Applied Ocean Research 155 (2025) 104439

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

Research paper

Reduced order modeling of wave energy systems via sequential Bayesian experimental design and machine learning

Eirini Katsidoniotaki [a,b,c,*], Stephen Guth [a], Malin Göteman [b,c], Themistoklis P. Sapsis [a]

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

ARTICLE INFO

Keywords:
Extreme events
Gaussian process regression (GPR)
LSTM neural networks
Surrogate models
Active sampling
Sequential Bayesian experimental design
Wave energy system
CFD simulations

ABSTRACT

Marine energy technologies face significant challenges in ensuring their survivability under extreme ocean conditions. Quantifying extreme load statistics on marine energy structures is essential for reliable structural design; however, this is a challenging task due to the scarcity of high-quality data and the inherent uncertainties associated with predicting rare events. While computational fluid dynamics (CFD) simulations can accurately capture the nonlinear dynamics and loads in extreme wave–structure interactions, providing high-fidelity data, extracting statistical information through these models is computationally impractical. This study proposes a reduced-order modeling framework for marine energy systems, enabling efficient analysis across diverse scenarios, and facilitating the quantification of extreme load statistics with significantly reduced computational cost. Specifically, a hybrid reduced-order or surrogate model for a wave energy converter is developed to map extreme sea states and design parameters to the resulting loads in the mooring system. The term "hybrid" refers to the combination of Gaussian Process Regression (GPR) and Long Short-Term Memory (LSTM) neural networks. The model is developed using two distinct approaches: (1) a baseline approach that relies on existing CFD data for training and validation, and (2) an active learning approach that strategically selects the most informative CFD samples from regions of the input space associated with extreme mooring loads. This procedure iteratively refines the model while minimizing prediction uncertainty, making it particularly effective for real-world applications where obtaining each sample requires substantial time and resources. The developed model demonstrates its exceptional ability to efficiently predict complex load time series, including instantaneous peaks, at speeds significantly faster than traditional modeling methods. Subsequently, the model is utilized to effectively evaluate Monte Carlo samples, providing accurate estimates of the probability of extreme mooring loads. Understanding the expected extreme loads is essential during the design phase of marine energy systems, enabling cost reduction by optimizing strength margins, refining overly conservative safety factors, and enhancing overall system reliability.

1. Introduction

Ocean waves hold immense potential for global electricity production (Clément et al., 2002), yet the current power capacity harnessed from waves remains limited. Although several wave energy conversion systems have been conceptualized, the majority has not reached commercial viability (Guo and Ringwood, 2021; Ning and Ding, 2022; International Renewable Energy Agency (IRENA), 2020). 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 WEC components, like mooring lines, can experience loads that are up to 100 times higher than average loading (Czech and Bauer, 2012). Estimating the probability of extreme structural loads is important for informing optimal design decisions, balancing reliability and cost-effectiveness.

In the presence of high and steep ocean waves, marine energy systems face nonlinear dynamic phenomena such as breaking waves, wave overtopping, slamming and turbulence effects, which can give rise to extreme structural loads. While CFD simulations enable accurate modeling of force-on-structure in extreme ocean conditions (Ransley, 2015; Yu and Li, 2013; Katsidoniotaki et al., 2021, 2022c; Katsidoniotaki and Göteman, 2022), their computational cost is very expensive, making it impractical to analyze the statistics of extreme forces. Frequency-domain and linear time-domain models are commonly used in the

* Corresponding author at: Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02138, USA.
E-mail address: eirka289@mit.edu (E. Katsidoniotaki).

https://doi.org/10.1016/j.apor.2025.104439
Received 25 September 2023; Received in revised form 19 December 2024; Accepted 20 January 2025
Available online 6 February 2025

preliminary design phase of marine energy systems due to their computational efficiency, allowing for the analysis of linear dynamics (Bosma et al., 2012; Yu et al., 2014; van Rij et al., 2019). However, these models fail to capture complex phenomena, resulting in significant uncertainties when modeling marine energy systems under extreme wave conditions (Yu and Li, 2013; Babarit et al., 2012; Katsidoniatski et al., 2021). This underscores the need for a balanced approach that integrates computational efficiency with accuracy.

Reduced-order models, also known as surrogate models, approximate complex physical systems by learning the relationships between input and output variables. These models are designed to replicate the actual behavior of the system while significantly reducing computational cost. They are typically data-driven models using data from high-fidelity numerical simulations or actual field measurements. By offering a simplified yet accurate representation, surrogate models are particularly valuable in scenarios requiring extensive evaluations, such as optimization, uncertainty quantification, and real-time predictions (Sobester et al., 2008). Numerous studies have successfully developed machine-learned surrogate models for various ocean engineering applications. For instance, GPR (Guth and Sapsis, 2022) and deep neural operators (Pickering et al., 2022) have been used to model ship load statistics in irregular seas. Surrogate models for offshore wind turbines have employed methods like artificial neural networks and GPR to estimate extreme load statistics (Katsidoniotaki and Sapsis, 2023), fundamental frequencies (Quevedo-Reina et al., 2023), fatigue damage (Li and Zhang, 2020), and optimal dimensioning (Shen et al., 2022). Other approaches, such as generalized polynomial chaos (Moura Paredes et al., 2020) and LSTM neural networks (Wang et al., 2022), have been applied to predict dynamic mooring forces in floating structures. In the context of marine energy systems, surrogate modeling has also seen significant advancements. For example, an LSTM neural network was developed to estimate the power output of a WEC (Mousavi et al., 2021), while deep learning techniques have been employed to optimize energy absorption in WECs (Anderlini, 2017; Li et al., 2018). Furthermore, digital twins for WECs have been developed using surrogate modeling techniques (Liu et al., 2022; Katsidoniotaki et al., 2022a). These examples demonstrate the growing application of surrogates in various aspects of offshore engineering, including marine energy systems.

Existing methods for constructing surrogate models often rely on randomly selecting training samples, without adequately considering the cost of acquiring these samples or the relevance of the information they contribute to the model (Shinozuka, 1983). Within the framework of Bayesian experimental design (BED), active learning offers a promising alternative by optimizing data sampling through the sequential selection of a minimal yet highly informative set of training samples, thereby reducing the need for large datasets. Active learning has proven to be effective for uncertainty quantification, particularly in estimating probability distributions in nonlinear dynamic systems (Katsidoniotaki et al., 2022b; Katsidoniotaki, 2022). The use of active learning to develop surrogate models for extreme events has great potential as it offers a promising solution to the challenges of data scarcity. In active learning, the acquisition function plays a pivotal role by directing the sampling process to select the most informative data points, tailored to the specific requirements of the application. In the context of extreme events, several studies have explored the formulation of the acquisition function to effectively identify the most informative data points of extreme occurrences, while aiming to minimize prediction uncertainty (Pickering et al., 2022; Blanchard and Sapsis, 2021b,a; Sapsis, 2021, 2020; Mohamad and Sapsis, 2018). In ocean engineering context, active learning has been effectively applied to predict the statistics of the extreme structural response of ships under random wave conditions (Pickering et al., 2022; Guth et al., 2022).

This work presents a novel framework for developing a "hybrid" surrogate model for a WEC, by combining GPR and LSTM neural networks. First, the hybrid model is constructed using **a baseline approach**, which utilizesrelies on existing CFD data for training—without active learning. The model effectively captures the relationship between extreme wave conditions, structural parameters, and the corresponding loads in the mooring line. The input parameters include the significant wave height and wave period associated with 50-year return period waves, as well as the damping coefficient and spring stiffness of the power take-off (PTO) system in the WEC. The 4-dimensional input space is directly mapped to the mooring force, enabling the surrogate model to predict how wave conditions and WEC design parameters influence mooring system forces. Second, the same hybrid model is developed but this time the selection of training samples is guided by **an active learning approach**, which prioritizes the most informative samples associated with extreme mooring loads while aiming to minimize prediction uncertainty. This approach optimizes the use of resources and computational time, particular important for real-world applications. Subsequently, the developed model is employed to reconstruct extreme load statistics on critical WEC components. The novelty of this study lies in overcoming the limitations of previous approaches in two key areas: (1) modeling of the wave energy system balancing computational cost and accuracy, and (2) effectively estimating the extreme load statistics–emphasizing the tails of the probability distribution. The developed framework can contribute in the optimal design process of marine energy technologies by providing an approach for cost-effective and reliable design, accelerating the development and establishment of these technologies.

The paper structure is as follows: Section 2 offers a detailed description of the WEC system. Section 3 serves a comprehensive review of the methods utilized in this work. Section 4 presents the hybrid surrogate model developed on existing CFD simulation data. Section 5 focuses on the active learning scheme for the construction of the hybrid surrogate model. In Section 6, the results of this study are presented, showcasing the performance and capabilities of the developed framework in predicting extreme force statistics. Finally, Section 7 summarizes the key findings and draws meaningful conclusions.

2. Wave energy converter in extreme sea states

Fig. 1 presents a schematic depiction of the point-absorber WEC (Leijon et al., 2008) considered in this study. The device comprises two main components: a buoy and a direct-driven linear generator housed in a capsule on the seabed, referred to as the PTO system. These two components are connected by a mooring line made of steel wire rope, and this is a critical component of the wave energy device. Within the PTO, the translator undergoes reciprocal motion, following the heaving motion of the buoy, inducing 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}, of the translator-stator, which governs the WEC's motion and the produced power.

The authors have previously developed and validated a high-fidelity CFD model to simulate the response of the WEC in extreme wave conditions (Katsidoniotaki et al., 2021, 2022c). While the CFD model provides accurate insights into the complex dynamic behavior of the device, the computational cost of those simulations and reliance on high-performance computing resources pose restrictions. This study proposes a reduced-order, or surrogate, modeling approach to efficiently evaluate a quantity of interest while significantly minimizing reliance on computationally expensive numerical simulations. In this context, the quantity of interest is the load experienced by the mooring line (rope) during the interaction of the WEC with extreme waves. Key

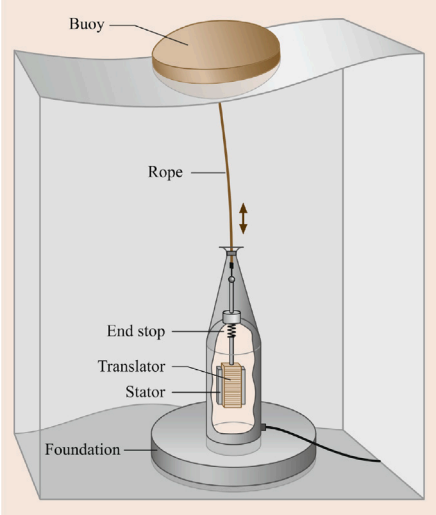


Fig. 1. A schematic depiction of the point-absorber wave energy converter considered in this study (Dhanak and Xiros, 2016).

factors influencing the mooring line loads include the wave characteristics – such as significant wave height (H_s) and peak wave period (T_p) – and system parameters like the end-stop spring stiffness (K_{es}) and the damping coefficient (D_{PTO}). These parameters serve as inputs for the surrogate model, with the range of values for each parameter detailed in Table 1.

