Ocean Engineering 266 (2022) 112633

Contents lists available at ScienceDirect

Ocean Engineering

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

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

Stephen Guth, Themistoklis P. Sapsis *

Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, United States of America

ARTICLE INFO

Keywords:
Nonlinear ship dynamics and loads
Heavy tails and extreme events
Gaussian Process Regression
Wave episodes
Reduced-order modeling
Karhunen–Loève expansion

ABSTRACT

For a number of structural and fatigue problems in marine science, engineers need accurate statistics for kinematic and dynamical quantities, such as displacements and bending moments. For many such problems the response statistics depend nonlinearly on input stochastic processes, such as sea surface elevation. Nonlinearity prevents the use of standard linear methods, leading to the adoption of expensive experiments and time-consuming numerical simulations. In order to avoid this cost we present a machine learning framework to minimize the training set requirements. This framework consists of two parts. First, we use the Karhunen–Loève theorem to represent stochastic sea states with finite-time wave episodes, which have low dimensionality that nonetheless captures the features important to hydrodynamics and structural mechanics. However, the choice of the wave episode duration is 'caught' between the Scylla of slow Karhunen–Loève series convergence for long time durations and the Charybdis of missing transient behaviors when the interval is short. To combat this dilemma, we propose a division into a region, designed for parametric interpolation and machine learning, and a stochastic prelude region, designed to probabilistically model transients. The second part of the framework consists of a Gaussian Process Regression (GPR) model designed to learn the mapping from wave episodes to structural outputs. GPR is able to take advantage of the low dimensional parametric representation of the sea state in order to converge with reasonably-sized training sets (on the order $n \approx 100$). At the same time, we use a low-dimensional representation in order to represent the stochastic response time series. The principal advantages of the Gaussian process surrogate are the blazing speed of evaluation – ten thousand times faster than the direct method – and the built in uncertainty quantification. Taken together, we can reconstruct the statistics of the responses by sampling sea states via the Karhunen–Loève construction, estimating the corresponding model outputs using the trained GPR, and estimating statistics of interest through Monte-Carlo calculation on the surrogate model. Finally the developed framework allows for trivial computation of the statistics for different sea spectra, i.e. without the need to retrain the Gaussian Process Regression scheme.

1. Introduction

The design of ocean vessels relies extensively on simulated hydrodynamics and structural responses in order to estimate both performance characteristics (Naess and Moan, 2013) and safety characteristics such as fatigue lifetime (Serebrinsky and Ortiz, 2005; Khan and Ahmad, 2007; Chasparis et al., 2009) and capsize risk (Belenky et al., 2016, 2018). The vertical bending moment in particular is an important target for modeling structural properties (Sapsis et al., 2020, 2021; Belenky et al., 2021). The challenge however is that ship structural responses can have strongly non-Gaussian statistics, especially when these are associated with large magnitude responses due to large waves. While there is a rich history of calculating second order statistics of ocean waves (Longuet-Higgens, 1957; Longuet-Higgins, 1975; Sharma and Dean, 1979; Tayfun, 1980; Forristall, 2000; Fedele and Tayfun, 2009), accurately calculating the ship's non-Gaussian statistical responses, and especially the extreme event properties remains a challenging task.

While finite element codes coupled with explicit time domain techniques (Shin et al., 2003) are able to accurately simulate the structural response to specific instances of sea conditions, the sea surface elevation is in general a stochastic process. Any simulation program with the goal of accurately estimating the statistical vessel response in the steady state will require a strategy for constructing representative sea surfaces. The literature has proposed a number of decomposition approaches to this problem (Belenky et al., 2012; Kyul Joo et al., 2018), and many

* Corresponding author.
E-mail address: sapsis@mit.edu (T.P. Sapsis).

https://doi.org/10.1016/j.oceaneng.2022.112633
Received 15 April 2022; Received in revised form 19 August 2022; Accepted 16 September 2022

previous investigators have described particular representations of sea surface waves, examples of which include the random phase model, stochastic wavegroups (Boccotti, 1989; Phillips et al., 1992), critical wavegroups (Sclavounos, 2012; Anastopoulos et al., 2016; Anastopoulos and Spyrou, 2016, 2019), and reduced order wavegroups (Mohamad and Sapsis, 2018; Cousins and Sapsis, 2016; Farazmand and Sapsis, 2017). The use of wavegroup based methods foregrounds the impact of irregular waves on the dynamical and structural response of the vessel, particularly the memory effects.

Evading the iron cost of Monte-Carlo sampling, which is the scope of this paper, requires both reconstructing the true statistical quantities of interest from a limited data set, as well as carefully choosing experimental designs to produce that data set. The reconstruction step draws from the theory of data-driven surrogate modeling techniques (Forrester et al., 2008; Vu et al., 2017; Rudy et al., 2021), and in particular theory of Gaussian Process Regression (GPR) (Rasmussen and Williams, 2006; Stevens, 2018). Like related techniques, such as Envelope Peaks Over Threshold (Campbell et al., 2016; Weems et al., 2019; Belenky et al., 2019), surrogate modeling must be carefully designed to accurately capture the tails of the distribution. The design step draws from literature on optimal experimental design both in the traditional Bayesian formulation (Chaloner and Verdinelli, 1995; Huan and Marzouk, 2013) and in forms that are either output aware or specifically tailored for recovering extreme events (Sapsis, 2020; Blanchard and Sapsis, 2021b; Sapsis, 2021). In particular, sequential design and active search take this to its natural conclusion, where early simulation results are used to fine tune subsequent experimental designs to maximize information (Mohamad and Sapsis, 2018; Blanchard and Sapsis, 2021a).

Our goal here is to design a machine learning approach to compute extreme event statistics with very few simulations. The challenge however is that statistics are defined by the nonlinear interaction of the ship with the stochastic waves. Neither of these two ingredients can be simplified, especially when the goal is to accurately recover the tails of the distribution. The first component of the presented framework consists of a finite-dimensional probabilistic representation of the stochastic sea state that is capable to capture the transient features of the nonlinear dynamics. This representation has the form of finite-time and finite-dimensional random wave episodes 'equipped' with a stochastic prelude whose role is to bring the system into the statistical steady state. Using this random wave episode representation and a high-fidelity numerical solver we obtain the corresponding response, which is also represented in a finite-dimensional form. Using wave episode-response pairs in finite-dimensional form we then machine learn a probabilistic map in the form of a GPR. The map is trained with a relatively small set of wave episode-response pairs and it allows to obtain a highly accurate description of the statistics with minimal computational cost.

The structure of the paper is as follows. In Section 2, we describe our approach to modeling sea state as both a stochastic process with known statistics, and also a parametrically controlled finite-time wave episode with a stochastic prelude. We leverage this approach to construct a set of experimental designs which (a) have known statistical relations to the steady state and (b) can be meaningfully interpolated across. In Section 3, we discuss the numerical code we use to obtain ship motions and loads under prescribed wave time-series, as well as some challenges we encounter when trying to develop a simple correspondence between the input time series and the output time series. In Section 4, we discuss GPR, a machine learning tool commonly used in design problems. In particular, we explain how we use GPR to reconstruct full output time series for the considered problem. Next, in Section 5, we present validations of the wave episode construction approach. In particular, we compare simulation results from the random phase model to the random wave episode representation, and we compare resampled statistics from the Gaussian process surrogates to directly sampled Monte-Carlo statistics. Finally, in Section 6 we investigate convergence when varying various algorithm parameters, among them the size of the training set, the number of retained output time series modes, and two important parameters used to model the wave episodes that represent the stochastic sea state: their duration and their dimensionality.

2. Stochastic-preluded wave episodes for representing irregular seas for nonlinear ship dynamics

Our goal is to develop a finite-dimensional and finite-time stochastic representation of the irregular sea state (i.e. excitation) that would be appropriate to employ in order to capture the response statistics of chaotic dynamical systems, e.g. ship motions and dynamics in irregular seas. A marine vessel in irregular seas can be seen as a nonlinear dissipative system that is subjected to multi-frequency excitation, which is a typical situation for chaotic behavior (e.g. a duffing oscillator subjected to even mono chromatic excitation is a chaotic system Wiggins, 2003)

Why do we need to obtain a representation of the irregular sea state in terms of random wave episodes, i.e. finite-time waves having random parameters? This is motivated by the need to optimally design wave tank experiments where the duration of each experimental run will be of finite duration, as short as possible. Also, it is obvious that this representation should involve a finite number of random parameters, preferably a small number, so it can be practical for realistic settings.

