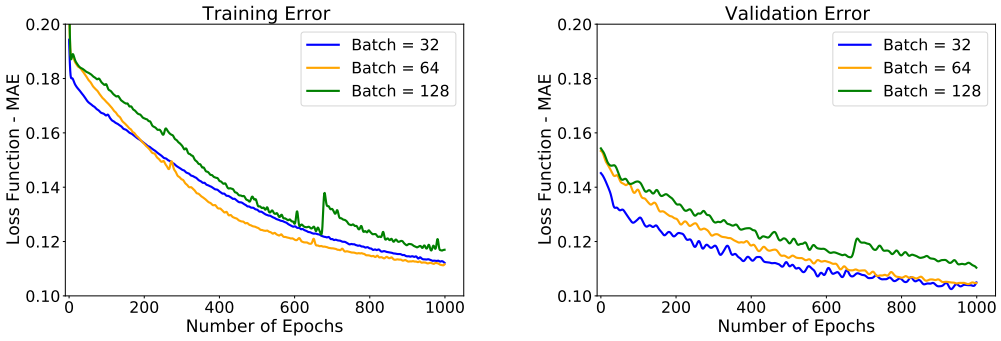
4. Batch size: The batch size is a hyperparameter of gradient descent algorithm that defines the number of training samples the LSTM processes before updating its internal parameters (weights and biases). A common approach is to take the batch size equal to 32, or 64 or 128 samples. In this study, the batch size is evaluated through the loss function (see section 2.3.1). Figure 9 shows that the loss function presents better convergence for batch size equal to 32 and 64 - both for validation and training samples. However, the LSTM training process is accelerated for bigger batch size, therefore, size equal to 64 is chosen in the this study.

5. Epochs: The number of epochs is a hyper-parameter that controls the number of complete passes through the total training dataset. Along with the learning rate, the choice of training epochs balances between overfitting and underfitting the LSTM network. In this study, the LSTM model is trained for 1000 epochs.

6. Learning rate: The learning rate is an important hyperparameter that controls how much the weights and bias are updated during the training, and consequently affects the training time. In this study the Adam gradient descent optimization algorithm is utilized since it provides computational efficiency, straightforward implementation, little memory requirements - among other benefits [26]. Adam uses adaptive learning rate for faster convergence and takes the initial value of 0.001.

Table 2: Summary of optimal LSTM hyperparameters from section 4.2.

Original Sampling Rate	$0.01s$
Time Resolution	$0.02s$
Time-Window	20 steps
Hidden Units	64
Batch Size	64
Training Epochs	1000
Initial Learning Rate	0.001



(a) Loss function error in training data

(b) Loss function error in validation data

Figure 9: Selecting batch size hyper-parameter for the LSTM model based on the loss function mean absolute error (MAE).

5 Hybrid Surrogate Model via Active Learning

Figure 10 provides a schematic depiction of the active learning strategy presented in section 2.5, which is adapted for the construction of the hybrid surrogate model.

1. The GPR model is pre-train on a few randomly selected samples $\mathcal{D}_0 = \{\boldsymbol{x}_i, y_i\}_{i=1}^{t_{init}}$, where $t_{init} = 5$, which are obtained from the real system (ground truth).

2. The GPR model is further trained using the Bayesian experimental design method that employs the active learning scheme for effective sample selection. During this step, the existing GPR model evaluates random input samples, $\{\boldsymbol{x}_i\}_{i=1}^{s}$, and based on the acquisition function the next-best input sample $\boldsymbol{x}^*$ is selected satisfying a specific condition (e.g. identify areas in the input parameter space where extremes are more probable to occur). The corresponding value, y^*, is then evaluated through the real system. The new pair $\{\boldsymbol{x}^*, y^*\}$ is added in the existing training dataset and the GPR model is retrained. This is a loop that continues for a number of user-defined iterations, $t_{iter} = 80$. Finally, the GPR model is retrained on the dataset $\mathcal{D}_{\mathcal{G}} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{t_{iter}}$, that consists of the pairs selected via active learning.

3. Now, the LSTM model is trained on the dataset $\mathcal{D}_{\mathcal{L}} = \mathcal{D}_0 \cup \mathcal{D}_{\mathcal{G}}$. It is important to note that the LSTM receives as input the GPR posterior mean force, $\bar{y}_{\mathcal{G},i}(t)$, and output the real system force $y_i(t)$.

4. The resulting hybrid surrogate model is able to predict the force in the mooring line for any other input vector, $\boldsymbol{x}$ – in orders of magnitude faster than the classical CFD simulations.

5. In addition, the hybrid surrogate is used for the reconstruction of the force statistics (i.e. PDF).

The selection of appropriate acquisition functions is important for an effective surrogate model, thus the acquisition functions discussed in section 2.5.1 are evaluated in this study. To quantify the uncertainty of the developed surrogate model, 15 Bayesian experiments run independently each

other with each differing in the choice of initial training samples $\mathcal{D}_0$ and consequently the dataset $\mathcal{D}_\mathcal{G}$.

5.1 Computation of the acquisition function

The active learning strategy aims to reduce the amount of necessary training data to built a reliable surrogate model. An important step is the sequential sample selection at each iteration of the active search loop and the acquisition function is the criterion for this selection. At this step, a large number of input samples is evaluated through the GPR surrogate model, and the sample which satisfies the acquisition function condition (see section 2.5.1) is selected as the next-best sample.

At each iteration, acquisition samples are generated via Latin Hypercube Sampling (LHS) from the input domain $x_{\mathcal{LHS}} = \{x_1, ...x_{100}\}$, with $n_{\mathcal{LHS}} = 100$ samples. Next, the $x_{\mathcal{LHS}}$ is fed to the current iteration GPR model in order to calculate the corresponding set of posterior means, $\overline{y}_G(x_{\mathcal{LHS}})$, and variances, $\sigma^2(x_{\mathcal{LHS}})$. At the same time, each LHS sample $\boldsymbol{x}_i$ has a corresponding probability $p_X(\boldsymbol{x}_i)$, where $p_X(\cdot)$ is the input distribution constructed in section 5.2. Finally, the probability density of $\bar{y}_\mathcal{G}$, $p_{\bar{y}_\mathcal{G}}(y)$ is estimated using the kernel density estimator (KDE) technique,

$$p_{\bar{y}_\mathcal{G}}(y) = \text{KDE}(\text{data} = \overline{y}_G(x_{\mathcal{LHS}}), \text{weight} = p_x(x_{\mathcal{LHS}})). \tag{28}$$

Equation 28 is used to compute the weight function $w(x)$ in Equation 25 by evaluating the PDF $p_{\bar{y}_\mathcal{G}}(y)$ at $y = \overline{y}_G(x_{\mathcal{LHS}})$, and then taking the ratio with $p_X(x_{\mathcal{LHS}})$. Finally, the next-best point $x^* \in x_{\mathcal{LHS}}$ is the one that satisfies

$$x^* = \underset{x^* \in x_{\mathcal{LHS}}}{\text{argmax}}\, a_{US-LW}(x). \tag{29}$$

5.2 Distribution of the input parameters

The nominal distribution, p_x, is defined for each sample. In this study, the input space is characterised by a 4-dimensional input vector, $\boldsymbol{x} = [H_s,\, T_p,\, D_{PTO},\, K_{es}]$. As suggested by [15], the significant wave height, H_s, follows a *Weibull* distribution,

$$f_{H_s}(h) = \frac{\beta}{\rho}\left(\frac{h}{\rho}\right)^{\beta-1} \exp\left\{-\left(\frac{h}{\rho}\right)^{\beta}\right\}, \text{for h}>\eta \tag{30}$$

where β is the shape parameter and ρ is the scale parameter. The wave period, T_p, follows a *lognormal* distribution conditional to significant wave height, H_s

$$f_{T_p|H_s}(t|h) = \frac{1}{\sqrt{2\pi}\sigma(h)t} \exp -\frac{(\ln t - \mu(h))^2}{2\sigma(h)^2} \tag{31}$$

where parameters $\mu(h)$ and $\sigma(h)$ are given by

$$\mu(h) = a_1 + a_2 h^{a_3} \tag{32}$$

$$\sigma(h)^2 = b_1 + b_2 \exp\{-b_3 h\} \tag{33}$$

The values for all the parameters (β, ρ, η, a_1, a_2, a_3, b_1, b_2, b_3) obtained from measurements in North Sea [15] and summarized in Table 3. Last, the PTO damping, D_{PTO}, and upper end-stop spring stiffness, K_{es}, are taken to follow a uniform distribution due to their nature as design parameters.