The model is developed using two distinct approaches for selecting training samples: **(1) Baseline Approach** (Section 4): The model is trained using a fixed dataset derived from existing CFD results, without employing active learning. **(2) Active Learning** (Section 5): The model undergoes iterative training following an active learning procedure for sample selection. In this approach, the training samples are not preselected but are identified sequentially by an algorithm that prioritizes the most informative samples. These samples are chosen to enhance the model's ability to capture extreme forces while minimizing prediction uncertainty. As demonstrated in Section 6, the developed model is employed to accurately reconstruct extreme mooring load statistics, offering a comprehensive and reliable understanding of the extreme loads the device is expected to experience.

2.1. Distribution of the input parameters

The surrogate model inputs are represented as a 4-dimensional vector $\boldsymbol{x} = [H_s, T_p, D_{PTO}, K_{es}]$. The significant wave height, H_s, follows a *Weibull* distribution (Haver and Winterstein, 2008),

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

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

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

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

$$\mu(h) = a_1 + a_2 h^{a_3} \tag{3}$$

$$\sigma(h)^2 = b_1 + b_2 \exp\{-b_3 h\} \tag{4}$$

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 (Haver and Winterstein, 2008) and summarized in Table 2. 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.

Table 1
Surrogate model's input parameters.

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]

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.

3. Preliminaries

3.1. 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 variable, to the output, y, which represents the quantity of interest. The black-box approximates the real function, $y = f(x)$, with a surrogate model $\hat{f}(x)$, which is a statistical model that learns the input–output relationship based on training datasets. 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 for any new input x_*, the output y_* is provided in orders of magnitude faster. In Sections 3.2 and 3.4, two machine learning techniques used in this study are explained in further detail.

3.2. Gaussian process regression

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

$$y = \hat{f}(\boldsymbol{x}) + \epsilon. \tag{5}$$

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

$$\hat{f}(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right). \tag{6}$$

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, $m(\boldsymbol{x})$, and the covariance function, $k(\boldsymbol{x}, \boldsymbol{x}')$, of the process $\hat{f}(\boldsymbol{x})$ at any point $\boldsymbol{x} \in \mathbb{R}^d$ (Williams and Rasmussen, 2006).

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right], \\ k(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}[(\hat{f}(\boldsymbol{x}) - m(\boldsymbol{x}))(\hat{f}(\boldsymbol{x}') - m(\boldsymbol{x}'))] \end{aligned} \tag{7}$$

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

$$\begin{aligned} D_{train} &= (\mathbf{X}, \mathbf{y}) = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n}, \boldsymbol{x}_i \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}, \\ D_{test} &= \mathbf{X}_* = \{\boldsymbol{x}_{*,i}\}_{i=1}^{n'}, \boldsymbol{x}_{*,i} \in \mathbb{R}^d. \end{aligned} \tag{8}$$

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

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right), \tag{9}$$

Table 2
Weibull and Lognormal distribution parameters derived from measurements in the North Sea (Haver and Winterstein, 2008), describing the significant wave height and peak wave period for a given sea state.

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

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

$$\bar{\mathbf{y}}_* = K(\mathbf{X}_*,\mathbf{X})[K(\mathbf{X},\mathbf{X})+\sigma_n^2 I]^{-1}\mathbf{y}, \tag{10}$$

$$\sigma^2 = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})[K(\mathbf{X},\mathbf{X})+\sigma_n^2 I]^{-1}K(\mathbf{X},\mathbf{X}_*). \tag{11}$$

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*. In the context of CFD data, epistemic uncertainty reflects the GP's lack of confidence in regions of the input space where CFD simulations are sparse or underexplored. On the other hand, the aleatoric uncertainty accounts for inherent numerical noise or variability in the CFD simulations, such as discretization errors, which are present even when the data is abundant.

3.2.1. 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 $\boldsymbol{x}$ and $\boldsymbol{x}'$ are close to each other, it is expected that $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ 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.

$$cov(f(\boldsymbol{x}), f(\boldsymbol{x}')) = k(\boldsymbol{x},\boldsymbol{x}') \tag{12}$$

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

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}')^2}{2l^2}\right) \tag{13}$$

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

$$\begin{aligned} k(\boldsymbol{x},\boldsymbol{x}') &= \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{n=1}^{d}\frac{(x_n - x_n')^2}{l_n^2}\right] \\ &= \sigma_f^2 \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^\top L^{-1}(\boldsymbol{x}-\boldsymbol{x}')\right) \end{aligned} \tag{14}$$

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

3.2.2. 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

$$\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top[K(\mathbf{X},\mathbf{X})+\sigma_n^2]^{-1}\mathbf{y} - \frac{1}{2}\log[K(\mathbf{X},\mathbf{X})+\sigma_n^2] - \frac{N}{2}\log 2\pi. \tag{15}$$

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

3.3. Principal component analysis for dimension reduction

In this study, the target quantity to predict is the mooring line force, $y(t)$, yet GPR is more effective in lower-dimensional spaces and not inherently designed for time series data. To address this issue, dimensionality reduction techniques are often employed in such cases to extract the most informative features from the data while discarding redundant information. A widely used technique is Principal Component Analysis (PCA), sometimes referred to as the Karhunen-Loève (KL) decomposition or Proper Orthogonal Decomposition (POD) (Gerbrands, 1981). PCA is a statistical method that transforms the data into a set of orthogonal components, known as principal components, that capture the directions with maximum variance within the dataset. By selecting the components that preserve the majority of the variance, PCA enables the representation of the data with fewer variables while preserving important patterns in the data.

At this stage, an essential step prior to developing the GPR model for predicting the mooring line force is the application of PCA to represent the mooring line force in a lower-dimensional space. Specifically, all the available data are integrated into the matrix $Y \in \mathbb{R}^{p\times m}$, where p represents the number of time steps, and m denotes the number of samples (i.e., available data from CFD simulations). The matrix is structured as follows:

$$Y = \begin{bmatrix} y_1(t_1) & \cdots & y_k(t_1) & \cdots & y_m(t_1) \\ y_1(t_2) & \cdots & y_k(t_2) & \cdots & y_m(t_2) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_1(t_p) & \cdots & y_k(t_p) & \cdots & y_m(t_p) \end{bmatrix} \tag{16}$$

where each column corresponds to each sample.

The matrix Y may be expressed as a product of three matrices via singular value decomposition (SVD) as

$$Y = USV^\intercal \tag{17}$$

where $U \in \mathbb{R}^{p\times p}$ is a unitary matrix containing the left singular values of Y. These vectors are the eigenvectors of the matrix $YY^\intercal$, and they represent the directions in the input space that maximize variance, known also as principal components. $S \in \mathbb{R}^{p\times m}$ is a rectangular diagonal matrix with the singular values of the Y which are the square roots of the eigenvalues of the matrix. The diagonal elements of S denoted as λ_i, represent the variance captured by each principal component. $V \in \mathbb{R}^{m\times m}$ is a unitary matrix containing the right singular vectors of Y, which are the eigenvectors of $Y^\intercal Y$.

The eigenvalues λ_i are sorted in descending order, and their corresponding eigenvectors U are arranged accordingly. The largest eigenvalue corresponds to the first principal component and accounts for the highest variance in the data, while each subsequent component captures progressively less variance. Dimensionality reduction is achieved by selecting a subset of principal components that preserve a sufficient amount of the total variance. This is quantified by the cumulative variance

$$\text{Cumulative variance} = \frac{\sum_{i=1}^{n}\lambda_i}{\sum_{j=1}^{r}\lambda_j} \geq 0.99, \tag{18}$$

where r is the rank of the matrix Y, representing the number of non-zero eigenvalues n is the number of retained PCA coefficients. Fig. 2 shows the cumulative variance with the number of PCA modes showing that it starts to converge for $n > 10$. In this study, based on a threshold of 99% cumulative variance, $n = 12$ components are retained.

Next, the data matrix Y is projected onto the retained principal components, i.e. the orthonormal eigenbasis defined by $u \in \mathbb{R}^{p\times n}$ (the

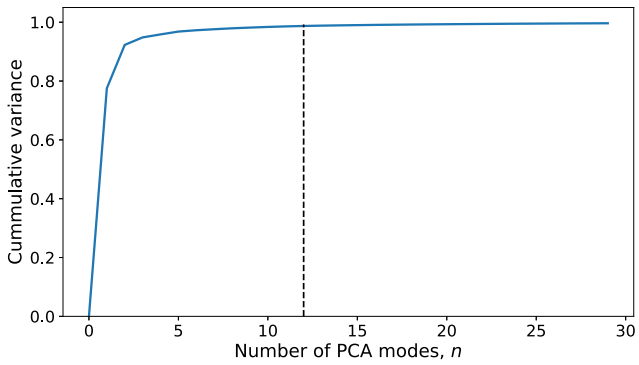


Fig. 2. Cumulative retained energy of PCA truncation order n, based on the eigenspectrum of the mooring force dataset. In this work, $n = 12$ is chosen to retain > 99% of the signal energy.

first n columns of U). The projection transforms the data into a new feature subspace and is performed as follows:

$$Q = Y^{\intercal} \cdot u \tag{19}$$

where $Q \in \mathbb{R}^{m \times n}$ matrix that contains the principal component coefficients for each sample, effectively reducing the dimensionality from p to n.

Having the PCA coefficients, the mooring force time series can be reconstructed using the following expression:

$$y_{rec,k}(t) = \sum_{i=1}^{n} q_{k,i} \cdot u_i(t), \quad t \in [0, t_p]. \tag{20}$$

where $q_{k,i}$ is the ith PCA coefficient for the kth sample from the matrix Q, and u_i is the corresponding ith eigenvector (principal component). The $y_{rec,k}(t)$ closely approximates the original force $y(t)$, depending on the proportion of energy preserved during the dimensionality reduction process. When 99% of the energy is retained, the reconstructed force is expected to be nearly identical to the original force obtained from the CFD simulations.