Representing a stationary and ergodic stochastic process, such as the irregular sea state, in terms of finite-time wave episodes is a straightforward task using, for example, the Karhunen–Loève (KL) expansion. However, there are many challenges when this finite-time representation of the irregular sea state has to be employed to quantify the statistical steady-state statistics of a chaotic system, such as the nonlinear ship response. This is because a random wave episode representation should not only have consistent statistics with the irregular sea state (given spectrum), but it has to take into account that the chaotic system has finite memory and dynamics characterized by transient features. For example, when any given wave episode excites the dynamic system, the latter may be in any point of the chaotic attractor. This random initial state, however, cannot be practically implemented in an experimental setting, i.e in a typical wave tank the model is initially at rest. For this reason, the random wave episode needs to be 'equipped' with a stochastic prelude that precedes the random wave episode and brings the system from rest to a random state, consistent with the chaotic attractor.

This stochastic prelude is designed so that (i) it has consistent statistics with the excitation, i.e. the irregular sea state, (ii) it is smoothly connected with each wave episode, and (iii) it excites the system for enough time so that it reaches the statistical steady state before it encounters the actual wave episode. To achieve these properties we assume that the stochastic prelude has a Gaussian prior with the spectrum of the irregular sea state. Furthermore, we condition this prior on each wave episode, similarly with what is done in a standard GPR scheme. To present the notion of the spectrum-consistent stochastic-preluded wave episodes we begin with a brief review of important notions, such as typical sea state spectra and representation of a sea state in terms of a KL expansion into finite-time wave episodes.

2.1. Sea state spectrum

Following general practice, we represent particular sea states by a time series of the sea surface elevation at a particular point ($\xi = 0$). We further represent this sea surface elevation as a zero-mean, statistically stationary, Gaussian process. Finally, we characterize this stochastic process by its power spectral density, which is defined as the Fourier transform of the autocorrelation function, given by

$$R_x(\tau) = \mathbb{E}[x(t_1)x(t_2)] = \mathbb{E}[x(t)x(t+\tau)]. \tag{1}$$

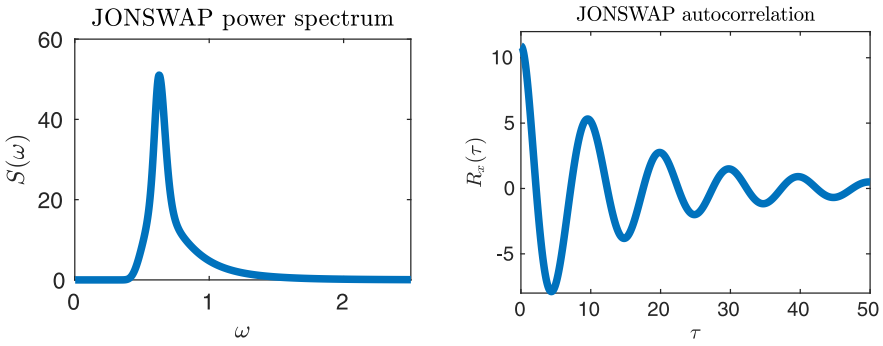

Fig. 1. (a) One-sided power spectral density for the JONSWAP spectrum. (b) JONSWAP time autocorrelation function.

Here we consider the JONSWAP spectrum (Fig. 1) (Hasselmann et al., 1973) with modal frequency ω_m:

$$S_J(\omega) = \frac{\alpha g^2}{\omega^5} \exp\left[-\frac{5}{4}\left(\frac{\omega_m}{\omega}\right)^4\right] \gamma^r \tag{2}$$

$$r = \exp\left[-\frac{(\omega-\omega_m)^2}{2\sigma^2\omega_m^2}\right]$$

where constants are given by

$$\alpha = 0.076\left(\frac{U_{10}^2}{Fg}\right)^{0.22}, \qquad \omega_m = 22\left(\frac{g^2}{U_{10}F}\right)^{\frac{1}{3}}, \quad \gamma = 3.0,$$

$$\sigma = \begin{cases} 0.07 & \omega \le \omega_m \\ 0.09 & \omega > \omega_m, \end{cases}$$

F is the wind fetch, and U_{10} is the wind speed measured at 10 m above the sea surface.

For the remainder of this work, we present results from a JONSWAP spectrum with $\alpha = 0.06$ and modal period $T_m = \frac{2\pi}{\omega_m} = 10$ s (together implying significant wave height $H_s = 13.2$ m and modal frequency $w_m = 2\pi/T_m = 0.63$ rad $\times$ s^{-1}). This very extreme sea state was chosen to emphasize the nonlinear interactions associated with very steep waves. The procedure described may be generalized to other wave spectra without change.

2.2. Random phase model

For stationary Gaussian random processes, we can create samples from the power spectrum using the random phase model:

$$x(t) = \sum_{i=1} \alpha_i \cos(\omega_i t + \phi_i), \tag{3}$$

where α_i, ω_i are deterministic amplitudes and frequencies, while ϕ_i are random phases that are uniformly distributed in $[0, 2\pi]$. We can derive the corresponding power spectral density for this model as

$$S_{xx}(\omega) = \sum_{i=1} \frac{\alpha_i^2}{2}\delta(\omega - \omega_i) \tag{4}$$

where $\delta(x)$ is the Dirac delta function.

We can use the last expression to select the amplitudes α_i so that the model approximates any given spectral density, $S(\omega)$. This will result the following representation:

$$x(t) = \sum_{i=1} \sqrt{2S(\omega_i)\delta_{\omega_i}} \cos(\omega_i t + \phi_i), \tag{5}$$

where $\delta_{\omega_i} = \omega_{i+1} - \omega_i$ is the bin width associated with the frequency discretization.

A direct approach for the quantification of statistics for ship responses and structural loads involves the simulation of the system using a large number of sea state realizations based on the random phase model. However, the random phase model for sampling from random sea states has two drawbacks. First, it requires a large number of modes in order that the discrete samples adequately represent the Gaussian random process. This means that random phase model is equivalent to a very high dimensional parametrization of the space of random processes (where the parameters are the random phases). Second, the parametrization is opaque: each ϕ_i controls the phase of one mode in the sum, but it is impossible to interpret how the phases impact the shape of the random process realization. To this end, we turn our attention to an expansion over finite-time wave episodes.

2.3. Karhunen–Loève (KL) expansion of irregular waves in finite-time wave episodes

An alternative approach for representing random waves is to sample random processes over a finite interval (Sclavounos, 2012; Anastopoulos et al., 2016; Anastopoulos and Spyrou, 2016). We recall the Karhunen–Loève (KL) theorem (Karhunen, 1947; Loève, 1948) that is the basic building block for such construction.

Theorem 2.1 (*Karhunen Loève*). *Consider the stationary stochastic process $x(t)$ which is zero mean and square integrable. Define the covariance function $R_x(\tau)$ with the corresponding integral operator over the interval $[0, T]$,*

$$T_{R_x}\phi(t) = \int_0^T R_x(t-s)\phi(t)ds, \quad t \in [0,T]. \tag{6}$$

Then by Mercer's Theorem for every interval $[0,T]$ the operator T_{R_x} has an orthonormal basis of eigenvectors $\{\hat{e}_{i,T}(t)\}, i = 1, \ldots$ and corresponding eigenvalues $\{\lambda_i\}$. Moreover, the coefficients

$$\alpha_i = \int_0^T x(t)\hat{e}_{i,T}(t)dt, \quad i = 1, \ldots \tag{7}$$

are centered orthogonal random variables:

$$\mathbb{E}[\alpha_i\alpha_j] = 0 \ \text{ for } \ i \neq j \ \text{ and } \ Var(\alpha_i) = \mathbb{E}[\alpha_i^2] = \lambda_i. \tag{8}$$

Furthermore, we can expand the random process $x(t)$ as

$$x(t) = \sum_{i=1}^{\infty} \alpha_i \hat{e}_{i,T}(t), \quad t \in [0,T]. \tag{9}$$

In summary, the eigenvectors of the covariance matrix of the sea surface form an orthonormal basis. The decomposition of $x(t)$ onto this basis produces a set of centered, orthogonal (in the random sense) coefficients with variance λ_i. We note that, for Gaussian random variables, orthogonality is equivalent to independence.