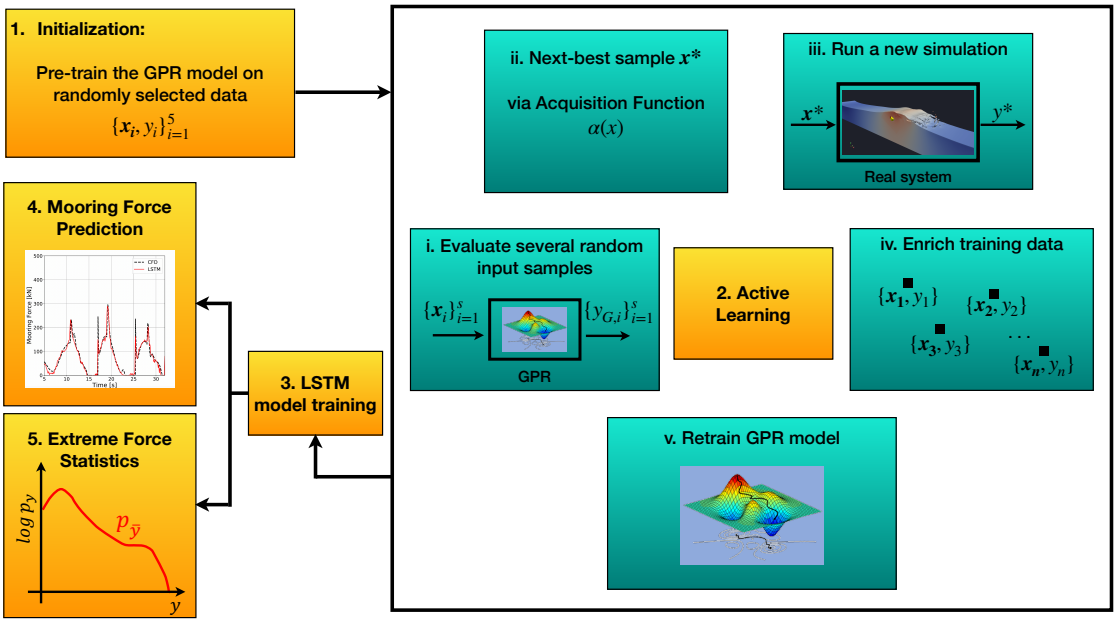


Figure 10: 1. Initialization: The GPR surrogate model is pre-trained on a few random samples and learns a sparse representation of the underlying system. 2. Active learning strategy: (i) The current GPR model evaluates a big number of Latin Hypercube samples for discovering the regions of the parameter space that lead to extreme events. (ii) The acquisition function identifies the next-best sample that reduces the model's uncertainty (exploration) and provides the best information about extremes (exploitation). (iii) The next-best sample is evaluated through a process that provides the real solution. This process can be a physical experiment, a numerical simulation or another model that the user trusts. (iv) The next-best input sample and the corresponding output are added in the existing training dataset. (v) The GPR model is retrained on the updated dataset. The steps (i)-(v) are repeated until either the solution converges or for a user-defined number of iterations – a limit that depends on the available resources. 3. The LSTM neural network is trained on the dataset obtained from the active learning loop plus the initial samples. The LSTM and GPR are coupled into a hybrid surrogate model that is used in the next steps. 4. The hybrid model is able to provide the accurate prediction of the force times series for any input vector. 5. To quantify the force statistics, thousands of Monte Carlo samples are evaluated through the surrogate model and the PDF is constructed on the corresponding outputs.

Finally, for the simulation designs, the vector $\boldsymbol{x}$ is constraint to live in a four dimensional hyperbox, with bounds given by Table 1. For the wave parameters H_s and T_p, the hyperbox bounds are chosen to include most of the probability mass associated with the unbounded distributions described in this section. For the PTO parameters D_{PTO} and K_{es}, the hyperbox bounds are chosen to include 'reasonable' engineering values.

Table 3: Parameters of the Weibull and lognormal distributions that describe the sea state characteristics. The values are determined from measurements in the North Sea [15].

Parameters	β	ρ	η	a_1	a_2	a_3	b_1	b_2	b_3
	1.550	2.908	3.803	1.134	0.892	0.225	0.005	0.120	0.455

5.3 Ground truth model for comparison

In practice, for any input vector, $\boldsymbol{x}$, the corresponding mooring line force, $y(t)$, can be obtained from the high-fidelity CFD model (section 3.4). However, the hybrid surrogate model developed in section 4 has been successfully validated to approximate the real system, yet is computationally faster compared to the classical CFD modeling. Therefore, this surrogate can be viewed as the real system (ground truth solution) at the active learning process. Specifically, Figure 10 (2.iii) shows that the algorithm calls the real system to evaluate the next-best sample, $\boldsymbol{x}^*$, and provide the corresponding mooring force, y^*. To the purpose of the present study, the choice of the hybrid surrogate model of section 4 allows for the accurate output evaluation without the punishing computational cost of repeated CFD simulations.

5.4 Extreme Force Statistics in the Mooring Line

The quantification of extreme statistics of a quantity of interest, i.e., PDF of extreme load on a critical component, are critical in the design and reliability assessment of nonlinear dynamical systems [32]. For practical applications, the definition of the PDF of a quantity of interest is a challenging process due to limitations in the classical modeling practices. That is to say, the availability of high-fidelity numerical models and/or experimental tests to provide the outputs for several inputs is limited due to cost and resources.

By employing the developed surrogate model, the extreme force statistics can be effectively reconstructed. In this study, the quantity of interest is the maximum peak force, y_{max}, (i.e., second peak in the force time series)

$$y_{max} = \max_{t\in[0,T]} |y_{\mathcal{L}}(t)|. \tag{34}$$

First, $n_{MC} = 10^5$ input samples $\{\boldsymbol{x}_i\}_{i=1}^{n_{MC}}$ are drawn with Monte Carlo sampling from the hyperbox defined in section 5.2. Next, the input samples are pushed into the GPR model which provides the initial prediction of the mooring force, $\bar{y}_{\mathcal{G}}(t)$, which in turn is the input to the LSTM model that computes the corrected force, $y_{\mathcal{L}}(t)$. Finally, the force PDF, p_s, is reconstructed via the KDE based on the maximum mooring force, y_{max}, for each output $y_{\mathcal{L}}(t)$. This KDE construction has superior statistical convergence compared to a Monte Carlo histogram, especially for the distribution tails (extreme events).

5.4.1 Error metric

To evaluate the ability of the active learning algorithm to quantify the extreme force PDF, the real system PDF (i.e., ground truth), p_y, is compared with the reconstructed PDF, p_s, at each iteration t of the algorithm. The log-pdf error is reported as the metric to quantify the discrepancy between the PDFs

$$e(t) = \int |\log_{10} p_{s_t}(y) - \log_{10} p_y(y)| dy \tag{35}$$

Both PDFs are estimated on 10^5 samples via Monte Carlo design (see section 5.4). The error metric includes the logarithms in order to emphasize the comparison of the tails of the PDFs. In this study, as 15 Bayesian experiments are conducted and the p_{s_t} expresses the median.