3.4. 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 Fig. 3. 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.

In Fig. 3, 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 (Fig. 4) without facing the issue of gradients 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. Eqs. (21)–(27) show the mathematics at each gate and the definition of the new long- and short-memories.

$$\text{Forget gate:} \quad f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right) \tag{21}$$

$$\text{Input gate:} \quad i_t = \sigma\left(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i\right) \tag{22}$$

$$\tilde{c}_t = \tanh\left(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c\right) \tag{23}$$

$$\text{Cell state output:} \quad c_t = i_t * \tilde{c}_t + f_t * c_{t-1} \tag{24}$$

$$\text{Output gate:} \quad o_t = \sigma\left(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o\right) \tag{25}$$

$$\tilde{h}_t = \tanh\left(c_t\right) \tag{26}$$

$$\text{Hidden layer output:} \quad h_t = o_t * \tilde{h}_t \tag{27}$$

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.

The above description refers to a single LSTM unit that accepts as input the value for one time step. As shown in Fig. 4, to make a prediction based on a sequence of inputs, $x(t_1, t_2, \ldots, t_n)$, 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, $y_L(t)$, which is the LSTM predicted mooring line force,

$$y_L(t) = W \cdot h_t + b \tag{28}$$

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

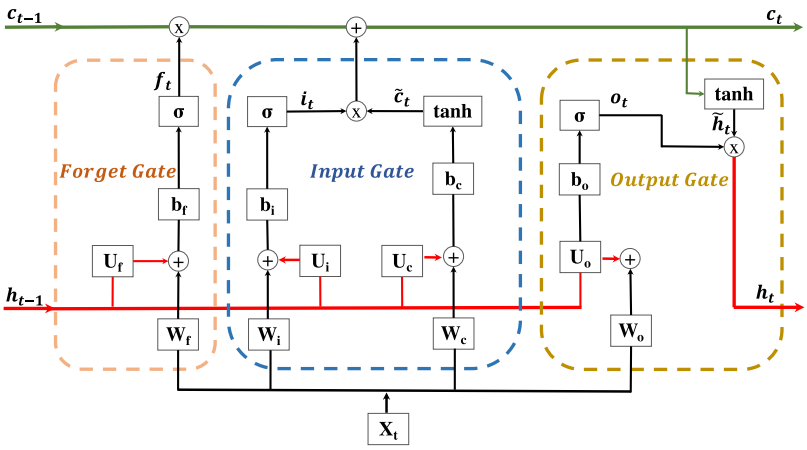


Fig. 3. The structure inside an LSTM unit.

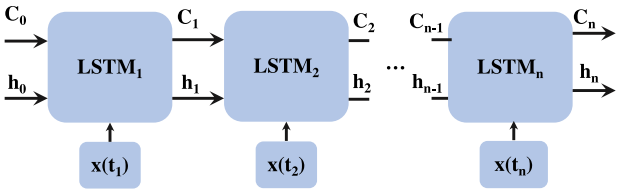


Fig. 4. An LSTM layer that consists of n LSTM units.

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, tailored for this specific application.

3.4.1. 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

$$\mathrm{MAE} = \frac{1}{b}\frac{1}{p}\sum_{i=1}^{b}\sum_{j=1}^{p}|y_L(t_j) - y(t_j)|_i \tag{29}$$

where y_L is the LSTM prediction and y is the CFD solution. As it is explained in Appendix in "LSTM hyperparameters selection", for LSTM training purposes the datasets is split into b batches and each batch comprises of p time steps. The error between the predicted and the real output is computed for each time step, t, and the batch-average error is calculated. Finally, the *overall loss function* MAE is the average of the batch-average errors.

3.5. Active learning

A key question that arises when developing machine-learned surrogate models is how to strategically select training samples to achieve two main goals: ensuring accurate predictions, particularly for extreme events, and addressing the practical challenges of acquiring training data. Traditional sampling approaches rely on querying the physical system (i.e., the black-box function) at a large number of input points. However, this method is highly resource-intensive, requiring extensive high-fidelity simulations and/or physical experiments, making it impractical due to significant time and computational demands.

This section introduces the active learning scheme within Bayesian experimental design to effectively construct a surrogate model. Fig. 5 provides a high-level overview of the active learning process. The surrogate model is iteratively trained on sequentially selected datasets of input–output pairs, denoted as $\mathcal{D}_{t-1} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})\}$. In the left-side plot of Fig. 5, significant uncertainty is observed between points x_1 and x_2, as indicated by the region enclosed by the dashed lines. Additionally, the high values of $y(x)$ in this region emphasize the need to develop a model with high prediction accuracy specifically in this area. The active learning algorithm determines the next-best sample by actively identifying and selecting the most informative training point from the input parameter space, $x \in \mathcal{X}$. This point is chosen to minimize the surrogate model's uncertainty while improving the prediction of extreme values. The new sample is added to the training dataset, resulting in $\mathcal{D}_t = \{\mathcal{D}_{t-1} \cup (x_t, y_t)\}$ and the surrogate model is retrained. The active learning process continues for a number of iterations t_{iter}, progressively improving model accuracy and reducing uncertainty in the targeted regions. This strategy ensures that each data point contributes maximally to the learning process, effectively addressing the practical limitation of acquiring large datasets—a common challenge in computationally expensive simulations such as CFD. The pseudoscope of the active learning procedure is provided in Algorithm 1.

4. Hybrid surrogate model via baseline approach

Fig. 6 illustrates the architecture of the hybrid surrogate model, which combines GPR and LSTM neural networks. The model establishes a mapping between the input vector $\boldsymbol{x} = [H_s, T_p, D_{PTO}, K_{es}]$ and the mooring force time series. As described in Section 3.3, to facilitate the development of GPR models, PCA is first applied to the available mooring force data $y(t)$ (from CFD simulations) in order to reduce the dimensionality to a finite set of PCA coefficients $\{q_i\}_{i=1}^{12}$. Each GPR model is trained to output $\bar{q}_i = \mathbb{E}[q_i|\boldsymbol{x}]$, representing the mean of the posterior distribution. The predicted PCA coefficients $\bar{q}_i$ from all the GPR models are then combined using Eq. (20) to reconstruct the mooring line force $y_G(t)$, which serves as the *indirect* GPR prediction. Next, the LSTM receives the reconstructed force $y_G(t)$ and outputs a refined prediction $y_L(t)$, which more closely matches the actual CFD solution $y(t)$.

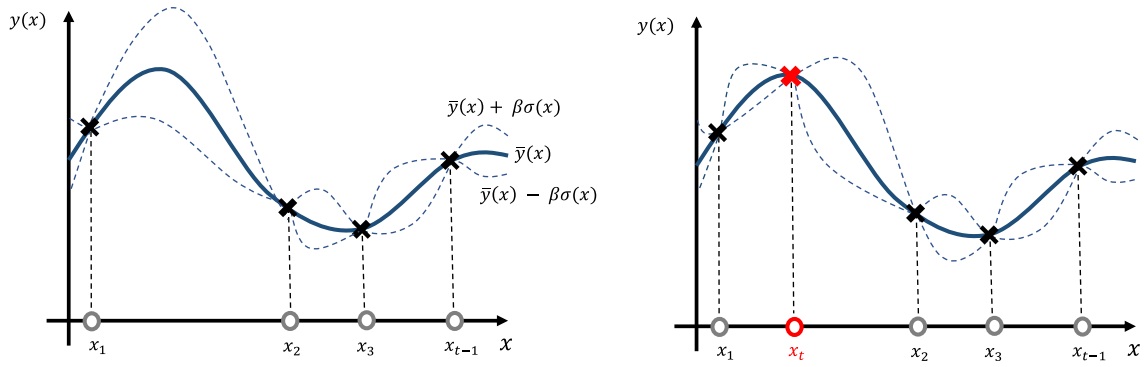


Fig. 5. Estimation of the output value, $\bar{y}(x)$, for any input, x, using the developed surrogate model, with uncertainty limits represented by dashed lines. **Left**: The surrogate model is constructed using the dataset $D_{t-1} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})\}$. **Right:** The active learning process identifies an optimal new sample, x_t, which reduces the surrogate model's uncertainty and improves the accuracy of extreme value predictions.

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

The model is developed using a baseline approach, meaning no sampling method is employed to select the training data. Instead, it is trained directly on available CFD data from a prior study, consisting of 73 datasets–65 allocated for training and the remainder for validation. The objective of this approach is to rigorously evaluate the performance of the hybrid modeling framework illustrated in Fig. 6 for predicting complex time series dynamics. Specifically, it aims to determine whether the combination of GPR and LSTM improves prediction accuracy.

5. Hybrid surrogate model via active learning

The objective is to develop the hybrid surrogate model illustrated in Fig. 6, this time incorporating the active learning scheme as outlined in Section 3.5. This approach strategically selects the training datasets, moving beyond the limitations of relying solely on pre-existing data. Fig. 7 provides a schematic representation of the methodological framework guiding the development of the hybrid surrogate model through the active learning process.

1. **Initialization:** The GPR components of the hybrid surrogate model is pre-train using a small, randomly selected dataset, $\mathcal{D}_0 = \{\boldsymbol{x}_i, y_i\}_{i=1}^{t_{init}}$, where $t_{init} = 5$. These initial samples are obtained from the physical system (ground truth).
2. **Active Learning:** The GPR model is further trained using the active learning process for effective sample selection. (i) The GPR model evaluates a set of random input samples, $\{\boldsymbol{x}_i\}_{i=1}^{s}$ and predicts the corresponding mooring line force, $\{y_{G,i}\}_{i=1}^{s}$. Further explanation of this step is provided in Section 5.1. (ii) From all these samples, the acquisition function selects the next-best input sample $\boldsymbol{x}^*$. The selection satisfies a specific condition, such as identifying regions in the input parameter space where extremes are more probable to occur, while minimizing prediction uncertainty. Further description is provided in Section 5.2. (iii) The physical system (ground truth) is then employed to evaluate the sample $\boldsymbol{x}^*$ providing the corresponding output, y^*. Section 5.3 explains the ground truth model considered at this step. (iv) The new input–output pair $\{\boldsymbol{x}^*, y^*\}$ is added to the existing training dataset. (v) The GPR model is retrained on the updated dataset. Steps (i) - (v) repeat for a user-defined number of iterations, $t_{iter} = 80$. At the last iteration of the active learning process, the GPR model has been trained on the samples strategically selected through active learning $\mathcal{D}_{\mathcal{G}} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{t_{iter}}$.
3. **LSTM training:** The LSTM model is subsequently trained on the dataset $\mathcal{D}_{\mathcal{L}} = \mathcal{D}_0 \cup \mathcal{D}_{\mathcal{G}}$. Notably, the LSTM model takes as input the GPR posterior mean force, $y_G(t)$, and outputs the refined prediction $y_L(t)$, which closely approximates the response of the physical system.
4. **Prediction:** The hybrid surrogate model is able to predict the mooring line force for any input vector, $\boldsymbol{x}$, with computation time orders of magnitude faster than CFD simulations. Specifically, it is employed to predict the mooring force for thousands of input scenarios $\boldsymbol{x}$.
5. **Statistics evaluation:** Having the model's prediction for all these input scenarios, the extreme load statistics can be effectively reconstructed. Further description is provided in Section 5.4.