By retaining in the truncation a sufficient number of coefficients we expect to model the largest part of the variance in the reconstructed random process. This can be measured by the sum of the retained eigenvalues, λ_i, relative to the overall variance of the process (integral of the spectral density). In this way, we obtain a finite-dimensional

approximation of the original stochastic sea state in terms of stochastic finite-time wave episodes having the form:

$$x(t) \simeq \sum_{i=1}^{n} \alpha_i \hat{e}_{i,T}(t), \quad t \in [0,T]. \tag{10}$$

In particular, given a known set of basis vectors $\hat{e}_{i,T}(t)$, we can change back and forth between the time series representation $x(t), t \in [0,T]$ and the coefficients representation $\alpha_i, i = 1,2,\ldots$. The forward step (coefficient representation to time series) can be performed easily with Eq. (10), while the backward step is the discrete analogue of the projection given in Eq. (7).

We considered two methods for determining the basis vectors $\hat{e}_{i,T}(t)$ (and associated eigenvalues λ_i). First, we performed singular value decomposition (SVD) of an approximate discrete correlation matrix

$$\hat{R}_x(t_i,t_j) = \frac{1}{n}\sum_{k=1}^{n} x_k(t_i)x_k(t_j), \tag{11}$$

where $\{x_k(t) : k \in [1,\ldots,n]\}$ are independent samples from the random process on the interval $[0,T]$, computed with the random phase mode. The eigenvectors and eigenvalues of the matrix $\hat{R}_x(t_i,t_j)$ correspond to $\hat{e}_{i,T}(t)$ and λ_i respectively. Second, we implemented the Sclavounos procedure (Sclavounos, 2012) for directly computing the basis vectors and eigenvalues. We found the primary differences between the two techniques to be parity; the Sclavounos procedure conserved even and odd basis vectors, while the SVD procedure allowed the two to mix. All results presented here make use of the SVD procedure for calculating the KL basis, but we expect identical results from the Sclavounos basis.

Two important questions here are first, how we should choose the length of the wave episode interval, T, and second, how we should choose the truncation order n.

2.3.1. Selection of the wave episodes duration, T

First we note that the obtained representation (10) should be capable of capturing the correlation structure of the stochastic sea state. This places a lower bound for T: it should be longer than the decorrelation time of the stochastic sea state. However, another factor that enforces a lower bound on T is the memory of the chaotic system. This should also be adequately modeled and this can only happen if the excitation wave episodes are sufficiently long.

For our purposes, we define the memory length of the chaotic dynamics to be the decay time of a fictitious impulse response function of the system (in this cases the wave excited ship). That is, if a large wave impacts the ship, we should look for how long it will take for the vessel dynamics to have 'mostly forgotten' the impact. The vessel dynamics are, of course, nonlinear, and for this reason we cannot define an impulse response function. Nonetheless, this provides a useful heuristic for thinking about how short we can make T. Note that if T is shorter than the required memory then when we measure the ship dynamics in the wave episode interval $[0,T]$, our measurements will be primarily influenced by the effects of the initial conditions (or the waves before the interval $[0,T]$).

2.3.2. Selection of the truncation order, n

The other parameter that needs to be selected is the truncation order, n. The truncation orders controls the fraction of the energy of the sea state our projection retains, or equivalently, the 'shape complexity' of the wave episodes inside of the projected space. At the same time, n also represents the dimensionality of the parametrized space of possible wave episodes. Therefore, our choice of n should balance between curse of high dimensionality on the one hand, and accurately representing a wide enough variety of wave episodes on the other.

An additional wrinkle is the relationship between T and n. Roughly speaking, as T increases, a larger n is necessary to maintain the same amount of energy (Fig. 2). In terms of retained energy content, the necessary n increases approximately linearly with T. In terms of shape complexity however, the minimum required n grows more slowly. This

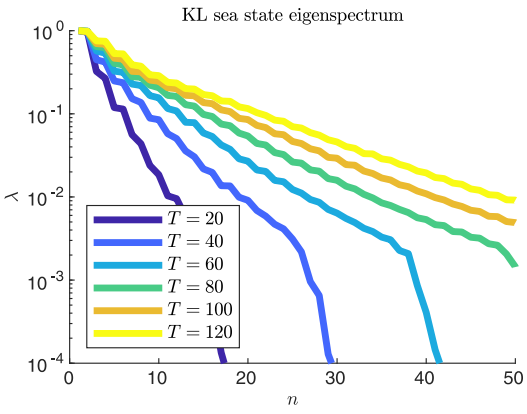

Fig. 2. Eigenvalues of the KL expansion in wave episodes of length T for representing irregular seas described by the JONSWAP spectrum.

interplay between n and T suggests that we need to be cautious in selecting T, since it should not be very small (as this will cause problems with modeling the memory of the input stochastic process and/or the chaotic system) nor very large (as this will lead to a very high dimensional representation). In Section 6.1, we analyze quantitatively the effect of T and n in connection to the performance of the stochastic regression model employed to model the dynamical response of the chaotic system.

2.4. Stochastic-prelude for finite-time wave episodes

Beyond the issues related to the adequate representation of the memory effects of the system and the correlation structure of the sea state, it is essential to model the effect of initial conditions of the system, when the finite-time wave episode begins to excite it. A simple modeling approach would be to include an additional number of random parameters that would represent the initial state of the model. However, due to the statistical relationship between the sea surface across time, the latter has complicated a priori statistics.

By 'complicated a priori statistics,' we mean that while the sea surface elevation (or even the ship motions) at an arbitrary time can be described by simple Gaussian statistics, more complicated is the question of the sea surface elevation before $t = 0$, given that we know that the sea surface elevation at time $t = 0$ is given by $x(t = 0)$. Moreover, even in the case where one could approximate the statistics, enforcing initial conditions is not practical for ship dynamics when the experiments take place in a wave tank.

To overcome these limitations we 'equip' the generated finite-time wave episodes with a stochastic prelude: a smooth backwards-in-time extension of the finite-time wave episode which has consistent statistics with the stochastic process, x. Specifically, let the probability measure of the stochastic process, $x(t), t \in \mathbb{R}$ be given by

$$\mathbb{P}[x(t), t \in \mathbb{R}]. \tag{12}$$

We can think of the above probability measure as the joint pdf for the sea surface elevation at every combination of time instances, i.e. for any M and $(t_1,\ldots,t_M)$, $\mathbb{P}[x(t), t \in \mathbb{R}]$ contains information for the joint pdf $p_{\vec{x}}(x(t_1),\ldots,x(t_M))$. For what follows, we assume that the stochastic process is Gaussian, zero-mean and statistically stationary with prescribed spectrum, typical for ocean waves.

We define the conditional stochastic process, $x(t|\boldsymbol{\alpha}), t \in \mathbb{R}, \boldsymbol{\alpha} \in \mathbb{R}^n$, on a given (finite) set of KL coefficients that prescribe a wave episode in the interval $[0,T]$, as well as the associated *conditional probability measure*:

$$\mathbb{P}[x(t), t \in \mathbb{R}|\alpha_1,\ldots,\alpha_n] \triangleq \mathbb{P}\left[x(t), t \in \mathbb{R} \,\middle|\, \int_0^T x(s)\hat{e}_{i,T}(s)ds = \left\{ \begin{array}{l} \alpha_i,\ i = 1,\ldots,n \\ 0, i = n+1,\ldots \end{array} \right\}\right]. \tag{13}$$

For the time interval $[0, T]$ we have zero (conditional) uncertainty, since the function is fully defined through the KL coefficients. On the other hand, as we depart from this interval, i.e. for $t < 0$ or $t > T$, the uncertainty increases gradually. The rate of this increase depends on the decorrelation time of the stochastic process. For sufficiently large distance from the interval $[0, T]$ the statistics of the conditioned process are identical with the unconditional probability measure, (12). We call the segment of the conditional stochastic process that corresponds to $t < 0$ the *stochastic prelude* (SP) of the wave episode. We note that our primary interest is in the prelude region (that is, $t < 0$), but that the same analysis holds for the $t > T$ region as well.