6 Results

6.1 Hybrid Surrogate Model for Time Series Prediction

The performance of the hybrid surrogate model, developed in section 4 and trained on data from 65 CFD simulations, is analyzed in this section. The model aims to predict the mooring force at a point-absorber WEC operating in 50-year extreme waves.

The GPR component of the hybrid surrogate maps the input vector, $\boldsymbol{x}$, to the conditional posterior mean of each PCA coefficient, $\bar{q}_i{}_{i=1}^{12}$. Subsequently, the PCA coefficients are combined using Eq. (21) to construct the force time series. Figure 11 (Left) illustrates a comparison between the force predicted by the GPR model and the CFD solution for four validation cases. The qualitative assessment reveals that the GPR model is capable of approximating the complex force time series; however, it underestimates the three main peaks. This underestimation can be attributed, in part, to the finite PCA truncation, which excludes very high frequency components. Additionally, the constrained size of the dataset further restricts the predictions of the first and third peaks.

In this study, the focus is on extreme events, making it crucial to accurately capture peak forces. This motivation leads to the inclusion of the LSTM component in the hybrid surrogate model. The LSTM neural network takes the mooring force time series from the GPR model as input and provides an improved prediction as output. Figure 11 (Right) presents a qualitative comparison of the time series predicted by the LSTM, GPR, and CFD models, demonstrating that the LSTM component enhances the accuracy of the predictions.

For a quantitative assessment of the prediction accuracy in the peak regions, the coefficient of determination, denoted as R^2, is utilized as a goodness-of-fit metric. This metric measures how well the quantity of interest predicted by the surrogate model compares to the corresponding values obtained from the CFD solution. The results are presented in Figure 12 in the form of a scatter plot. The plot illustrates the distribution of the three predicted peak forces.

The GPR model (represented by blue dots) demonstrates a high level of accuracy in predicting the second peak ($R^2 = 0.852$). However, its performance decreases for the first ($R^2 = 0.188$) and third peaks ($R^2 = 0.299$). On the other hand, the LSTM model (represented by red dots) significantly improves the prediction accuracy. The coefficient of determination, R^2, increases to 0.822 for the first peak (from 0.188), 0.934 for the second peak (from 0.852), and 0.703 for the third peak (from 0.299). This enhancement demonstrates the effectiveness of the LSTM model in improving the accuracy of the predictions.

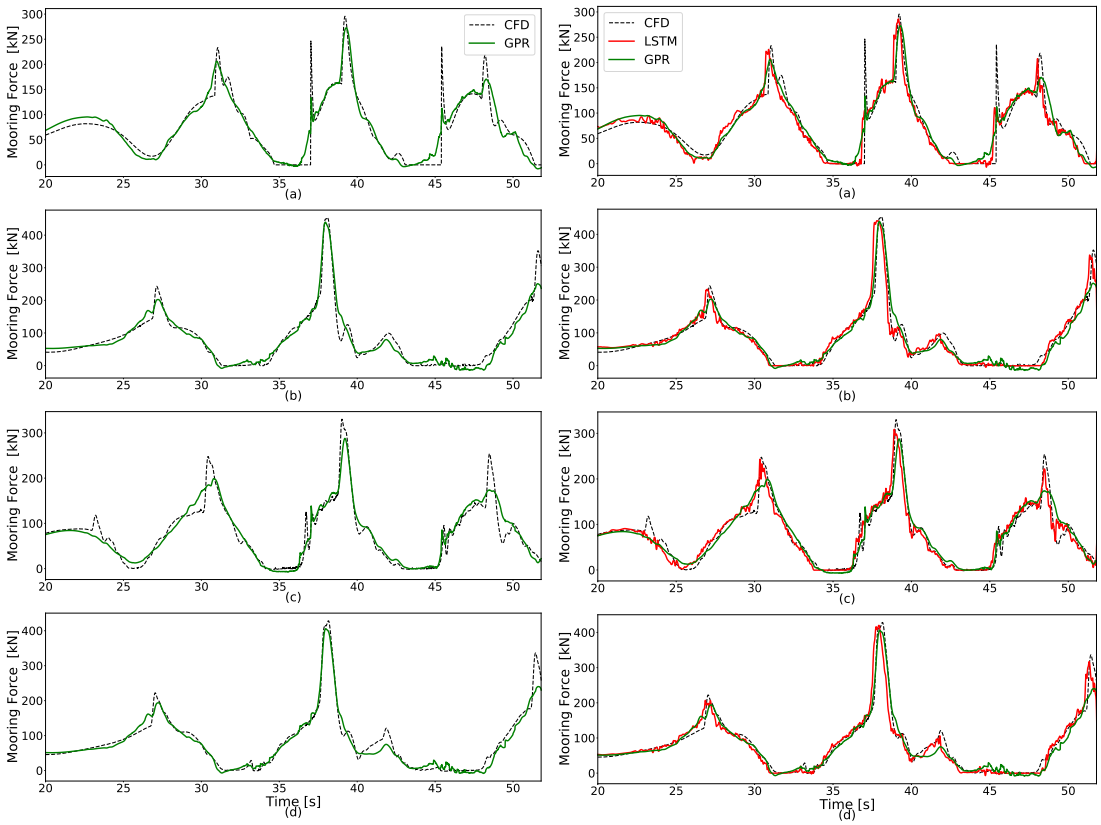
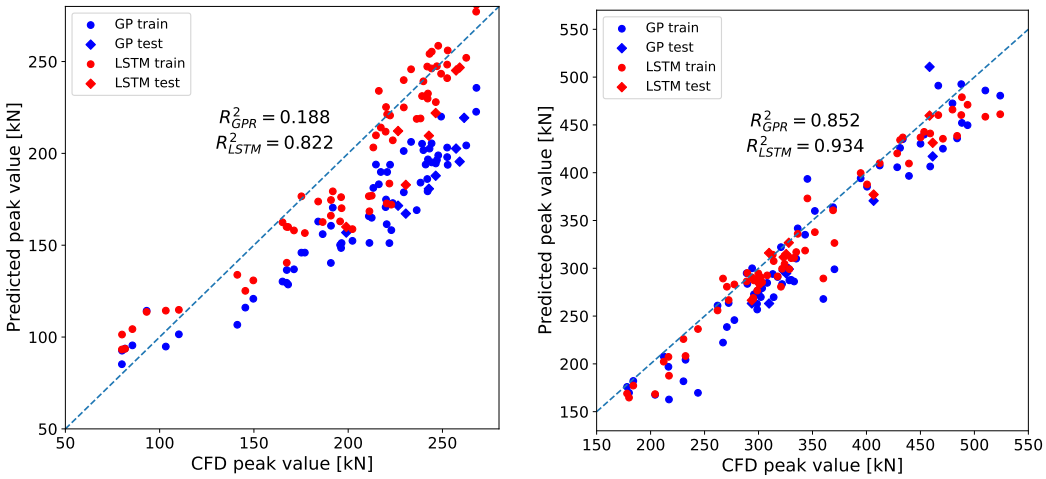
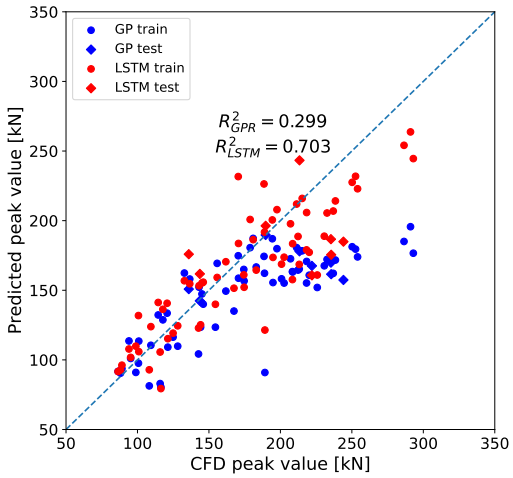


Figure 11: [**Left**] Comparison of the GPR mooring force (green) with the CFD solution (dashed black) for four validation cases. [**Right**] Improved time series prediction by the LSTM model (red), particularly for the three main peaks.