It is important to note that the overarching goal remains to construct a hybrid surrogate model, as depicted in Fig. 6. In this framework, the GPR model predicts the PCA coefficients, and the mooring line force, $y_G(t)$, is reconstructed using Eq. (20). To achieve this, the GPR model is trained on the PCA coefficients derived from the available $y(t)$ samples at each stage of the process.

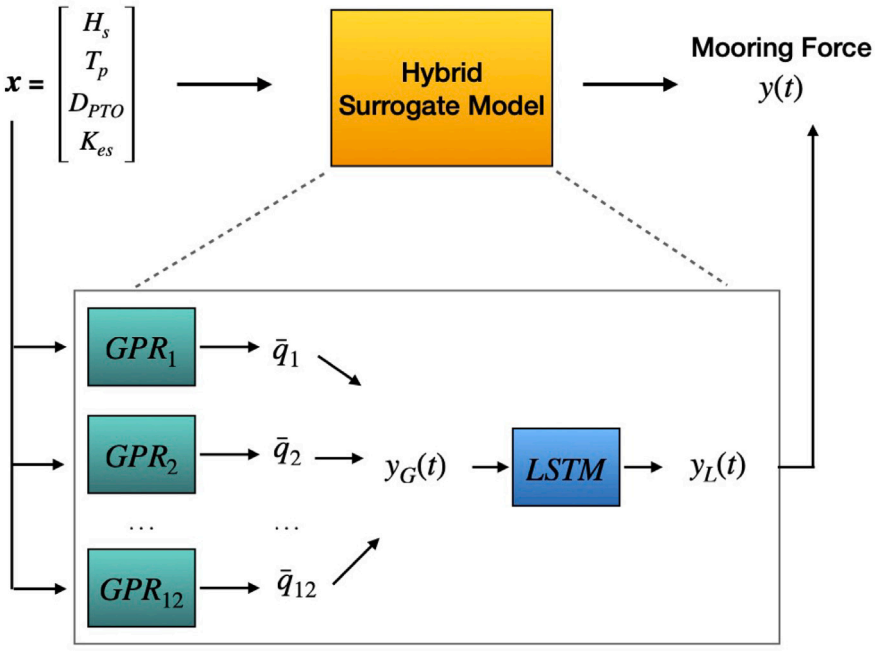


Fig. 6. The hybrid surrogate model establishes a mapping between the input parameters $\boldsymbol{x}$, which represent wave characteristics and the design parameters of the wave energy device, and the mooring line force, $y(t)$. The term "hybrid" refers to the combination of GPR and LSTM neural networks. To implement GPR, the dimensionality of the force time series is reduced using PCA, representing the data with 12 PCA coefficients, which leads to the development of 12 separate GPR surrogate models. The predicted GPR posteriors, $\{\bar{q}_i(\boldsymbol{x})\}_{i=1}^{12}$, are then used to reconstruct the mooring line force, $y_G(t)$. The latter serves as the input for the LSTM model, which further refines the prediction, $y_L(t)$, bringing it closer to the actual force.

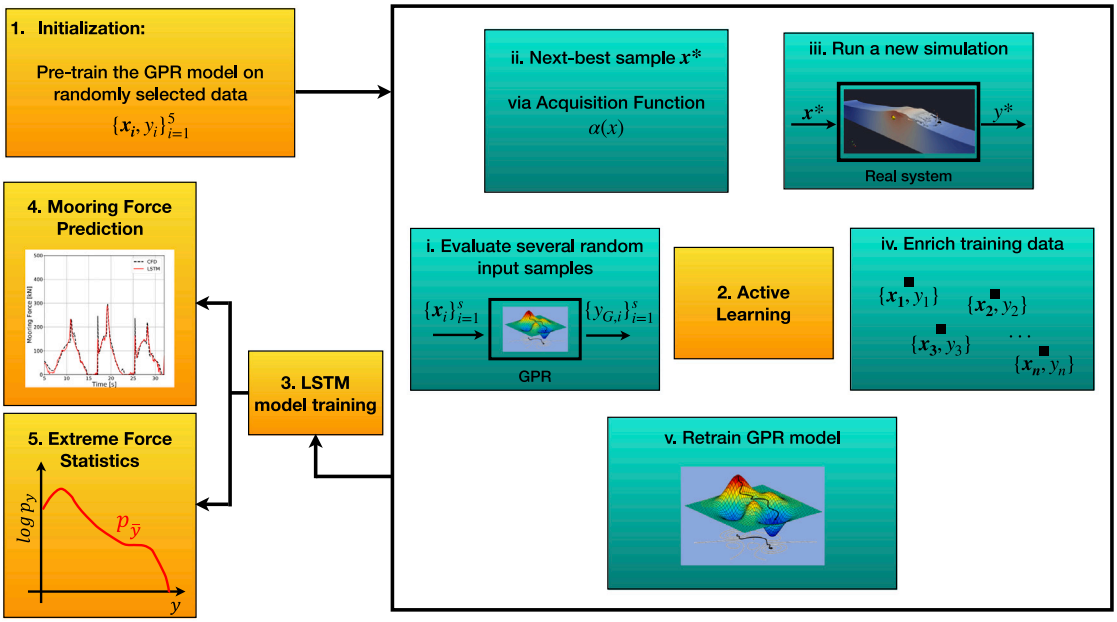


Fig. 7. 1. Initialization: The GPR surrogate model is pre-trained on a few random samples and learns a sparse representation of the underlying system. 2. Active learning strategy: (i) The current GPR model evaluates a big number of Latin Hypercube samples for discovering the regions of the parameter 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.

Additionally, the selection of appropriate acquisition functions is important for an effective surrogate model, thus the acquisition functions discussed in Section 5.2 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. Acquisition function prerequisites

At the first step of the active learning loop, a large number of input samples is evaluated through the GPR surrogate model, and the sample which satisfies the acquisition function condition (see Section 5.2) is selected as the next-best sample.

At each iteration, acquisition samples are generated via Latin Hypercube Sampling (LHS) from the input domain $x_{\mathcal{LHS}} = \{x_1, \dots, x_{100}\}$, 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, $y_G(x_{\mathcal{LHS}})$, and variances, $\sigma^2(x_{\mathcal{LHS}})$. The next-best point $x^* \in x_{\mathcal{LHS}}$ is the one that satisfies

$$x^* = \underset{x^* \in x_{\mathcal{LHS}}}{\operatorname{argmax}}\, a(x) \tag{30}$$

where $a(x)$ is any acquisition function defined in Section 5.2. As it is demonstrated in the following section, the acquisition function incorporates variances, $\sigma^2(x_{\mathcal{LHS}})$, the probability distribution of $y_{\mathcal{G}}$, $p_{y_G}(y)$. The latter is estimated using the kernel density estimator (KDE) technique, which applies a weighted factor based on the probability of the input parameters $p_x(x_{\mathcal{LHS}})$, as discussed in Section 2.1.

$$p_{y_{\mathcal{G}}}(y) = \text{KDE}(\text{data} = y_G(x_{\mathcal{LHS}}), \text{weight} = p_x(x_{\mathcal{LHS}})). \tag{31}$$

Algorithm 2 describes precisely the steps for the definition of the $p_{y_{\mathcal{G}}}$ which is called importance distribution. The $p_{y_{\mathcal{G}}}$ is estimated at every iteration in Algorithm 1.

5.2. Acquisition function

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,

$$\alpha_{US}(x) = \sigma^2(x). \tag{32}$$

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 (Blanchard and Sapsis, 2021a). 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 (Sapsis, 2020; Blanchard and Sapsis, 2021a). In practice, the LW-US is the US acquisition function multiplied by the likelihood-ratio, $w(x)$,

$$\alpha_{US-LW}(x) = \sigma^2(x) w(x). \tag{33}$$

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

$$w(x) = \frac{p_x(x)}{p_{y_{\mathcal{G}}}(y_{\mathcal{G}}(x))}, \tag{34}$$

where $p_{y_{\mathcal{G}}}$ denotes the GP posterior distribution mean, y_G, estimated using Eq. (31), and p_x is the PDF of the input parameters, x, with more details provided in Section 2.1. The p_x is usually referred to as *nominal distribution* and $p_{y_{\mathcal{G}}}$ *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_{y_G}). Conversely, for samples that have similar probability for extreme/rare events (same p_{y_G}), it promotes those with higher probability of occurrence (larger p_x).

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

1: **Input:** GPR model and nominal distribution, $p_x(x)$.
2: **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 $\{y_{G_k}\}_{k=1}^M$.
5: Perform KDE on the outputs, $\{y_{G_k}\}_{k=1}^M$, and obtain the PDF $p_{y_{\mathcal{G}}}$.
6: For a particular input-output pair $(x, y_G(x))$, the corresponding density value, $p_{y_G}(y_G(x))$, is defined by the PDF constructed at the previous step.
7: **end do**
8: **return** Importance distribution $p_{y_G}(y_G(x))$.

5.3. Ground truth

At each iteration of the active learning process, the physical system is evaluated at the next-best sample, $\boldsymbol{x}^*$, yielding the corresponding output, y^*. In this study, while high-fidelity CFD simulations are intended to serve as the physical system, computational resource constraints necessitate the adoption of an alternative approach. Specifically, the hybrid surrogate model developed in Section 4 is used as the ground truth. As demonstrated in Section 6.1, it effectively replicates CFD simulation results and significantly enhances computational efficiency. The key takeaway is that users can choose any model they trust as the ground truth, provided it reliably represents and assimilates the real system.