We emphasize that when we employ the developed wave episodes to characterize the response of the dynamical system these have to be accompanied with their stochastic prelude. The stochastic prelude will bring the dynamical system into a random state, so that when it encounters each prescribed wave episode, the effect of the stochastic initial state is taken into account. How far back should the SP extend to? That depends on the memory of the dynamical system. Assuming a dissipative chaotic model, as is the case for a typical ship dynamics model, this memory will be finite. In the next section, where we consider a specific application, we describe in more detail this parameter.

2.4.1. Construction of the stochastic prelude using Gaussian process conditioning

The next step of our analysis involves the numerical construction of the stochastic preludes. We begin with the wave episode in $[0, T]$, which is fully specified by the KL coefficients $\alpha_1, \dots, \alpha_n$. We generate random realizations of the stochastic prelude ($t < 0$), by sampling from a Gaussian Process posterior distribution (Rasmussen and Williams, 2006) in an iterative extrapolation mode. In brief, will we use the known values of $x(t)$ inside the wave episode, along with any previously samples from the stochastic prelude, in order to iteratively sample the next point from the stochastic prelude in an extrapolation mode.

Let the length of the stochastic prelude be given by T_{SP} (i.e., the stochastic prelude is the interval $[-T_{SP}, 0]$). We begin by choosing a discretization of both the wave episode and stochastic prelude into n_W and n_{SP} points respectively, with spacing Δt. Next, we choose a memory length, $n_m \leq n_{SP}$. The memory length depends on the decay rate of the correlation function for the stochastic process, $x(t)$. For numerical reasons, we also require that $n_m \leq n_W$. Additionally, we require that $n_m \Delta t$ be comparable with the typical decorrelation length of the stochastic process. Finally, a very small Δt leads to poorly conditioned matrix inversion step.

Then, we compute a Töplitz covariance matrix

$$\Sigma = \begin{bmatrix} R(0) & R(\Delta t) & R(2\Delta t) & \cdots & R((n_m-1)\Delta t) \\ R(\Delta t) & R(0) & R(\Delta t) & \cdots & R((n_m-2)\Delta t) \\ R(2\Delta t) & R(\Delta t) & R(0) & \cdots & R((n_m-3)\Delta t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R((n_m-1)\Delta t) & R((n_m-2)\Delta t) & R((n_m-3)\Delta t) & \cdots & R(0) \end{bmatrix}, \tag{14}$$

where $R(\tau)$ is the autocorrelation (equal to the inverse Fourier transform the power spectrum) of the stochastic process. We also compute the marginal row

$$\Sigma^* = \begin{bmatrix} R(n_m \Delta t) & R((n_m-1)\Delta t) & R((n_m-2)\Delta t), & \dots & R(\Delta t) \end{bmatrix}. \tag{15}$$

Note that while Σ and Σ^* depend on the memory length n_m, they are independent of the wave episode in $[0, T]$, i.e. the KL coefficients $\alpha_1, \dots, \alpha_n$.

We extrapolate backwards from the interval $[0, T]$ in an iterative manner:

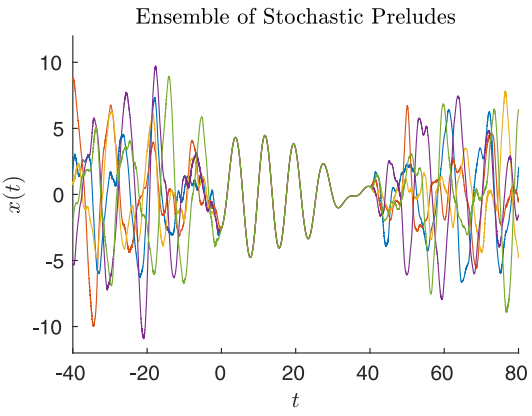

Fig. 3. Five distinct stochastic-prelude (SP) realizations ($t < 0, t > 40$) corresponding to a single wave episode, $t \in [0, 40]$.

1. Using the n_m points (denoted with the vector, $\hat{X} = (x(t_1), \dots, x(t_{n_m}))$), we calculate the conditional mean and variance of the next point in the sequence. These are given by

$$\mu_{t_0|t_1,\dots,t_{n_m}} = \Sigma^* \Sigma^{-1} \hat{X} \tag{16}$$

$$\sigma^2_{t_0|t_1,\dots,t_{n_m}} = R(0) - \Sigma^* \Sigma^{-1} \Sigma^{*\mathsf{T}}. \tag{17}$$

2. We sample the next point, in the sequence, $x(t_0)$, from its conditional distribution:

$$\mathcal{N}(\mu_{t_0|t_1,\dots,t_{n_m}}, \sigma^2_{t_0|t_1,\dots,t_{n_m}}).$$

3. We shift the 'reading frame' by Δt and repeat, until we cover the whole duration of the stochastic prelude.

Note that, like Σ and Σ^*, we may precompute $\sigma^2_{t_0|t_1,\dots,t_{n_m}}$, since the posterior sampling variance is independent of the sample. Similarly, we need only perform the matrix inversion for the vector expression $\Sigma^* \Sigma^{-1}$ once.

Let us briefly summarize application of the preluding technique described in this section. We begin with a time series defined on the interval $[0, T]$. Using this wave episode, we iteratively extend this time series backward in time (into the prelude region $t \in [-T_{sp}, 0]$) by sampling the sea state at each time from a conditional distribution that is consistent with both the power spectrum of the sea state, as well as the already fixed portion of the wave episode.

2.4.2. Statistical structure of the stochastic prelude

In Fig. 3 we show an example of a set of distinct overlaid stochastic preludes, corresponding to a single wave episode, i.e. single set of KL coefficients. For any given set of KL coefficients, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$, we define the $\boldsymbol{\alpha}-$conditional mean of each wave episode:

$$\bar{x}(t|\boldsymbol{\alpha}) = \mathbb{E}[x(t)|\boldsymbol{\alpha}], \quad t \in \mathbb{R}. \tag{18}$$

and its $\boldsymbol{\alpha}-$conditional variance or stochastic prelude variance:

$$\sigma^2_{SP}(t|\boldsymbol{\alpha}) = \mathbb{E}[(x(t|\boldsymbol{\alpha}) - \mathbb{E}[x(t)|\boldsymbol{\alpha}])^2], \quad t \in \mathbb{R}. \tag{19}$$

Clearly, within the interval $[0, T]$ the $\boldsymbol{\alpha}-$conditional wave episode is deterministic, i.e. its value is given by the KL expansion, (10), and the $\boldsymbol{\alpha}-$conditional variance, $\sigma^2_{SP}(t|\boldsymbol{\alpha})$, is always zero. Outside this interval, where the stochastic prelude is defined, the $\boldsymbol{\alpha}-$conditional statistics are given by the Gaussian Process extrapolation we constructed in the previous subsection. Specifically, the mean of the stochastic prelude is a smooth extrapolation of the wave episode, and gradually converges to zero, since away from $[0, T]$ there is no influence from the KL

coefficients and the conditional process has identical statistics with $x(t)$, i.e.

$$\bar{x}(t|\boldsymbol{\alpha}) = 0 \quad \text{for} \quad t \ll 0 \quad \text{or} \quad t \gg T. \tag{20}$$

The stochastic prelude variance, $\sigma^2_{SP}(t|\boldsymbol{\alpha})$, is 0 in the interval $[0, T]$ and gradually increases to the variance of the stochastic process, σ^2_x; see Fig. 4, blue curve. We also note that as we concluded in Eq. (17) the variance of the stochastic prelude is independent of the wave episode, i.e. the KL coefficients. Therefore, $\sigma^2_{SP}(t|\boldsymbol{\alpha}) = \sigma^2_{SP}(t)$.

Equally as interesting the conditional statistics, we have the energy of the generated wave episodes and their stochastic preludes, i.e. their variance over the random KL coefficients, $\boldsymbol{\alpha}$. We recall that the KL coefficients follow a Gaussian distribution, $p_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$.

We have for the mean of the wave episodes:

$$\mathbb{E}^{\boldsymbol{\alpha}}[\bar{x}(t|\boldsymbol{\alpha})] = \int_{\mathbb{R}^n} \bar{x}(t|\boldsymbol{\alpha}) p_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) d\boldsymbol{\alpha} = 0, \quad t \in \mathbb{R}. \tag{21}$$

The above can be easily seen in $[0, T]$ using the KL expansion and the fact that the KL coefficients are zero mean. Outside $[0, T]$ we can show the same result using Eq. (16).