(a) First peak

(b) Second peak



(c) Third peak

Figure 12: Comparison of the three main peaks predicted by the GPR and LSTM models to the CFD solution, quantified by the coefficient of determination R^2. The LSTM model shows a significant improvement in the R^2 value.

To assess the predictive performance of the hybrid surrogate model for the entire force time series (not just the peaks), Figure 13 compares the force spectral density obtained from the GPR and LSTM machine learning models with the CFD simulations. It is worth noting that the GPR-based spectral density does not accurately capture certain frequency components, which can be attributed to the dimensionality reduction of the force using a low-rank vector of PCA coefficients. However, the LSTM component improves not only the accuracy of the peaks but also the overall prediction of the entire time series.

For a quantitative comparison of the spectra, the l_1 norm is employed as the error metric within the frequency range $[f_l, f_u] = [0.04, 0.16]$ Hz, given by

$$l_1 = \int_{f_l}^{f_u} |s_{surrog}(f) - s_{CFD}(f)| df, \tag{36}$$

where s_{surrog} and s_{CFD} represent the predicted and CFD force spectral densities, respectively, and f denotes the frequency. Figure 14 provides a summary of the l_1 norm metric for the eight validation samples. It can be observed that the error l_1 is reduced after the LSTM model is implemented, indicating improved accuracy in the prediction of the force spectral density.

6.1.1 Computational cost

The motivation behind the construction of the hybrid surrogate model is driven by the need to introduce a computationally inexpensive alternative when the direct simulation of wave-structure interaction would be computationally prohibitive. In this section, the resources and time required by CFD simulations and machine learning techniques are compared and discussed.

For a given input vector, $\boldsymbol{x} = [H_s, T_p, D_{PTO}, K_{es}]$, approximately 3000 CPU hours or 1 calendar day is demanded by the CFD code to simulate the wave-structure interaction and provide the corresponding force in the mooring line. The simulations are performed on the Tetralith HPC cluster, with 128 cores being occupied. In this study, results from 73 CFD simulations are employed for training and validating the surrogate models.

In contrast, the GPR surrogate model offers the great advantage of being a very inexpensive tool. The training of the GPR model on 65 datasets is a very quick process that takes just a few seconds, while the trained model also provides the solution in just a few seconds. This process is executed on a computer equipped with an Intel Xeon W-2135 CPU, NVIDIA Quadro P1000 GPU, and 64 GB RAM. The training of the LSTM neural network requires a longer time since the model is tasked with predicting a time series. Specifically, in this study, the training stage requires 10 hours. However, once the LSTM model is trained, it provides the output in approximately 1-2 minutes.

In summary, the output prediction for a massive number of input variables is successfully achieved by the hybrid surrogate model within just a few minutes. In contrast, the CFD simulations would demand many calendar days, consuming significant computational resources.

6.2 Development of a Surrogate Model via Active Learning

In this section, a hybrid surrogate model is developed (as detailed in Section 5), employing the active learning scheme within Bayesian experimental design to effectively select the proper training samples for ensuring the model's reliability in the estimation of extreme force statistics. At the core of this procedure lies the acquisition function, which guides the sample selection process. Two acquisition functions, namely US-LW and US, as outlined in Section 2.5.1, are utilized. For comparison purposes,

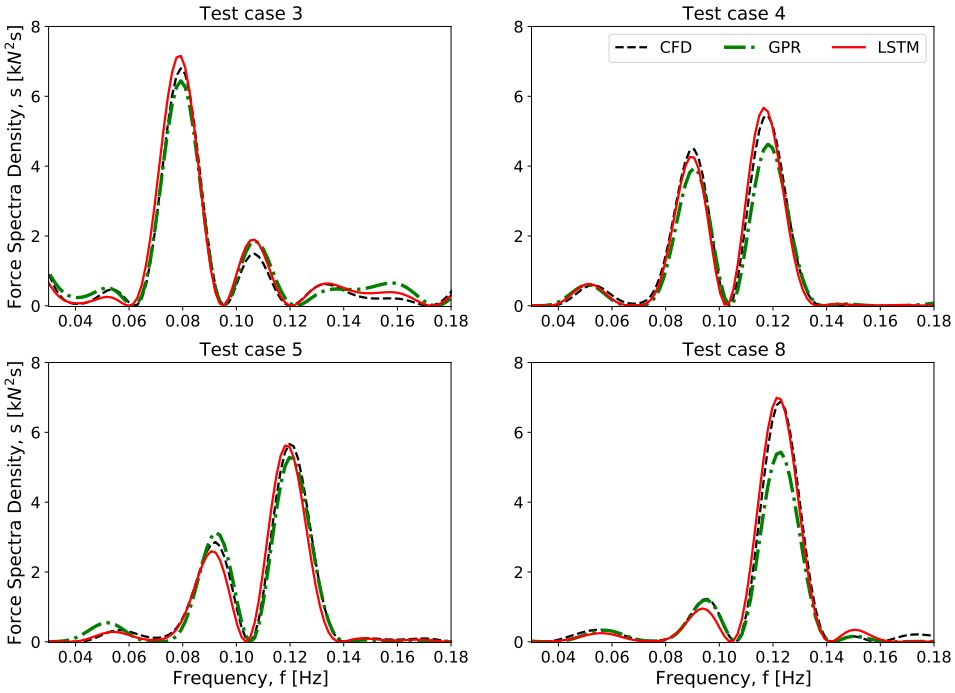


Figure 13: Comparison of the force spectra density obtained from the GPR (green) and LSTM (red) models, compared to the CFD solution (black), demonstrating better agreement with the LSTM model.

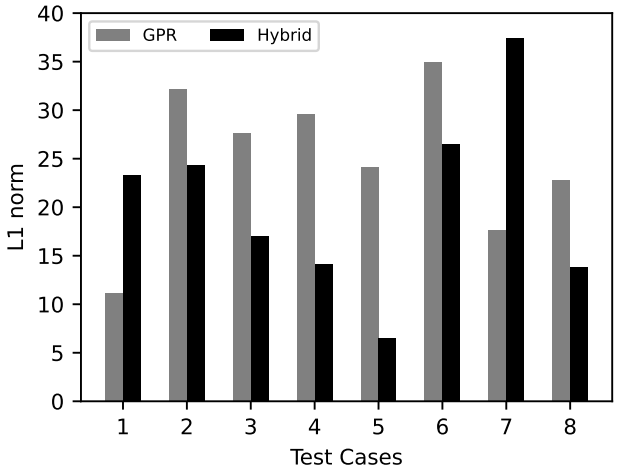


Figure 14: Comparison of the predicted force spectra density by the surrogate model and the actual spectra from CFD solution. Surrogate models include: GPR-based (gray) and hybrid model (black) which combines GPR and LSTM. The error evaluation is based on the L_1 norm metric.

the LHS method – a widely employed technique in experimental design – is also employed for the selection of training data. The results are compared within this section to provide a comprehensive evaluation.

Samples selection: Although the input space includes both wave conditions and PTO parameters, the wave characteristics (H_s, T_p) assume a more prominent role in guiding the algorithm's search. In Figure 15, a comparison is presented between the decisions made by the US and US-LW acquisition functions after 10 and 20 iterations. Both acquisition functions foster exploration of regions within the input space characterized by high uncertainty. However, the US-LW offers a distinct advantage in sample selection as it also takes into account the significance of the output concerning the input (exploitation). Conversely, the US's primary focus is on uncertainty reduction, neglecting the input distribution and output values. Consequently, this results in relatively uniform sampling, with the US exhibiting a preference for selecting points along the boundaries of the input parameter space – a trend consistently observed in [3]. In contrast, the US-LW opts for samples along a diagonal band and avoids an excessive emphasis on boundary points. Notably, the US-LW prioritizes experimental conditions that lead to the generation of substantial waves and significant structural forces.