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 (Mohamad and Sapsis, 2018). 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 force, y_{max}, which corresponds to the second peak in the force time series.

$$y_{max} = \max_{t \in [0,T]} |y_{\mathcal{L}}(t)|. \tag{35}$$

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 2.1. Next, the input samples are pushed into the GPR model which provides the initial

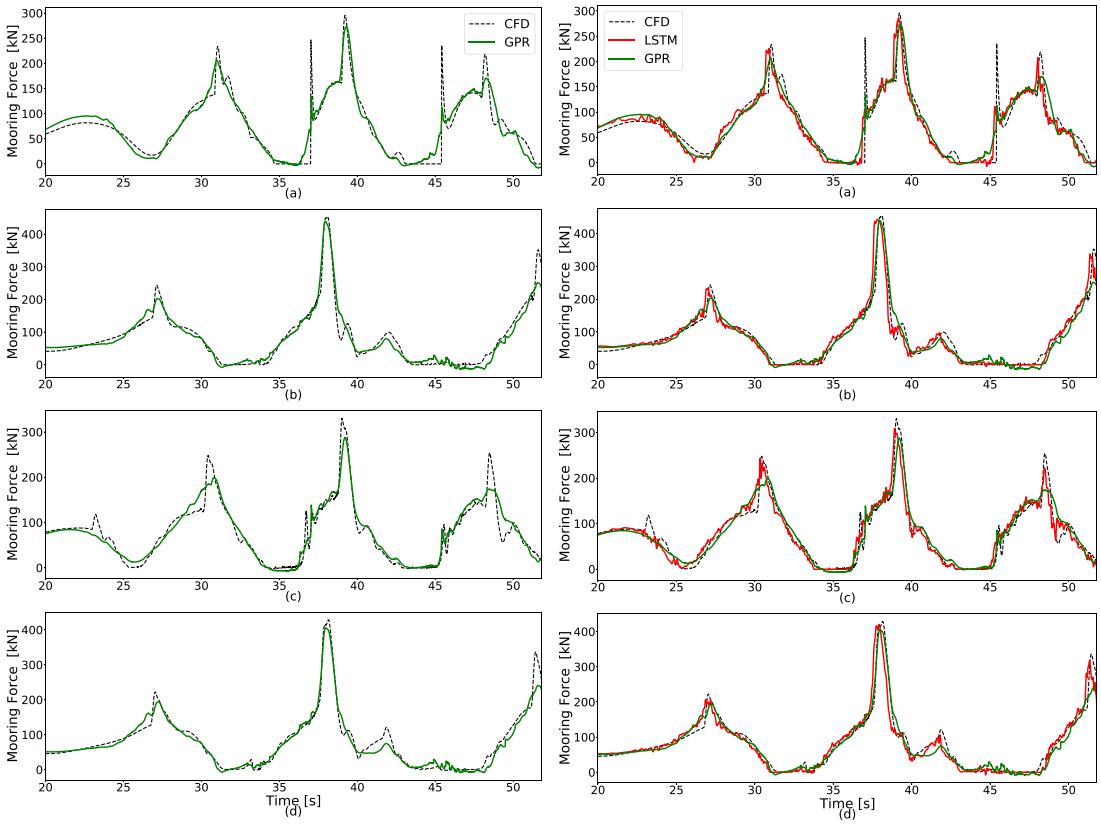


Fig. 8. **[Left]** Comparison of the GPR-predicted mooring force (green) with the CFD solution (dashed black) for four validation cases. **[Right]** Enhanced prediction accuracy in capturing the three main peaks achieved by incorporating the LSTM model (red).

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

$$e(t) = \int |\log_{10} p_{s_t}(y) - \log_{10} p_y(y)| dy \tag{36}$$

Both PDFs are estimated on 10^5 samples via Monte Carlo design (see Sections 5.3 and 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 is analyzed in this section. The GPR component of the hybrid surrogate maps the input vector, $\boldsymbol{x}$, to the conditional posterior mean of each PCA coefficient, $\{q_i\}_{i=1}^{12}$. Subsequently, the PCA coefficients are combined using Eq. (20) to construct the force time series. Fig. 8 (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.

As this study focuses on extremes, it is important to accurately capture those peaks. This motivates the inclusion of the LSTM component in the hybrid surrogate model. The LSTM neural network takes the mooring force time series generated by the GPR model as input and outputs an improved prediction. Fig. 8 (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.

The prediction accuracy in the peaks is quantitatively assessed using the coefficient of determination, R^2, which compares the surrogate model predictions to the CFD solution. The results are presented in Fig. 9 in the form of scatter plots. 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.

To assess the predictive performance of the hybrid surrogate model for the entire force time series (not just the peaks), Fig. 10 compares the force spectral density obtained from the GPR and LSTM machine learning models with the CFD solution. 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

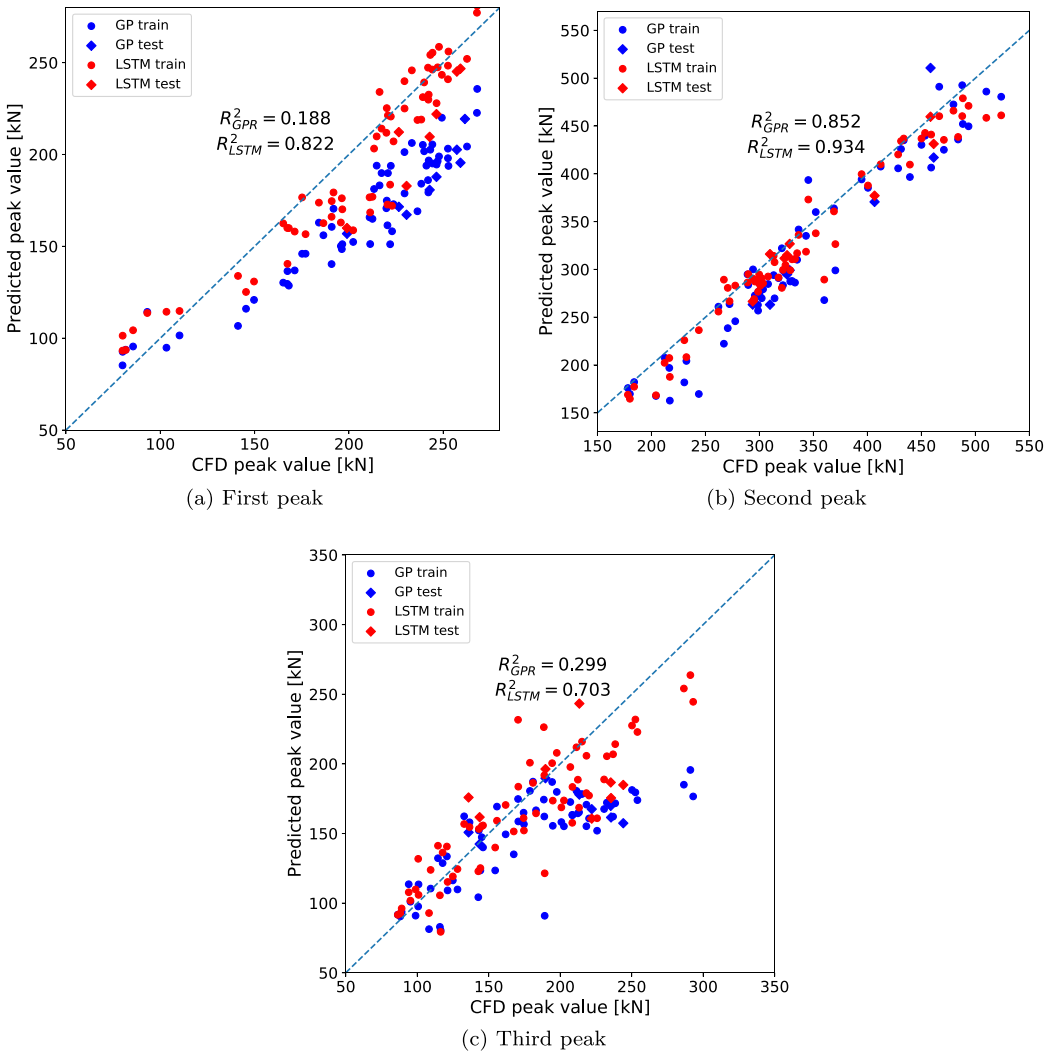


Fig. 9. Comparison of the three main peaks predicted by the GPR and LSTM models against the CFD solution, quantified by the coefficient of determination R^2. The addition of the LSTM model significantly improves the R^2 value.

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

$$l_1 = \int_{f_l}^{f_u} |s_{surrog}(f) - s_{CFD}(f)| df, \tag{37}$$

where s_{surrog} and s_{CFD} represent the predicted and CFD force spectral densities, respectively, and f denotes the frequency. Fig. 11 provides a summary of the l_1 norm metric for the eight validation samples. It can be observed that the error l_1 is generally reduced after the LSTM model is implemented, indicating improved accuracy in the prediction of the force spectral density.

6.1.1. LSTM hyperparameters selection

To create the optimal LSTM architecture, it is necessary to select the values for several hyperparameters, i.e., time series resolution, time-window, number of hidden units, batch size, number of epochs. In this study, a procedure similar to Wang et al. (2022) 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 hyperparameter *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. To accelerate the LSTM training process, it is examined the option to consider a coarser time series. However, a critical point is that the LSTM model should be still able to capture the extreme forces, which are the three main peaks in the time series, as highlighted by the circles in Fig. 18(b). In this study, two resolutions are examined, i.e., dt = 0.02 s, 0.04 s. In Fig. 18, the LSTM prediction is compared to the CFD solution for both resolutions and it is observed that the coarser resolution (dt = 0.04 s) underestimates the three main peaks while the finer (dt = 0.02 s) provides better accuracy. While time resolutions finer than 0.02 s could improve accuracy and capture detailed phenomena, such as the second-order instantaneous forces at$t$ = 17 s and t = 25.5 s, these effects are rarely observed in our sample data. Therefore, a resolution of 0.02 s provides a suitable balance between prediction accuracy and the computational cost of training the LSTM model.