On the one hand, for the variance of the wave episodes (over $\boldsymbol{\alpha}$) we have:

$$\sigma^2_W(t) \triangleq \mathbb{E}^{\boldsymbol{\alpha}}[(\bar{x}(t|\boldsymbol{\alpha}))^2] = \int_{\mathbb{R}^n} (\bar{x}(t|\boldsymbol{\alpha}))^2 \, p_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) d\boldsymbol{\alpha}, \quad t \in \mathbb{R}. \tag{22}$$

For $t \in [0, T]$ we have a tight bound, $\sigma^2_W(t) < \sigma^2_x$; we do not have equality since we are setting all the higher order KL coefficients (greater than n) to 0. Therefore, some variance will always be lost due to the finite truncation of the KL expansion. However, as n grows we have $\lim_{n\to\infty} \sigma^2_W(t) = \sigma^2_x, t \in [0, T]$.

On the other hand, outside the wave episode interval $[0, T]$ the variance due to the randomness of the KL coefficients will decay and eventually converge to zero; see Fig. 4, red curve. This is because, for a stochastic process $x(t)$ with finite decorrelation time, the influence of the KL coefficients on the stochastic preludes decays away from $[0, T]$. Specifically,

$$\bar{x}(t|\boldsymbol{\alpha}) = 0, \quad \text{for } t \ll 0 \quad \text{or} \quad t \gg T. \tag{23}$$

Thus, $\sigma^2_W(t) = 0$ for $t \ll 0$ or $t \gg T$.

Finally, using the law of total expectation we have

$$\sigma^2_{SP}(t) + \sigma^2_W(t) \leqslant \sigma^2_x, \tag{24}$$

where the equality holds for $n \to \infty$ or $t \to \pm\infty$. This property is also demonstrated in Fig. 4.

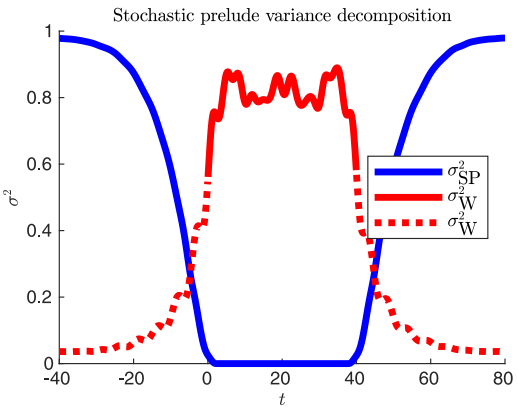

Fig. 4. Variance of the stochastic prelude, $\sigma^2_{SP}(t)$ (blue curve) and energy, $\sigma^2_W(t)$ of the generated wave episodes. Dashed red line corresponds to the part of the energy in the stochastic prelude region that is associated with the mean effect of the wave episode. Here we have $T = 40$, while we consider $n = 6$ (nonzero) KL modes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Hydrostatic details of the ONR Topsides flared variant hull.

Displacement	8520 kg
Maximum draft	7.61 m
Overall length	163 m
Maximum beam	22.0 m
Wetted surface area	323 m^2
Midship cross-sectional area	86.0 m^2
Block coefficient	0.312
Prismatic coefficient	0.606
Midships section coefficient	0.515

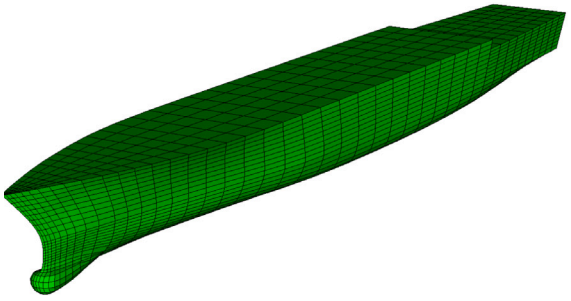

Fig. 5. The ONR Topsides Flared hull used for the LAMP simulations.

3. Nonlinear ship responses using LAMP

3.1. Overview

To characterize the ship hydrodynamics and structural responses we employ the Large Amplitude Motions Program (LAMP) v4.0.9 (May 2019). LAMP is numerical solver that computes 3-D potential flow solution of the wave-body interaction problem in order to calculate the time-domain motions and loading of floating bodies (Shin et al., 2003; Lin et al., 2007b,a, 2010). We note that our analysis does not rely on the specifics of LAMP and, in principle, any hydrodynamic structural simulation or tow tank physical experiment could replace LAMP, with perhaps minor changes to the format of the sea state representation, or a different numerical fidelity/computational runtime balance.

The primary setting for our simulation is a marine vessel with the Office of Naval Research Topsides flared variant geometry (Fig. 5) traveling with constant linear velocity at 10 knots through long crested (unidirectional) head seas with heading 180.0°. This symmetry restricts the available motions of the vessel to z displacement and pitch angle θ. Selected hydrostatic quantities are included in Table 1.

The induced Froude–Krylov forces (hydrodynamic pressure force) also induces a vertical bending moment (VBM), calculated in LAMP by rigid beam theory. Previous works have found that the distribution of VBM has an important asymmetry with non-trivial tails (Sapsis, 2020), so we will focus our statistical efforts at modeling the VBM statistics, calculated midship. On the other hand, the probability distribution for the pitch motion is very close to normal and for this reason we will not present relevant results.

3.2. Spatial wave episode representation

LAMP represents the episode as a sum-of-sinusoids with prescribed amplitude, frequency, and period. Our procedure above generates a time series — the wave episode, as well as a stochastic prelude and postlude. We transform between these representations using the Discrete Fourier Transform (DFT). We combine sine and cosine modes with identical frequency to get their phase $\phi_i = \text{atan}(a_i/b_i)$ (using signed

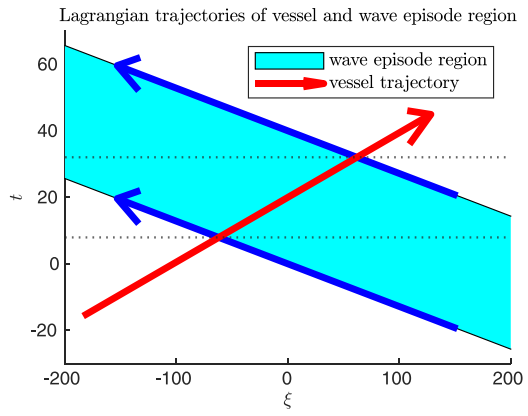

Fig. 6. Schematic representation of the propagating wave episode region in space and time, with overlaid trajectory of the vessel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

arc tangent), and choose the n_{LAMP} modes with the greatest combined amplitude.

The sum-of-sinusoids wave episode representation fully determines the wave elevation at $\xi = 0$. In order to extend this to other spatial locations, we employ for each of the traveling monochromatic waves the representation:

$$\cos(k_i\xi - \omega_i t + \phi_i), \tag{25}$$

where k_i is determined as the (positive) solution to the dispersion relation

$$\omega^2 = gk\tanh(kh). \tag{26}$$

For our data collection, we have chosen $n_{LAMP} = 1024$ to ensure a high-quality DFT reconstruction of the wave episode. Finally, over our time scale of interest (T), dispersive effects do not lead to significant changes in the shape of the propagating wave episode.

3.3. Effect of ship velocity

In our experimental setup, we use LAMP to simulate a marine vessel traveling at constant velocity through long crested (uni-directional) headwaves. This will create two complications, that must be carefully handled.

First, we must choose the position of the vessel at the beginning of the simulation in order to control when and where the vessel travels across the wave episode. For simplicity, we construct the wave episode to be spatially located around $\xi = 0$ at time $t = 0$. We begin the simulation at time $t_0 < 0$, in order to handle transient and initialization effects. Therefore, we should locate the vessel near $\xi_0 = t_0 v_x$ at the beginning of the simulation, so that the vessel passes through position $\xi = 0$ at time $t = 0$.