Solution convergence: To assess the uncertainty of the surrogate model, each acquisition function undergoes evaluation through 15 separate Bayesian experiments. Each experiment begins by utilizing a pre-trained GPR model and then proceeds through 20 active learning iterations. At each iteration, the log-pdf error metric (as detailed in section 5.4.1) gauges the surrogate model's performance by comparing the statistics of the reconstructed force PDF with the ground truth PDF. Figure 16 (Left) displays the log-pdf error for both the US and US-LW acquisition functions, presenting the mean log-pdf error resulting from multiple experiments. Moreover, a comparative assessment is

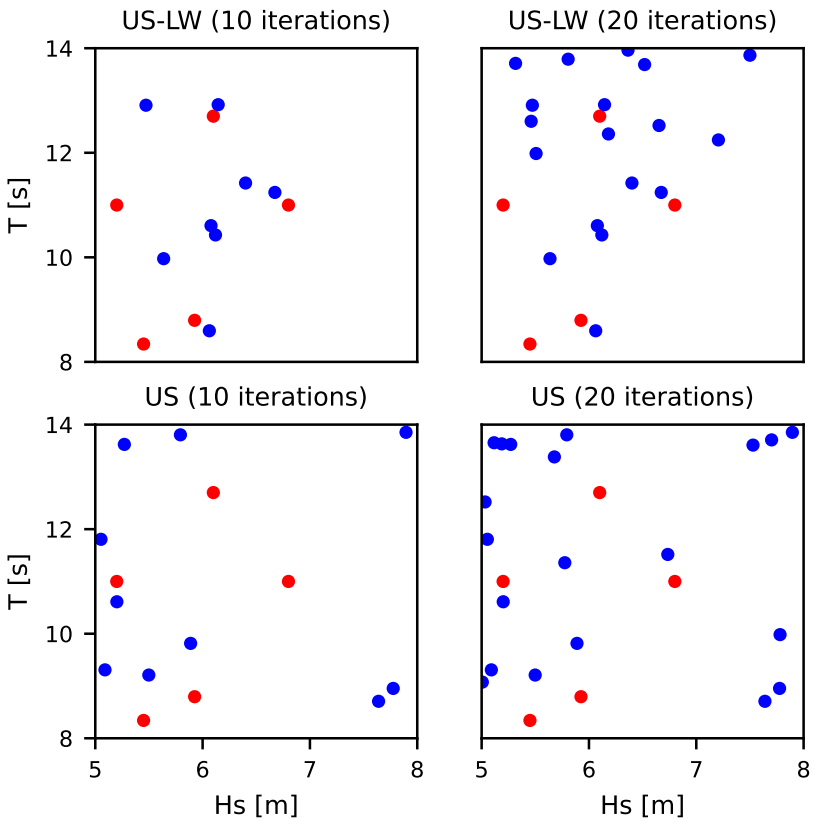


Figure 15: Progression of the sampling algorithm for US-LW (top row) and US (bottom row) after 10 iterations (left column) and 20 iterations (right column). The red dots represent the five initial random samples, while the blue indicate the points visited by the active learning algorithm.

carried out, contrasting the Bayesian sequential sampling algorithm with LHS. This comparison highlights the benefits of employing the US-LW acquisition function within the domain of Bayesian experimental design. The surrogate model built upon US-LW demonstrates superior accuracy in predicting mooring force statistics compared to both US and LHS, and it does so with faster convergence. Finally, an LSTM neural network is trained using the data sampled after 20 iterations of the active learning algorithm. Combining the GPR model from the 20th iteration with the LSTM model creates the hybrid surrogate model, which is used for predicting extreme force statistics. Figure 16 (Right) illustrates that the incorporation of the LSTM component in the surrogate model significantly reduces the error, resulting in better alignment of the reconstructed PDF with the Ground Truth PDF. Among all the methods employed for training data sampling, US-LW yields the lowest error, and the introduction of the LSTM component further reduces the error by 22.7%.

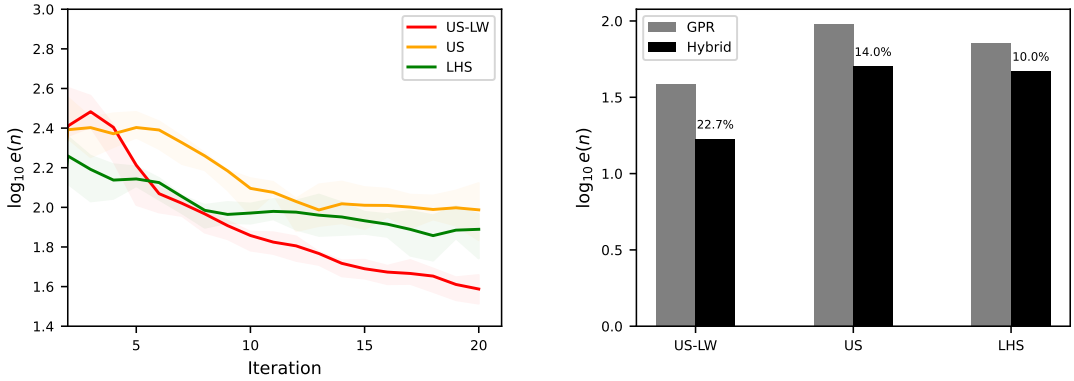


Figure 16: **[Left]** Log-pdf error per AS iteration for US-LW, US, and LHS, representing the mean error from 15 Bayesian experiments. It compares Ground Truth PDF with the PDF reconstructed by the GPR model at each iteration, with error bands indicating half of the median absolute deviation. **[Right]** After the 20th iteration, the hybrid surrogate model is trained and compared to the 20th iteration GPR model.

6.3 Surrogate Model Applications

6.3.1 Extreme Force Statistics

Main objective of this study is to quantify maximum mooring force statistics using random sets of input variables. Classical high-fidelity modeling techniques, such as CFD simulations, are impractical for this task due to their high computational demands. Instead, a surrogate model was developed through 20 active learning iterations, enabling rapid predictions of mooring forces for thousands of Monte Carlo input samples.

The extreme mooring force PDF reconstruction follows the procedure outlined in Section 5.4. Figure 17 [Left] presents the PDFs reconstructed by the 20th iteration GPR model, which was trained using data sampled through different methods, i.e., US-LW, US, and LHS. For a more precise comparison, the ground truth PDF is also included. Similarly, Figure 17 [Right] displays the PDF generated by the hybrid surrogate model, comprising the 20th iteration GPR model and an LSTM

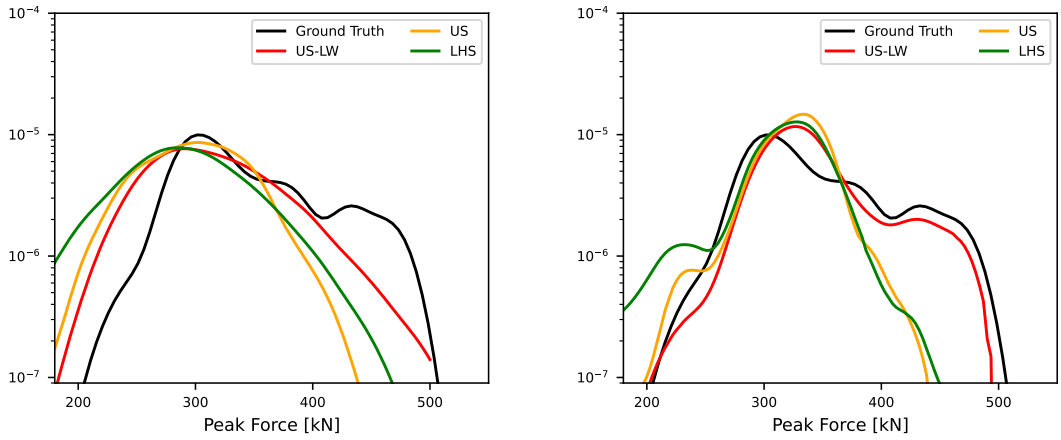


Figure 17: PDF of the maximum mooring line force estimated from the **[Left]** GPR model developed at the 20^{th} Active Learning iteration and **[Right]** hybrid surrogate model that consists of the GPR model of the 20^{th} iteration and an LSTM model. The PDF is reconstructed from GPR models trained on samples selected using US-LW, US and LHS. The PDF is compared to the Ground Truth PDF (black).

neural network. This combination significantly enhances the quality of the PDF, underscoring the significance of the hybrid model. Furthermore, Figure 17 illustrates the value of utilizing the US-LW acquisition function, consistently outperforming alternative methods.