2. Time-window: The total training data is divided into shorter sequences which are called time-windows. Each sub-sequence consists of

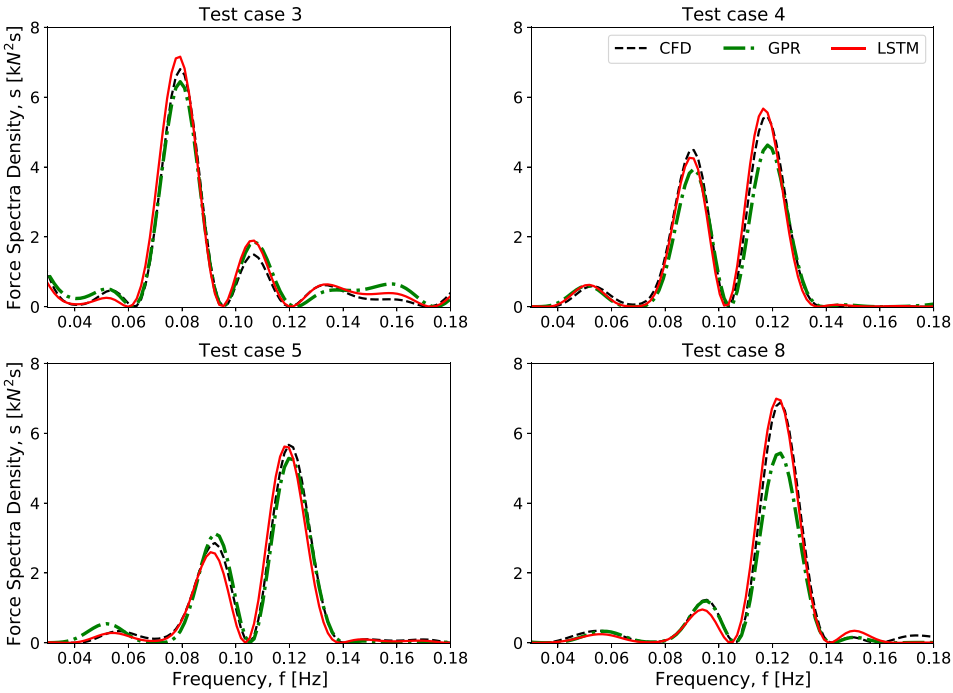


Fig. 10. Comparison of the force spectra density obtained from the GPR model (green) and the hybrid model which incorporates the LSTM component (red) against the CFD solution (black).

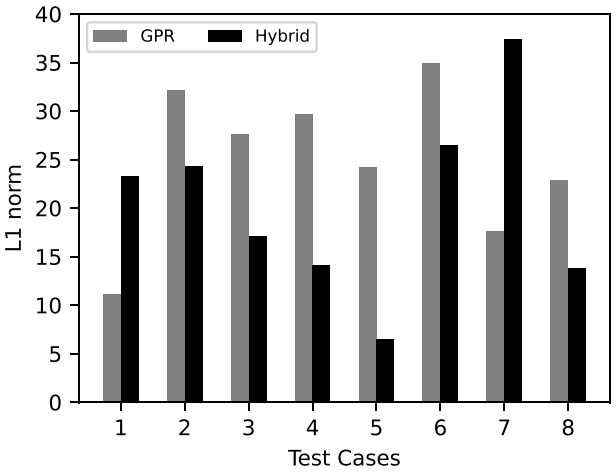


Fig. 11. Summary of the error between the predicted and actual force spectral density. The predictions are obtained using the GPR model (gray) and the hybrid model which incorporates the LSTM component (black). The error evaluation is based on the L_1 norm metric.

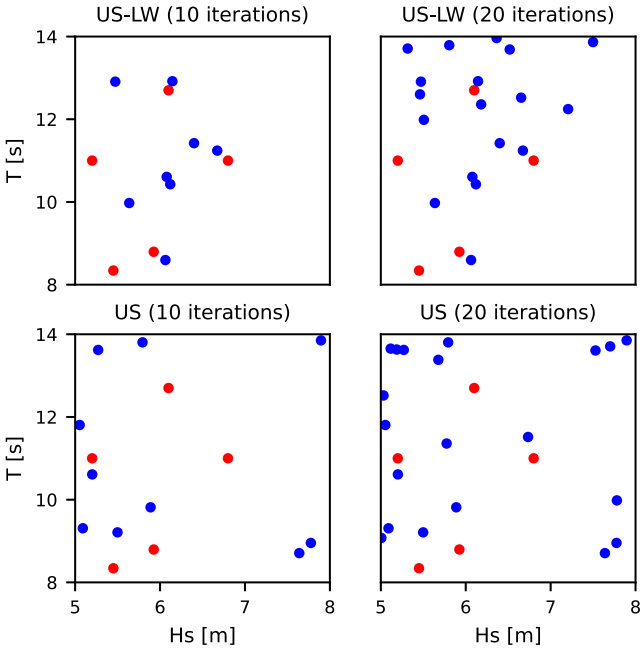


Fig. 12. Progression of the sampling algorithm for US-LW (top row) and US (bottom row) after 10 iterations (left column) and 20 iterations (right column). The red dots represent the five initial random samples, while the blue indicate the points visited by the active learning algorithm.

the number of time-steps that the LSTM model considers for predicting the next time-step value. In this stage, it is important to characterize 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 Fig. 19(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 3.4.

For example, for an input vector $\boldsymbol{x}$ with dimensions $d\times1$, 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 $h\times1$. 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 Fig. 19(b). Finally, 64 hidden

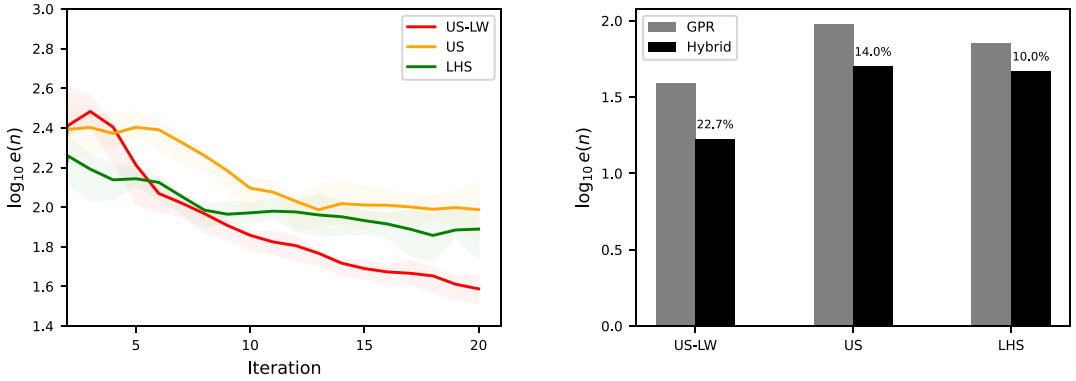


Fig. 13. **[Left]** Log-PDF error per AS iteration for US-LW, US, and LHS, representing the error computed as the difference between the ground truth PDF and the mean PDF reconstructed from the GPR predictions across an ensemble of 15 experiments. The error bands indicating half of the median absolute deviation. **[Right]** The final hybrid surrogate model is compared to the 20th iteration GPR model demonstrating the improvement achieved by incorporating the LSTM component.

Table 3
Summary of the optimal LSTM hyperparameters.

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

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 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 3.4.1). Fig. 20 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 (Kingma and Ba, 2014). Adam uses adaptive learning rate for faster convergence and takes the initial value of 0.001.

The selected LSTM hyperparameters are summarized in Table 3.

6.1.2. 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 h. However, once the LSTM model is trained, it provides the output in approximately 1-2 min.

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. Hybrid surrogate model built on active learning

In this section, the hybrid surrogate model, developed using the active learning process described in Section 5, is evaluated. A key component of active learning is the acquisition function, which guides the sample selection process. Two acquisition functions, US-LW and US, as detailed in Section 5.2, are considered. Additionally, the LHS method – a widely used technique for training data selection – is employed to compare the performance of the active learning approach against a conventional sampling method.

Sample 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 Fig. 12, 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 Blanchard and Sapsis (2021a). 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

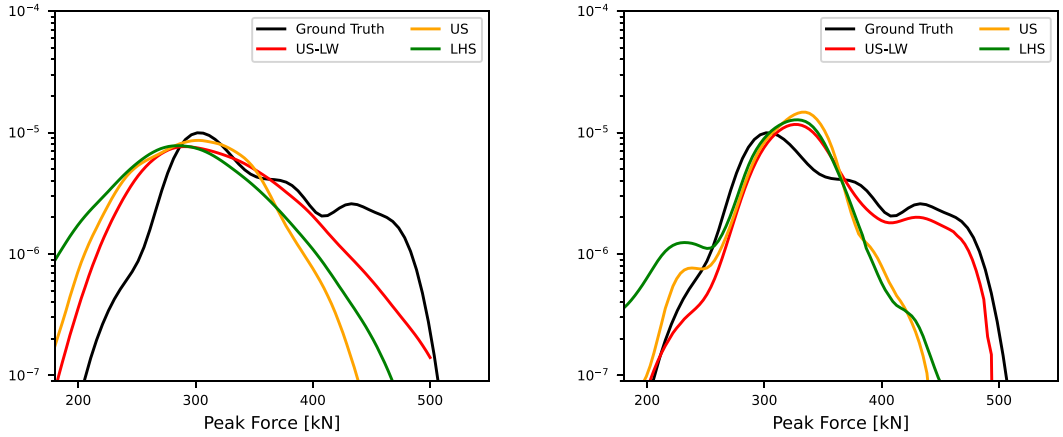


Fig. 14. PDF of the maximum mooring force as estimated from **[Left]** the 20th iteration GPR model and **[Right]** the hybrid surrogate model, using US-LW and US aquisition functions, as well as LHS. The black color indicates the ground truth PDF.