In Fig. 6, we show the trajectory of the vessel in red. In this model, we fix the *start* of the wave episode at $t = 0$, so that the vessel should pass through the origin at $t = \frac{T}{2}$. This leads to the choice of initial position

$$\xi_0 = \left(t_0 + \frac{T}{2}\right) v_x. \tag{27}$$

The second complication is the switch between Eulerian and Lagrangian reference frames for the *duration* of the wave episode. In the Eulerian reference frame, the wave episode at $\xi = 0$ begins at $t = 0$ and ends at $t = T$. However, as the wave episode travels, it has different arrival and departure times at different spatial locations.

In the Lagrangian frame, the vessel first intersects the wave episode 'earlier in space' than $\xi = 0$. This means that the time of intersection must be slightly *later* than $t = 0$. Likewise, the vessel exits the wave episode at $\xi_e > 0$, which occurs *before* $t = T$. We call this Lagrangian duration, T_L.

In Fig. 6, the gray dotted line represents the duration of the wave episode in the Lagrangian frame — that is, the duration between the vessel's entrance into the wave episode and its exit. Note that compared to the full 'width' of the shaded region, the gray lines are separated by the shorter time interval, T_L. While graphically it is apparent that the $T_L < T$ (for the case of head seas), how ought we determine its precise value?

We will calculate using the relative velocities. Let D be the distance that the wave episode will travel in time T, or

$$D = T v_G, \tag{28}$$

where v_G is the velocity of the traveling wave episode. For the JONSWAP spectrum, we approximate the velocity of the traveling wave episode by the group velocity corresponding to the spectrum modal peak. The Lagrangian duration is the time it takes for the counter travelling wave episode and vessel to meet if they began separated by the distance D. That is

$$T_L = \frac{D}{|v_S - v_G|}, \tag{29}$$

where v_S is the velocity of the vessel, which is assumed to have opposite sign as the wave episode velocity. For our simulation parameters, we find that T_L is approximately 40% the duration of T. Note that in the Lagrangian frame, the vessel sees approximately the same number of wave periods, but that the wave episode is 'compressed,' and the Lagrangian-effective wave period is shorter.

3.3.1. Selection of the stochastic prelude length

The stochastic prelude length, T_{SP}, has to be sufficiently large to bring the system in its statistical steady state, i.e. it has to be at least as large as the decorrelation time of the system. On the other hand, we do not want to choose T_{SP} very long, as this will lead to large computational cost of numerical simulations and also large cost to perform the Gaussian process extrapolation. Taking these considerations into account, we chose stochastic preludes of length $T_{SP} = T$ both before and after the wave episode. We also include a 'postlude' to ensure that no Gibbs phenomena rise when we transform the resulted wave form in DFT. We note that while we experimented with longer stochastic preludes (up to $T_{SP} = 4T$), we did not notice significant effects from lengthening the preludes.

4. Reduced order modeling with Gaussian process regression

4.1. Overview of Gaussian Process Regression (GPR)

GPR involves the construction of probabilistic surrogate model for set of data-points produced by the map:

$$y = \mathcal{F}(\boldsymbol{\alpha}) + \epsilon \tag{30}$$

where y is the output of interest, e.g. a modal coefficient, $\boldsymbol{\alpha}$ is the wave episode parameters, and ϵ is observation noise with intensity σ_n, such as the effect of the stochastic prelude on the output.

The training data set consists of the input data $\mathcal{A}$, and the corresponding output data, $\mathcal{Y}$. The input data is a $n_s \times n$ matrix, where the rows correspond to n_s distinct samples, and the columns correspond to the dimensionality of the wave episodes. We refer to a single sample from $\mathcal{A}$ as the n dimensional vector $\boldsymbol{\alpha}_i$. The output data $\mathcal{Y}$ is a $n_s \times n_{out}$ matrix, with n_s distinct samples and n_{out} output variables. For our analysis we reduce the problem of vector output to n_{out} independent scalar problems.

The training process, described below, produces a surrogate model that takes a new (i.e. not included in the training) wave episode $\boldsymbol{\alpha}$ and provides a distribution (posterior) for the output $\hat{y}(\boldsymbol{\alpha})$. By construction

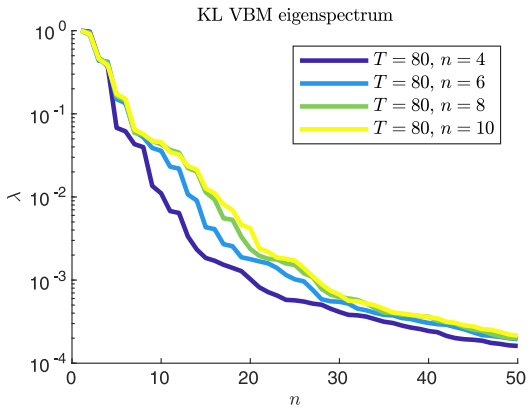

Fig. 7. Eigenspectrum of the KL basis associated with VBM intervals corresponding to different values of the wave episode n parameter.

this output distribution is Gaussian, with mean $\bar{y}(\boldsymbol{\alpha})$ and variance $\sigma_y^2(\boldsymbol{\alpha})$. That is to say

$$\hat{y}(\boldsymbol{\alpha}) \sim \mathcal{N}(\bar{y}(\boldsymbol{\alpha}), \sigma_y^2(\boldsymbol{\alpha}); \theta), \tag{31}$$

where the argument θ represents the parameter set for the Gaussian distribution.

The starting point for GPR is a prior normal distribution for the random field, $\hat{y}(\boldsymbol{\alpha})$. Specifically, we assume a zero mean prior and a covariance given by a kernel $k(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$. Conditioning this prior to the data leads to closed expressions for the mean and the variance (Rasmussen and Williams, 2006):

$$\bar{y}(\boldsymbol{\alpha}) = K_*^\intercal (K + \sigma_n^2 I)^{-1} Y \tag{32}$$

$$\sigma_y^2(\boldsymbol{\alpha}) = K_{**} - K_*^\intercal (K + \sigma_n^2 I)^{-1} K_* \tag{33}$$

where,

$$\begin{aligned} K_{i,j} &= k(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j), \\ K_i^* &= k(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}), \\ K^{**} &= k(\boldsymbol{\alpha}, \boldsymbol{\alpha}). \end{aligned} \tag{34}$$

Note that the quantities defined in Eqs. (34) are a $n_s \times n_s$ dimensional matrix, a $n_s \times 1$ dimensional vector, and a scalar respectively.

4.1.1. Kernel and hyperparameters

The kernel function or prior covariance, $k(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, is an abstract measure of similarity, conceptually similar to a metric. Here, we consider the squared exponential kernel (or exponential quadratic), given by

$$k_{SE}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|^2}{2l^2}\right). \tag{35}$$

The squared exponential kernel is smooth (infinitely differentiable), and has two parameters: σ, the magnitude of the uncertainty (a scale factor); and l, the length scale of the uncertainty fluctuations. We identify the ordered tuple $\theta = (\sigma, l)$ as the parameter set for this kernel. For vector inputs, a rescaling of the input data can lead to improved results but also a significantly larger number of hyperparameters: the *Automatic Relevance Determination* (ARD) kernel,

$$\begin{aligned} k_{SE}^{ARD}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) &= \sigma^2 \exp\left(-\frac{1}{2}\sum_i^n \frac{(\alpha_{1,i} - \alpha_{2,i})^2}{l_i^2}\right) \\ &= \sigma^2 \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)^\intercal M (\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\right), \end{aligned} \tag{36}$$

where $M = \mathrm{diag}(l_i^{-2})$. This kernel function can distinguish between the importance of different input dimensions. As a consequence, the dimension of the parameter set θ grows from 2 to $n+1$. For k_{SE}^{ARD}, θ is given by the ordered tuple $(l_1, l_2, \ldots, l_n, \sigma)$.

4.1.2. Choice of hyperparameters

In order to optimize a set of parameters, we need to define a loss function. For GPR, the natural loss function is the (negative) log marginal likelihood, given by

$$\begin{aligned} \log p(\mathcal{Y}|\mathcal{A}; \theta) = &-\frac{1}{2}(\mathcal{Y} - \bar{y}(\mathcal{A}))^\intercal (K(\theta) + \sigma_n^2 I)^{-1} (\mathcal{Y} - \bar{y}(\mathcal{A})) \\ &- \frac{1}{2}\log|K(\theta) + \sigma_n^2 I| - \frac{n}{2}\log 2\pi. \end{aligned} \tag{37}$$

This loss function is non-convex. To this end, we employ a conjugate gradient descent method. The presence of local optima causes the final 'optimized' solution to depend on the specific implementation of the optimizer and starting point–hyperparameter choices. To address the problem of local optima we perform the optimization procedure multiple times with different choices of starting θ.