6.3.2 Interpretation of the inferred surrogate model

Figure 18 and Figure 19 present contour plots illustrating the influence of input variables, $\boldsymbol{x} = [H_s, T_p, D_{PTO}, K_{es}]$, on the maximum mooring force. As it is difficult to visualize the four-dimensional input vector, a series of two-dimensional plots are generated, keeping two variables constant at a time. These contour plots offer insights into the individual impact of each input variable but also the interactions between them.

In particular, Figure 18 illustrates how variations in wave characteristics impact the maximum mooring force while holding the PTO parameters constant. These contour plots are useful during the preliminary design phase of a WEC system at a specific offshore site as they facilitate the assessment of maximum loads in the mooring system for any given set of PTO parameters. For instance, at an offshore site characterized by a prominent sea state with a wave period of $T_p = 12$ s and a significant wave height of $H_s = 5$ m, the maximum force in the mooring line is expected to fall within the range of approximately 200-240 kN, regardless of the combination of PTO values employed. Conversely, at an offshore location where the dominant sea state features $T_p = 14$ s and $H_s = 7$ m, the maximum force can vary between 480 and 680 kN, depending on the chosen PTO parameters. Hence, the selection of suitable PTO parameter values holds critical importance.

In the top-row subplots of Figure 18, PTO damping (D_{PTO}) remains constant while spring stiffness (K_{es}) varies. Similarly in the bottom-row subplots. Notably, it is observed that the mooring force undergoes relatively minor changes in this scenario. This behavior can be attributed to the

fact that the spring is not fully compressed in most sea states, resulting in minimal additional force being exerted on the mooring line. Conversely, the first-column figures maintain a consistent spring stiffness value (K_{es}) while altering the damping (D_{PTO}), and the same applies to the second-column figures. In this situation, the mooring force exhibits higher values with increased damping. This phenomenon can be explained by the fact that higher PTO damping applies a greater force on the buoy through the mooring line, dampening the buoy's motion during interactions with incoming waves.

Figure 19 provides insights into how adjustments in the PTO parameters (K_{es}, D_{PTO}) influence the extreme mooring force while keeping the sea state variables (H_s, T_p) constant. These contour plots offer valuable guidance for the development of a control algorithm aimed at adapting the damping of the PTO system to specific sea states, as discussed in [41]. In the top-row subplots, H_s remains consistent while T_p increases, and similarly in the bottom-row subplots. As the wave period T_p increases, the wave length expands, leading to a higher wave celerity and consequently an increased potential for wave energy. This results in higher loads on the structure. The first-column subplots maintain the same T_p value but increase H_s, and the same holds for the second-column subplots. In this scenario, the wave becomes steeper, providing an explanation for the occurrence of higher forces, as previously observed in [19] and [18].

7 Conclusions

When the direct modeling of a real-world offshore system becomes computationally prohibitive, the development of a reliable surrogate model becomes essential. Yet, in practical applications where each sample evaluation demands valuable time and resources, the careful selection of samples for the proper model development becomes crucial. In this study, an active learning scheme in Bayesian experimental design and advanced machine learning techniques are leveraged in a framework that effectively creates a surrogate model. The great advantage of this framework is that it circumvents the need for massive training data while reduces the prediction uncertainty. Specifically, the active learning scheme effectively guides the selection of training samples, uncovering regions within the parameter space that yield the most pertinent information about extremes and higher uncertainty. The resulting surrogate model is then deployed to quantify the statistics of the quantity of interest.

While the effectiveness of the developed framework has been previously demonstrated in [4,38], its application to realistic scenarios, such as a wave energy system subjected to 50-year waves, is clearly demonstrated. In offshore structures, the extreme loads on critical components, such as the mooring system, are a natural quantity of interest. In this study, the hybrid surrogate model for a point-absorber WEC is constructed combining two machine learning methods, specifically GPR and LSTM neural networks. This surrogate exhibits the capability to predict the complex mooring force with remarkable success, validated against CFD simulations, and does so at orders of magnitude faster compared to CFD. Subsequently, the surrogate proves invaluable in quantifying extreme mooring force statistics, evaluating thousands of Monte Carlo samples, and effectively reconstructing the PDF, with particular emphasis on accurately capturing the tails. Furthermore, the surrogate model can be used to offer insights into the system's behavior, i.e., how the input variables influence the quantity of interest.

The accurate quantification of extreme loads not only has the potential to reduce costs by refining conservative safety factors but also ensures that systems are designed to withstand these extreme conditions, preserving reliability. Additionally, reliable surrogate models can expedite the design process and optimize the structural design of emerging wave energy technologies.

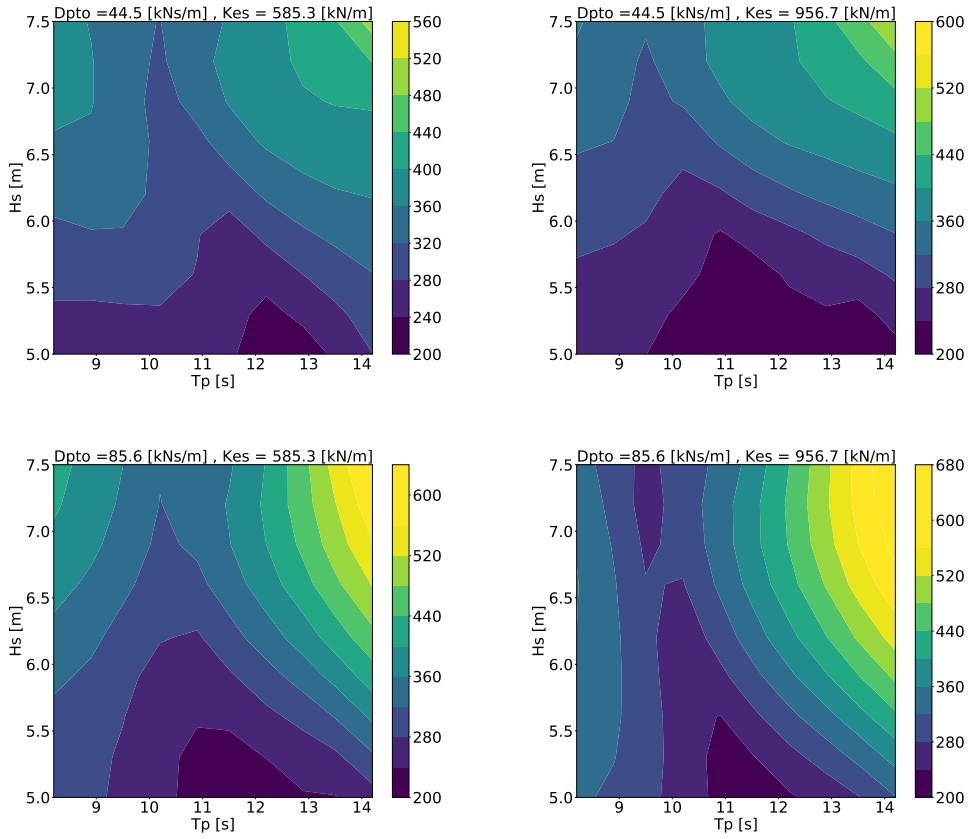


Figure 18: Visualization of surrogate model projections: Contour plots depicting the maximum mooring force. Colorbar represents values in [kN]. Each subplot illustrates force variations with wave characteristics (H_s, T_p), keeping other input parameters constant (D_{PTO}, K_{es}).