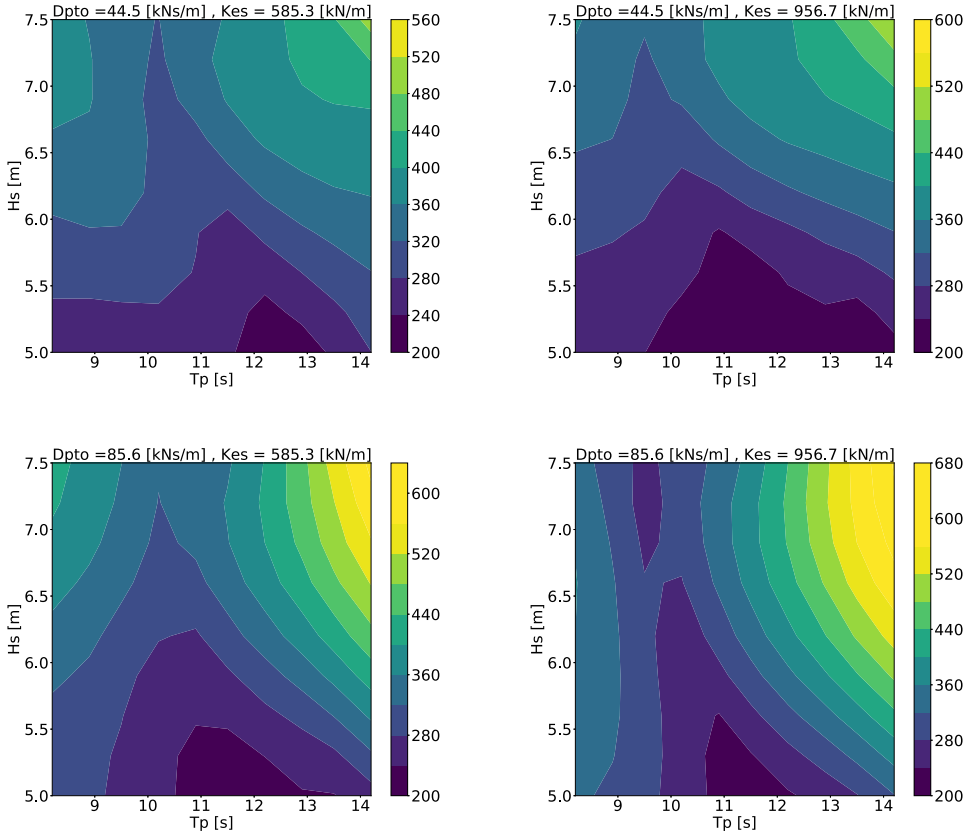


Fig. 15. Contour plots representing the maximum mooring force, with the colorbar indicating values in [kN]. Each subplot illustrates force variations as a function of wave characteristics (H_s, T_p), while keeping other input parameters (D_{PTO}, K_{es}) constant.

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. Fig. 13 (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 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. Fig. 13 (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,

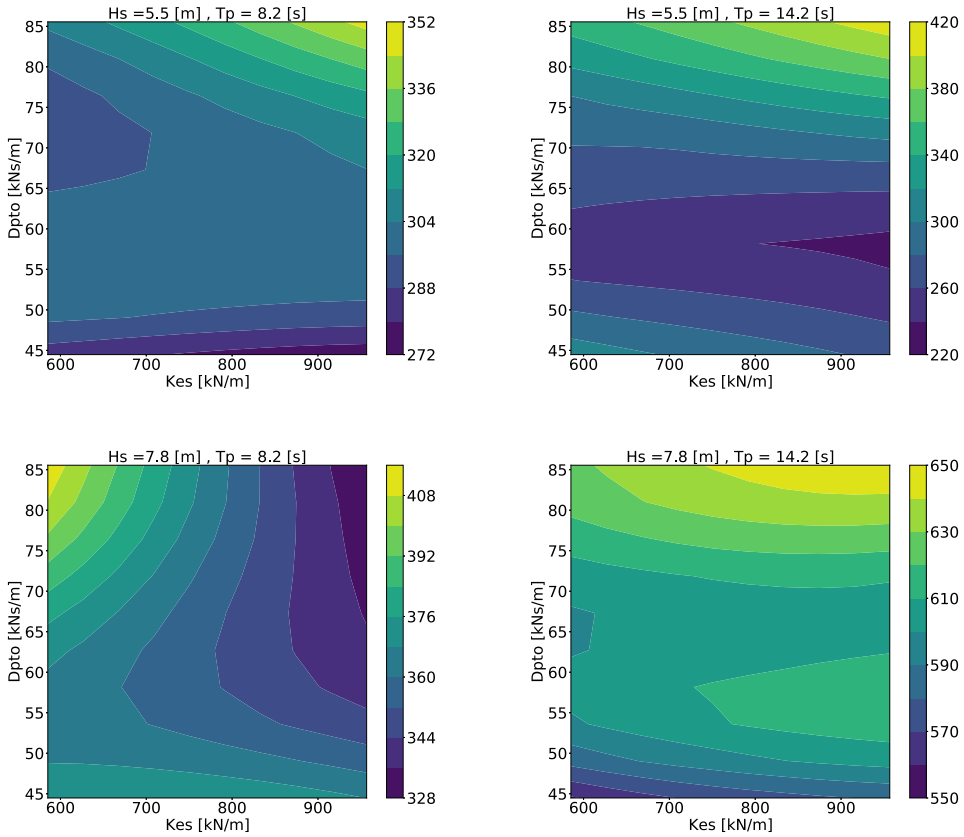


Fig. 16. Contour plots depicting the maximum mooring force, with the colorbar indicating values in [kN]. Each subplot illustrates force variations as a function of PTO parameters (D_{PTO}, K_{es}), while keeping wave characteristics (H_s, T_p) constant.

and the introduction of the LSTM component further reduces the error by 22.7%.

6.3. Applications

6.3.1. Extreme force statistics

The main goal of this study is to develop a robust methodology for accurately quantifying the statistical distribution of maximum mooring forces. High-fidelity modeling techniques, such as CFD simulations, are impractical for this task due to their significant computational demands. Instead, a hybrid surrogate model, developed following the active learning approach, enables this by effectively evaluating thousands of Monte Carlo samples.

The extreme mooring force PDF reconstruction follows the procedure outlined in Section 5.4. Fig. 14 [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, Fig. 14 [Right] displays the PDF generated by the hybrid surrogate model, comprising the 20th iteration GPR model and an LSTM neural network. This combination significantly enhances the quality of the PDF, underscoring the significance of the hybrid model. Furthermore, Fig. 14 illustrates that the US-LW acquisition function consistently outperforms alternative methods.

6.3.2. Interpretation of model's mapping

Figs. 15 and 16 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, Fig. 15 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 Fig. 15, 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.

Fig. 16 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 Rodríguez et al. (2019). 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 Katsidoniotaki et al. (2021), Katsidoniotaki and Göteman (2022).

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 Pickering et al. (2022), Blanchard and Sapsis (2021b), its application to realistic scenarios, such as a wave energy system subjected to extreme waves, is clearly demonstrated. In this study, the hybrid surrogate model is constructed combining two machine learning methods, the 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 parameters 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.

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. 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 data, ensuring less high-fidelity data is necessary for training purposes.

CRediT authorship contribution statement

Eirini Katsidoniotaki: Writing – original draft, Software, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stephen Guth:** Writing – review & editing, Methodology, Formal analysis. **Malin Göteman:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Themistoklis P. Sapsis:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research in this paper was supported by the Swedish Centre of Natural Hazards and Disaster Science (CNDS), the Onassis Foundation, Principality of Liechtenstein (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, United States grant N00014-21-1-2357.

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

Appendix

Extreme wave representation

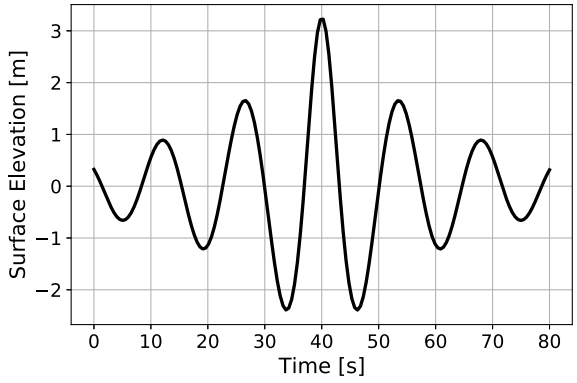
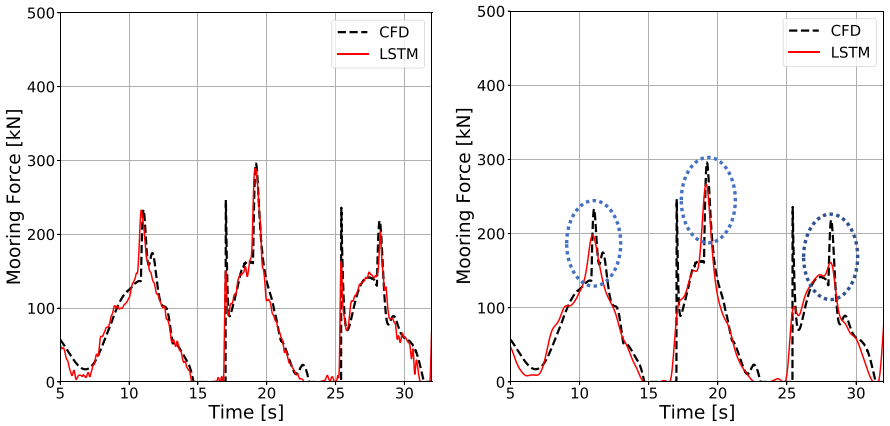


Fig. 17. Focused Wave Profile: Wave components focusing at $t = 40$ s with their amplitudes determined by NewWave Theory.

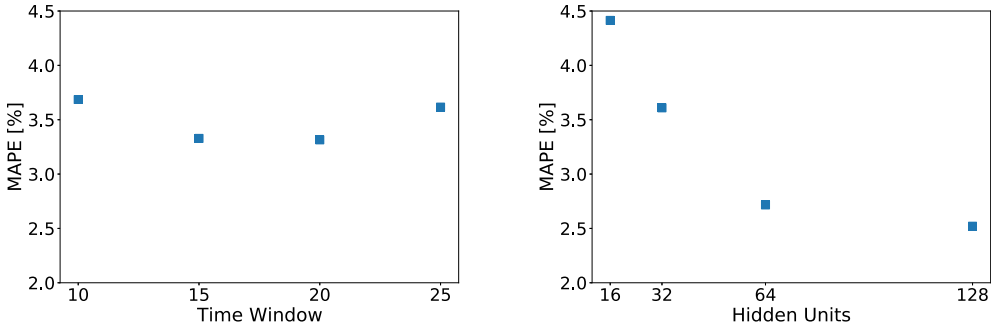
The waves are chosen along the 50-year environmental contour for a potential installation site in the North Sea (Wrang et al., 2021). Each wave is characterized by the significant wave height (H_s) and peak wave period (T_p). The numerical wave realization, typically represented as a 30-minute to 3-hour irregular wave episode, demands substantial computational resources in the context of CFD. Instead, as detailed in Katsidoniotaki et al. (2020), 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 (Tromans et al., 1991). 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 (Fig. 17). 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.