4.2. Representation of the response time series and transfer learning for different sea spectra

A LAMP simulation with a parameter vector $\boldsymbol{\alpha}$ produces a *time series* of the VBM. We cannot simply train a Gaussian Process model to 'learn' a time series, so we require an intermediate processing step that approximates a time series segment by a finite-dimensional vector.

The first option is to characterize the VBM time series through a scalar quantity. For instance, one may choose that an adequate representation of the VBM time series is the 'interval-max' (perhaps combined with 'interval-min'). That is, if we are interested in the maximum value that the VBM takes on our interval of analysis, then we have reduced the problem of modeling a time series to the simpler one of modeling a (small set of) scalar(s). The challenge with this approach is that the scalar summaries quantities that are most appropriate from a simulation perspective – mean, variance, interval extrema – are *not* the most appropriate quantities from a statistical perspective. In particular, mean and variance are not enough to reconstruct a pdf for the VBM, and the relationship between interval extrema and local extrema is very complicated, especially in the context of irregular waves.

The alternative is to represent the time series directly by a low rank vector, each component of which is a scalar that can be modeled with a Gaussian process surrogate. While this idea naturally suggests a Fourier series representation, the previously developed machinery of the KL decomposition is an even better fit. Specifically, we employ the following representation for the VBM:

$$M_y(t|\boldsymbol{\alpha}) = \sum_{i=1}^{n_{out}} q_i(\boldsymbol{\alpha})\hat{\mu}_{i,T}(t), \quad t \in [0, T]. \tag{38}$$

where $q_i(\boldsymbol{\alpha})$ are the KL coefficients which are functions of the excitation wave episode and $\hat{\mu}_{i,T}(t)$ are the VBM KL modes defined over the wave episode duration. Using the KL expansion as our low rank representation of the (output) VBM time series has two main advantages over Fourier series. First, the KL expansion is optimal in the sense that is guaranteed to contain the most energy in a finite truncation. This means we should expect to require fewer KL modes to adequately reconstruct the output time series that we would need had we used Fourier modes. Second, KL coefficients are linearly decorrelated. While we do *not* expect or assume that the VBM or the modal coefficients will follow Gaussian statistics, the KL coefficients will reduce the intermodal dependencies that we must account for otherwise. While it is possible to model the intermodal correlations directly (see Álvarez et al., 2012), in general the KL coefficients are close enough to independent (for the considered problem) to treat separately.

The drawback of the KL expansion is that we require adequate second order statistics in order to calculate the eigendecomposition of the covariance matrix for the output time series. For this work, we assume access to such information, which we calculated via experimental realizations of LAMP with randomly selected wave episode parameters. In general, we expect both that accurate second order statistics are

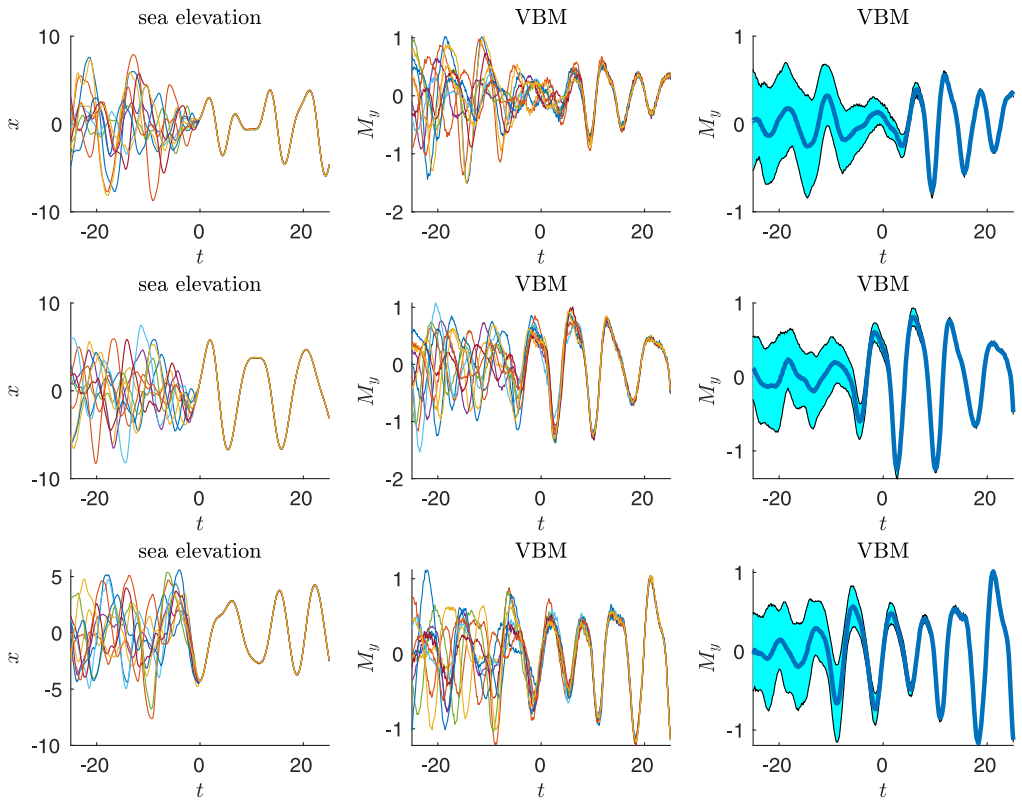

Fig. 8. First column: collection of stochastic preludes with a common wave episode region — each row shows a different wave episode. Second column: collection of corresponding VBM time series. Third column: mean and 1σ spread of the VBM. Note that time is measured with respect to the Lagrangian frame, i.e. with respect to the ship.

simpler to acquire, and that poorly resolved KL modes are *no worse than* Fourier modes. In particular, Principle Component Analysis (PCA) of the output stress data from many experiments will usually provide reasonable approximations to the KL modes.

Finally, an important parameter is the number of output modes necessary for accurate statistical reconstruction. The effects of varying the number of retained outputs modes is discussed in Section 6.3. For this work, we use $n_{out} = 12$, unless otherwise noted.

Finally, we emphasize a very important property of the developed approach related to the application of the trained GPR scheme for different sea spectra. In particular, the coefficients, $q_i(\boldsymbol{\alpha})$ do not depend *directly* on the specific input spectrum, i.e. on the probability distribution of $\boldsymbol{\alpha}$. Rather, the coefficient model represents a conditional description of the VBM *given* a wave episode with prescribed shape, which may represent relatively steep waves in a calm sea or relatively shallow waves in a storm. Thus, after the GPR scheme has been trained, the distribution of the modes, q_i, or the VBM can be trivially computed by employing the specific probability distribution for $\boldsymbol{\alpha}$, associated with the spectrum under consideration.

This transfer learning property assumes that the output KL modes for the VBM do not change significantly as the spectrum varies and also that the training of the GPR scheme has been performed with appropriate training data so that it is accurate for any value of $\boldsymbol{\alpha}$ that has non-negligible probability for the spectra considered. In other words, if one trains with wave episodes $\boldsymbol{\alpha}_j$ that have non-negligible probability for a high-energy sea spectrum, the trained GPR coefficients will also be accurate for wave episodes associated with low-energy sea spectra. Of course, the inverse is not true, although in this case one can complement the existing GPR scheme with new wave episodes so that it increases its accuracy for the high-energy spectrum.