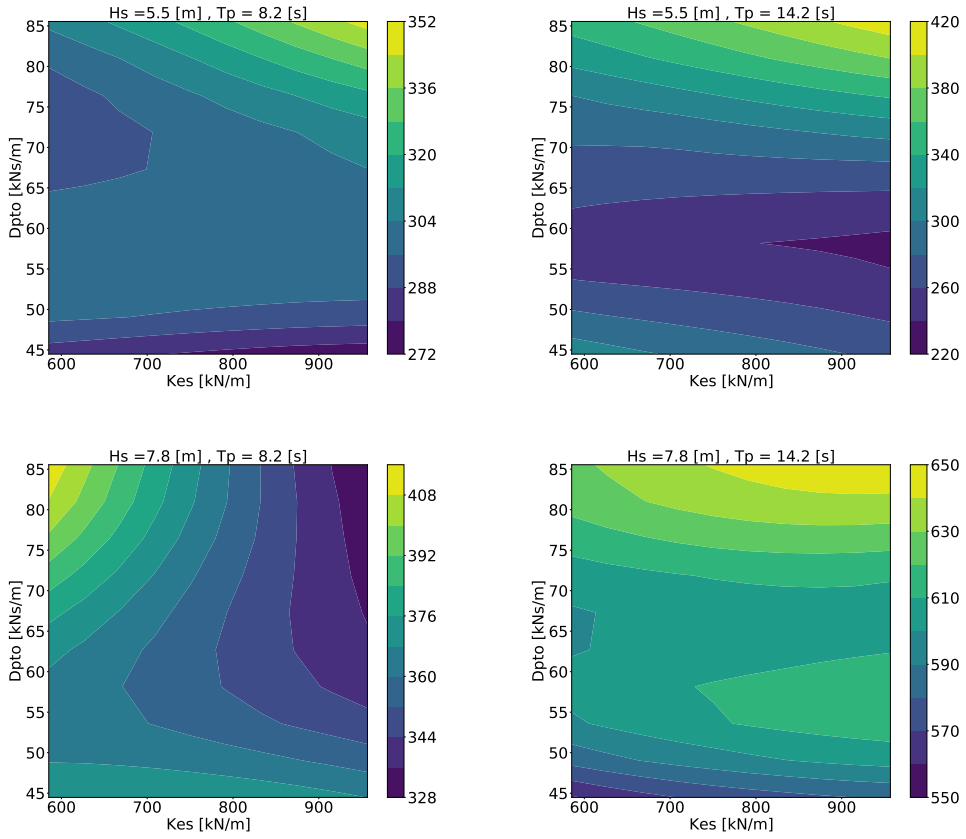


Figure 19: Visualization of surrogate model projections: Contour plots depicting the maximum mooring force. Colorbar represents values in [kN]. Each subplot illustrates force variations with PTO parameters (D_{PTO}, K_{es}), keeping wave characteristics (H_s, T_p) constant.

As a future direction, it is suggested to increase the dimensional input space by considering more design parameters. This step would require machine learning methods capable of accurately mapping a multi-functional input space to multi-functional outputs while ensuring accurate prediction in unseen data. Deep neural operators, such as DeepONets [31] and [37], are recommended for this purpose. Furthermore, the surrogate model can be further used for the development of an optimization procedure, which can be integrated into the design stage. Additionally, to address the practical difficulty of obtaining high-fidelity training data, a multi-fidelity surrogate model is suggested. This model leverages the advantages of low- and high-fidelity modeling tools, such as recursive co-kriging [36]. The low-fidelity part of the surrogate provides the solution trend, while the high-fidelity part learns the residuals (the difference between the low- and high-fidelity solution) and appropriately corrects the predictions. The multi-fidelity surrogate model will ensure that less high-fidelity data is necessary for training purposes.

Acknowledgments

The research in this paper was supported by the Swedish Centre of Natural Hazards and Disaster Science (CNDS), the Onassis Foundation (Scholarship ID: FZP 021-1/2019-2020, the Anna-Maria Lundins Scholarship (AMh2021-0023) and Liljewalch Scholarship. TPS and SG have been supported through the ONR grant N00014-21-1-2357.

The CFD simulations were performed on resources provided by the Swedish National Infrastructure for Comuting (SNIC) at the HPC cluster Tetralith at the National Supercomputer Centre, at Linköping University.

References

[1] Enrico Anderlini. *Control of wave energy converters using machine learning strategies*. PhD thesis, University of Edinburgh, 2017.

[2] Aurélien Babarit, Jorgen Hals, Made Jaya Muliawan, Adi Kurniawan, Torgeir Moan, and Jorgen Krokstad. Numerical benchmarking study of a selection of wave energy converters. *Renewable energy*, 41:44–63, 2012.

[3] Antoine Blanchard and Themistoklis Sapsis. Bayesian optimization with output-weighted optimal sampling. *Journal of Computational Physics*, 425:109901, 2021.

[4] Antoine Blanchard and Themistoklis Sapsis. Output-weighted optimal sampling for Bayesian experimental design and uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):564–592, 2021.

[5] Bret Bosma, Zhe Zhang, Ted KA Brekken, H Tuba Özkan-Haller, Cameron McNatt, and Solomon C Yim. Wave energy converter modeling in the frequency domain: A design guide. In *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*, pages 2099–2106. IEEE, 2012.

[6] Alain Clément, Pat McCullen, António Falcão, Antonio Fiorentino, Fred Gardner, Karin Hammarlund, George Lemonis, Tony Lewis, Kim Nielsen, Simona Petroncini, et al. Wave

energy in Europe: current status and perspectives. *Renewable and sustainable energy reviews*, 6(5):405–431, 2002.

[7] Richard Courant, Kurt Friedrichs, and Hans Lewy. On the partial difference equations of mathematical physics. *IBM journal of Research and Development*, 11(2):215–234, 1967.

[8] Balazs Czech and Pavol Bauer. Wave energy converter concepts: Design challenges and classification. *IEEE Industrial Electronics Magazine*, 6(2):4–16, 2012.

[9] Manhar R Dhanak and Nikolaos I Xiros. *Springer handbook of ocean engineering*. Springer, 2016.

[10] Wenyuan Fan and Henryk Anglart. On the closure requirement for vof simulations with rans modeling. *arXiv preprint arXiv:1911.09727*, 2019.

[11] Jan J. Gerbrands. On the relationships between SVD, KLT and PCA. *Pattern Recognition*, 14(1):375–381, 1981. 1980 Conference on Pattern Recognition.

[12] Bingyong Guo and John V Ringwood. A review of wave energy technology from a research and commercial perspective. *IET Renewable Power Generation*, 15(14):3065–3090, 2021.

[13] Stephen Guth, Bianca Champenois, and Themistoklis P Sapsis. Application of gaussian process multi-fidelity optimal sampling to ship structural modeling. In *34th Symposium on Naval Hydrodynamics Proceedings Washington DC 2022*, 2022.

[14] Stephen Guth and Themistoklis P. Sapsis. Wave episode based gaussian process regression for extreme event statistics in ship dynamics: Between the Scylla of Karhunen–Loève convergence and the Charybdis of transient features. *Ocean Engineering*, 266:112633, 2022.