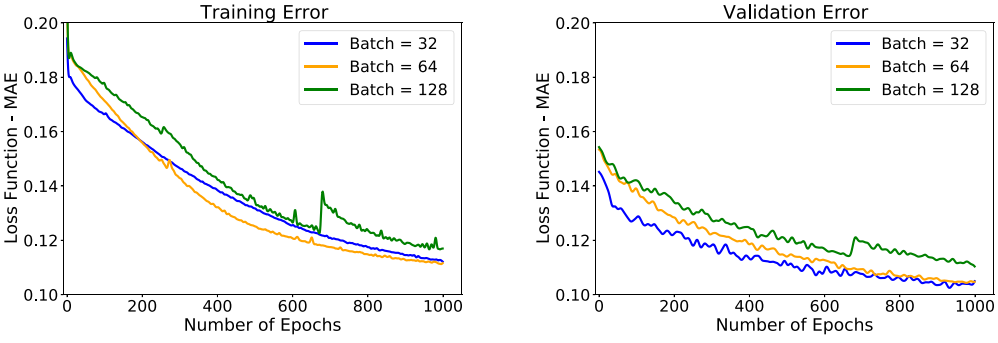
(a) Resolution 0.02 s (b) Resolution 0.04 s

Fig. 18. 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.



(a) Time Window (b) Hidden Units

Fig. 19. Selecting hyper-parameters for the LSTM model.



(a) Loss function error in training data (b) Loss function error in validation data

Fig. 20. Selecting batch size hyper-parameter for the LSTM model based on the loss function mean absolute error (MAE).

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.

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 (Fan and Anglart, 2019). 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 (Katsidoniotaki and Göteman, 2022). 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 Katsidoniotaki and Göteman (2022), Katsidoniotaki et al. (2021, 2022c). 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 (Katsidoniotaki et al., 2022c).

References

Anderlini, E., 2017. Control of Wave Energy Converters using Machine Learning Strategies (Ph.D. thesis). University of Edinburgh.

Babarit, A., Hals, J., Muliawan, M.J., Kurniawan, A., Moan, T., Krokstad, J., 2012. Numerical benchmarking study of a selection of wave energy converters. Renew. Energy 41, 44–63.

Blanchard, A., Sapsis, T., 2021a. Bayesian optimization with output-weighted optimal sampling. J. Comput. Phys. 425, 109901.

Blanchard, A., Sapsis, T., 2021b. Output-weighted optimal sampling for Bayesian experimental design and uncertainty quantification. SIAM/ ASA J. Uncertain. Quantif. 9 (2), 564–592.

Bosma, B., Zhang, Z., Brekken, T.K., Özkan-Haller, H.T., McNatt, C., Yim, S.C., 2012. Wave energy converter modeling in the frequency domain: A design guide. In: 2012 IEEE Energy Conversion Congress and Exposition. ECCE, IEEE, pp. 2099–2106.

Clément, A., McCullen, P., Falcão, A., Fiorentino, A., Gardner, F., Hammarlund, K., Lemonis, G., Lewis, T., Nielsen, K., Petroncini, S., et al., 2002. Wave energy in Europe: current status and perspectives. Renew. Sustain. Energy Rev. 6 (5), 405–431.

Czech, B., Bauer, P., 2012. Wave energy converter concepts: Design challenges and classification. IEEE Ind. Electron. Mag. 6 (2), 4–16.

Dhanak, M.R., Xiros, N.I., 2016. Springer handbook of ocean engineering. Springer.

Fan, W., Anglart, H., 2019. On the closure requirement for VOF simulations with RANS modeling. arXiv preprint arXiv:1911.09727.

Gerbrands, J.J., 1981. On the relationships between SVD, KLT and PCA. Pattern Recognit. (ISSN: 0031-3203) 14 (1), 375–381, 1980 Conference on Pattern Recognition.

Guo, B., Ringwood, J.V., 2021. A review of wave energy technology from a research and commercial perspective. IET Renew. Power Gener. 15 (14), 3065–3090.

Guth, S., Champenois, B., Sapsis, T.P., 2022. Application of Gaussian process multi-fidelity optimal sampling to ship structural modeling. In: 34th Symposium on Naval Hydrodynamics Proceedings Washington DC 2022.

Guth, S., Sapsis, T.P., 2022. 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 Eng. (ISSN: 0029-8018) 266, 112633.

Haver, S., Winterstein, S.R., 2008. Environmental contour lines: A method for estimating long term extremes by a short term analysis. In: SNAME Maritime Convention. OnePetro.

International Renewable Energy Agency (IRENA), 2020. Abu dhabi, innovation outlook: Ocean energy technologies.

Katsidoniatski, E., Yu, Y.-H., Goteman, M., 2021. 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).

Katsidoniotaki, M., 2022. Uncertainty quantification techniques with diverse applications to stochastic dynamics of structural and nanomechanical systems and to modeling of cerebral autoregulation (Ph.D. thesis). Columbia University.

Katsidoniotaki, E., Göteman, M., 2022. Numerical modeling of extreme wave interaction with point-absorber using openfoam. Ocean Eng. 245, 110268.

Katsidoniotaki, E., Nilsson, E., Rutgersson, A., Engström, J., Göteman, M., 2021. Response of point-absorbing wave energy conversion system in 50-years return period extreme focused waves. J. Mar. Sci. Eng. 9 (3), 345.

Katsidoniotaki, E., Psarommatis, F., Göteman, M., 2022a. Digital twin for the prediction of extreme loads on a wave energy conversion system. Energies 15 (15), 5464.

Katsidoniotaki, M.I., Psaros, A.F., Kougioumtzoglou, I.A., 2022b. Uncertainty quantification of nonlinear system stochastic response estimates based on the Wiener path integral technique: A Bayesian compressive sampling treatment. Probab. Eng. Mech. 67, 103193.

Katsidoniotaki, E., Ransley, E., Brown, S., Palm, J., Engström, J., Göteman, M., 2020. Loads on a point-absorber wave energy converter in regular and focused extreme wave events. In: International Conference on Offshore Mechanics and Arctic Engineering. 84416, American Society of Mechanical Engineers, V009T09A022.

Katsidoniotaki, S.G.E., Sapsis, T.P., 2023. Statistical modeling of fully nonlinear hydrodynamic loads on offshore wind turbine foundations using wave episodes and targeted CFD simulations through active sampling. Submitted.

Katsidoniotaki, E., Shahroozi, Z., Eskilsson, C., Palm, J., Engström, J., Göteman, M., 2022c. Validation of a CFD model for wave energy system dynamics in extreme waves. Under Revis. Ocean. Eng..

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Leijon, M., Boström, C., Danielsson, O., Gustafsson, S., Haikonen, K., Langhamer, O., Strömstedt, E., Stålberg, M., Sundberg, J., Svensson, O., et al., 2008. Wave energy from the north sea: Experiences from the lysekil research site. Surv. Geophys. 29 (3), 221–240.

Li, L., Yuan, Z., Gao, Y., 2018. Maximization of energy absorption for a wave energy converter using the deep machine learning. Energy 165, 340–349.

Li, X., Zhang, W., 2020. Long-term fatigue damage assessment for a floating offshore wind turbine under realistic environmental conditions. Renew. Energy 159, 570–584.

Liu, Y., Liu, X., Guo, J., Lou, R., Lv, Z., 2022. 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, IEEE, pp. 718–719.

Mohamad, M.A., Sapsis, T.P., 2018. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. Proc. Natl. Acad. Sci. 115 (44), 11138–11143.

Moura Paredes, G., Eskilsson, C., P. Engsig-Karup, A., 2020. Uncertainty quantification in mooring cable dynamics using polynomial chaos expansions. J. Mar. Sci. Eng. 8 (3), 162.

Mousavi, S.M., Ghasemi, M., Dehghan Manshadi, M., Mosavi, A., 2021. Deep learning for wave energy converter modeling using long short-term memory. Mathematics 9 (8), 871.

Ning, D., Ding, B., 2022. Modelling and Optimization of Wave Energy Converters. CRC Press.

Pickering, E., Guth, S., Karniadakis, G.E., Sapsis, T.P., 2022. Discovering and forecasting extreme events via active learning in neural operators. Nat. Comput. Sci. 2 (12), 823–833.

Quevedo-Reina, R., Álamo, G.M., Padrón, L.A., Aznárez, J.J., 2023. Surrogate model based on ANN for the evaluation of the fundamental frequency of offshore wind turbines supported on jackets. Comput. Struct. 274, 106917.

Ransley, E.J., 2015. Survivability of wave energy converter and mooring coupled system using CFD (Ph.D. thesis). Plymouth University.

van Rij, J., Yu, Y.-H., Guo, Y., Coe, R.G., 2019. A wave energy converter design load case study. J. Mar. Sci. Eng. 7 (8), 250.

Rodríguez, C.A., Rosa-Santos, P., Taveira-Pinto, F., 2019. Assessment of damping coefficients of power take-off systems of wave energy converters: A hybrid approach. Energy 169, 1022–1038.

Sapsis, T.P., 2020. Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples. Proc. R. Soc. A: Math. Phys. Eng. Sci. 476 (2234), 20190834.

Sapsis, T.P., 2021. Statistics of extreme events in fluid flows and waves.

Shen, Q., Vahdatikhaki, F., Voordijk, H., van der Gucht, J., van der Meer, L., 2022. Metamodel-based generative design of wind turbine foundations. Autom. Constr. 138, 104233.

Shinozuka, M., 1983. Basic analysis of structural safety. J. Struct. Eng. 109 (3), 721–740.

Sobester, A., Forrester, A., Keane, A., 2008. Engineering design via surrogate modelling: a practical guide. John Wiley & Sons.

Tromans, P.S., Anaturk, A.R., Hagemeijer, P., 1991. 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.

Wang, Z., Qiao, D., Yan, J., Tang, G., Li, B., Ning, D., 2022. A new approach to predict dynamic mooring tension using LSTM neural network based on responses of floating structure. Ocean Eng. 249, 110905.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. vol. 2, (3), MIT press Cambridge, MA.

Wrang, L., Katsidoniotaki, E., Nilsson, E., Rutgersson, A., Rydén, J., Göteman, M., 2021. Comparative analysis of environmental contour approaches to estimating extreme waves for offshore installations for the baltic sea and the north sea. J. Mar. Sci. Eng. 9 (1), 96.

Yu, Y.-H., Li, Y., 2013. Reynolds-averaged Navier–Stokes simulation of the heave performance of a two-body floating-point absorber wave energy system. Comput. & Fluids 73, 104–114.

Yu, Y.-H., Li, Y., Hallett, K., Hotimsky, C., 2014. Design and analysis for a floating oscillating surge wave energy converter. In: International Conference on Offshore Mechanics and Arctic Engineering. 45547, American Society of Mechanical Engineers, V09BT09A048.