[15] Sverre Haver and Steven R Winterstein. Environmental contour lines: A method for estimating long term extremes by a short term analysis. In *SNAME Maritime Convention*. OnePetro, 2008.

[16] International Renewable Energy Agency (IRENA), Abu Dhabi. *Innovation outlook: Ocean energy technologies*, 2020.

[17] Eirini Katsidoniatski, Yi-Hsiang Yu, and Malin Goteman. Midfidelity model verification for a point-absorbing wave energy converter with linear power takeoff. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2021.

[18] Eirini Katsidoniotaki and Malin Göteman. Numerical modeling of extreme wave interaction with point-absorber using openfoam. *Ocean Engineering*, 245:110268, 2022.

[19] Eirini Katsidoniotaki, Erik Nilsson, Anna Rutgersson, Jens Engström, and Malin Göteman. Response of point-absorbing wave energy conversion system in 50-years return period extreme focused waves. *Journal of Marine Science and Engineering*, 9(3):345, 2021.

[20] Eirini Katsidoniotaki, Foivos Psarommatis, and Malin Göteman. Digital twin for the prediction of extreme loads on a wave energy conversion system. *Energies*, 15(15):5464, 2022.

[21] Eirini Katsidoniotaki, Edward Ransley, Scott Brown, Johannes Palm, Jens Engström, and Malin Göteman. Loads on a point-absorber wave energy converter in regular and focused extreme wave events. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 84416, page V009T09A022. American Society of Mechanical Engineers, 2020.

[22] Eirini Katsidoniotaki, Zahra Shahroozi, Claes Eskilsson, Johannes Palm, Jens Engström, and Malin Göteman. Validation of a CFD model for wave energy system dynamics in extreme waves. *Under revision in Ocean Engineering*, 2022.

[23] Maria Katsidoniotaki. *Uncertainty quantification techniques with diverse applications to stochastic dynamics of structural and nanomechanical systems and to modeling of cerebral autoregulation.* PhD thesis, Columbia University, 2022.

[24] Maria I Katsidoniotaki, Apostolos F Psaros, and Ioannis A Kougioumtzoglou. Uncertainty quantification of nonlinear system stochastic response estimates based on the Wiener path integral technique: A Bayesian compressive sampling treatment. *Probabilistic Engineering Mechanics*, 67:103193, 2022.

[25] Stephen Guth Eirini Katsidoniotaki and Themistoklis P. Sapsis. Statistical modeling of fully nonlinear hydrodynamic loads on offshore wind turbine foundations using wave episodes and targeted CFD simulations through active sampling. *Submitted*, 2023.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Mats Leijon, Cecilia Boström, Oskar Danielsson, Stefan Gustafsson, Kalle Haikonen, Olivia Langhamer, Erland Strömstedt, Magnus Stålberg, Jan Sundberg, Olle Svensson, et al. Wave energy from the north sea: Experiences from the lysekil research site. *Surveys in geophysics*, 29(3):221–240, 2008.

[28] Liang Li, Zhiming Yuan, and Yan Gao. Maximization of energy absorption for a wave energy converter using the deep machine learning. *Energy*, 165:340–349, 2018.

[29] Xuan Li and Wei Zhang. Long-term fatigue damage assessment for a floating offshore wind turbine under realistic environmental conditions. *Renewable Energy*, 159:570–584, 2020.

[30] Yuqi Liu, Xiaocheng Liu, Jinkang Guo, Ranran Lou, and Zhihan Lv. Digital twins of wave energy generation based on artificial intelligence. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 718–719. IEEE, 2022.

[31] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[32] Mustafa A Mohamad and Themistoklis P Sapsis. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 115(44):11138–11143, 2018.

[33] Guilherme Moura Paredes, Claes Eskilsson, and Allan P. Engsig-Karup. Uncertainty quantification in mooring cable dynamics using polynomial chaos expansions. *Journal of Marine Science and Engineering*, 8(3):162, 2020.

[34] Seyed Milad Mousavi, Majid Ghasemi, Mahsa Dehghan Manshadi, and Amir Mosavi. Deep learning for wave energy converter modeling using long short-term memory. *Mathematics*, 9(8):871, 2021.

[35] Dezhi Ning and Boyin Ding. *Modelling and Optimization of Wave Energy Converters*. CRC Press, 2022.

[36] P. Perdikaris, D. Venturi, J. O. Royset, and G. E. Karniadakis. Multi-fidelity modelling via recursive co-kriging and gaussian–markov random fields. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150018, 2015.

[37] Ethan Pickering, Stephen Guth, George Em Karniadakis, and Themistoklis Sapsis. Discovering and forecasting extreme events via active learning in neural operators. *Nature Computational Science*, 2:823–833, 2012.

[38] Ethan Pickering, Stephen Guth, George Em Karniadakis, and Themistoklis P Sapsis. Discovering and forecasting extreme events via active learning in neural operators. *Nature Computational Science*, 2(12):823–833, 2022.

[39] Román Quevedo-Reina, Guillermo M Álamo, Luis A Padrón, and Juan J Aznárez. Surrogate model based on ann for the evaluation of the fundamental frequency of offshore wind turbines supported on jackets. *Computers & Structures*, 274:106917, 2023.

[40] Edward Jack Ransley. *Survivability of wave energy converter and mooring coupled system using CFD*. PhD thesis, Plymouth University, 2015.

[41] Claudio A Rodríguez, Paulo Rosa-Santos, and Francisco Taveira-Pinto. Assessment of damping coefficients of power take-off systems of wave energy converters: A hybrid approach. *Energy*, 169:1022–1038, 2019.

[42] Themistoklis P. Sapsis. Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2234):20190834, 2020.

[43] Themistoklis P Sapsis. Statistics of extreme events in fluid flows and waves. 2021.

[44] Qinshuo Shen, Faridaddin Vahdatikhaki, Hans Voordijk, Jeffrey van der Gucht, and Lex van der Meer. Metamodel-based generative design of wind turbine foundations. *Automation in construction*, 138:104233, 2022.

[45] Masanobu Shinozuka. Basic analysis of structural safety. *Journal of Structural Engineering*, 109(3):721–740, 1983.

[46] András Sobester, Alexander Forrester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

[47] Peter S Tromans, Ali R Anaturk, and Paul Hagemeijer. A new model for the kinematics of large ocean waves-application as a design wave. In *The first international offshore and polar engineering conference*. OnePetro, 1991.

[48] Jennifer van Rij, Yi-Hsiang Yu, Yi Guo, and Ryan G Coe. A wave energy converter design load case study. *Journal of Marine Science and Engineering*, 7(8):250, 2019.

[49] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[50] Ziming Wang, Dongsheng Qiao, Jun Yan, Guoqiang Tang, Binbin Li, and Dezhi Ning. A new approach to predict dynamic mooring tension using lstm neural network based on responses of floating structure. *Ocean Engineering*, 249:110905, 2022.

[51] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[52] Linus Wrang, Eirini Katsidoniotaki, Erik Nilsson, Anna Rutgersson, Jesper Rydén, and Malin Göteman. Comparative analysis of environmental contour approaches to estimating extreme waves for offshore installations for the baltic sea and the north sea. *Journal of Marine Science and Engineering*, 9(1):96, 2021.

[53] Yi-Hsiang Yu and Ye Li. Reynolds-Averaged Navier–Stokes simulation of the heave performance of a two-body floating-point absorber wave energy system. *Computers & Fluids*, 73:104–114, 2013.

[54] Yi-Hsiang Yu, Ye Li, Kathleen Hallett, and Chad Hotimsky. Design and analysis for a floating oscillating surge wave energy converter. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 45547, page V09BT09A048. American Society of Mechanical Engineers, 2014.