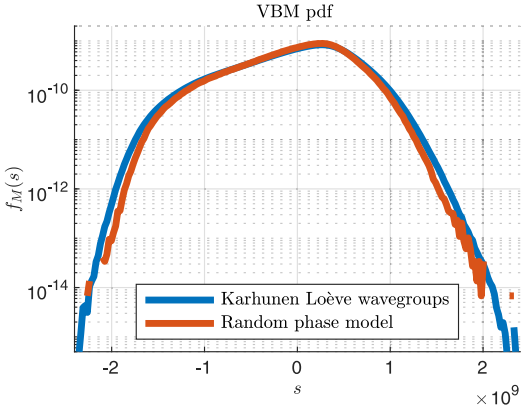

Fig. 9. Comparison of long-time Monte-Carlo histogram of the VBM (based on the random phase model) with short-time KL Monte-Carlo histogram of the VBM (based on the wave episode method). For the wave episode-based Monte-Carlo, we have $T = 60$ and $n = 30$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This spectrum transfer capability does have limits. In the simplest case, we may imagine two spectra which differ only in the significant wave height. For this case, the only source of error is a slight mismatch in the energy of stochastic prelude, and the spectrum transfer statistics are high quality. In the most complicated case, where the two spectra differ in modal period, the interprojection of the respective KL bases will be difficult, and require a very large n. For this reason, we recommend spectrum transfer data and reuse only in sea states

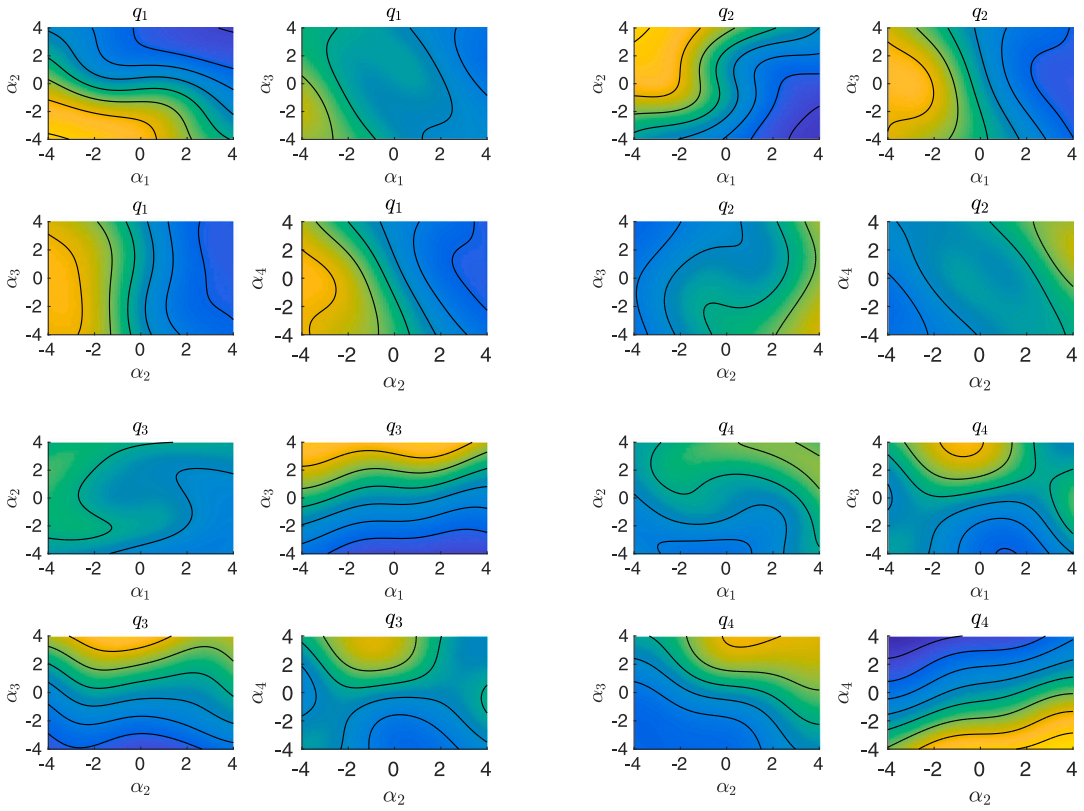

Fig. 10. Projections of the trained Gaussian process surrogate for $(n, T) = (4, 60)$. (a–d) Output modes are denoted as q_1 through q_4, while input wave episode parameters are denoted as α_1 through α_4. For each subplot shown, two input parameters vary while the other two are set to zero. All colors follow the same palette corresponding to $q_i \in [-4, 4]$.

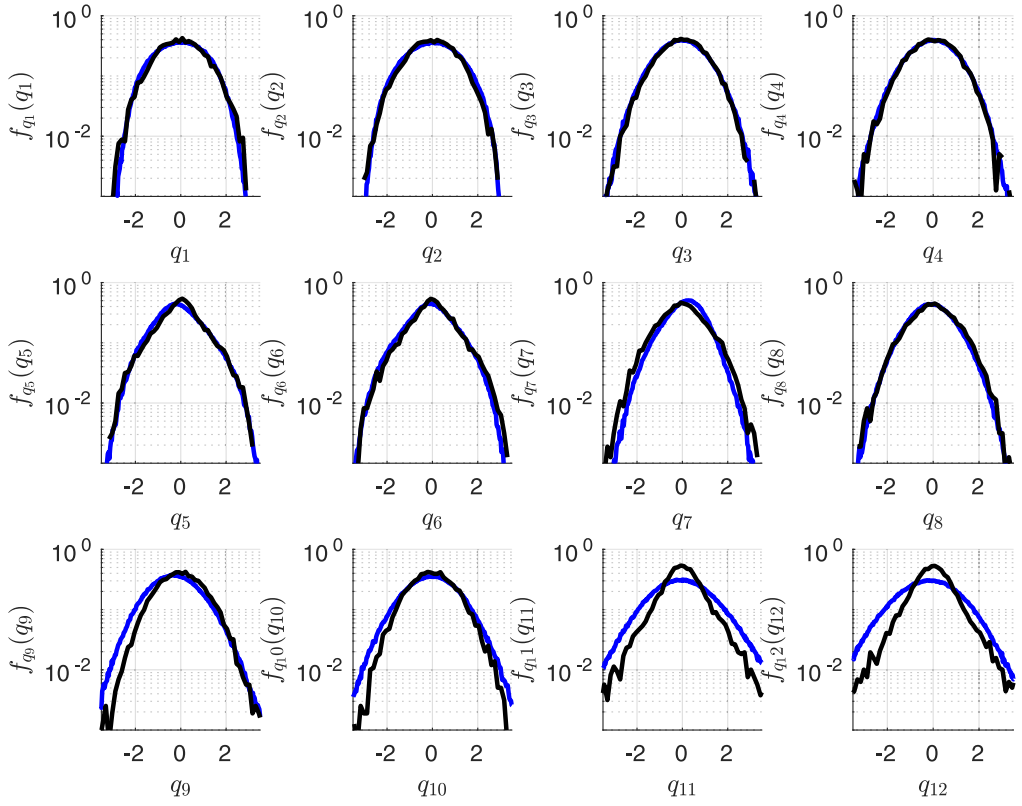

Fig. 11. Reconstructed pdf using the GPR for each output KL mode (blue), compared against the corresponding short-term KL Monte Carlo histogram (black). wave episode parameters are $(n, T) = (4, 60)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

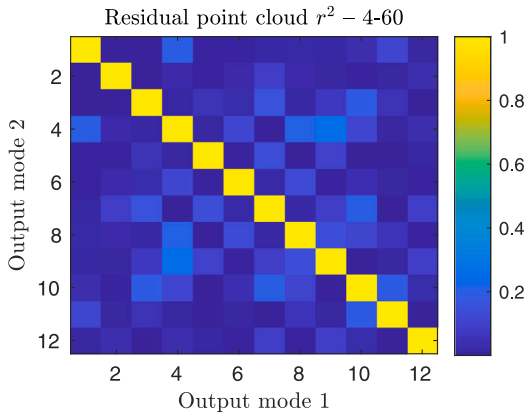

Fig. 12. Correlational structure of the surrogate residuals for $(n, T) = (4, 60)$. As the KL procedure decorrelates the VBM modes, this is primarily a validation check.

with quite similar modal periods. Finally, we have not investigated the intermediate situation where the modal period is identical but the spectral bandwidth differs. In this situation, a multi-fidelity approach to data reuse may be appropriate, in order to maximize data reuse while minimizing model errors.

To summarize, LAMP generates time series of the VBM as output for given wave episodes. The VBM time series is approximated through a set of KL coefficients. We use these data points to train a separate Gaussian process surrogate model for each output KL coefficient. To draw samples, we sample a coefficient from each GPR model separately, and then reconstruct the VBM time series using Eq. (10). Any statistical techniques we wish to perform (histograms, identify local extrema, etc.) are performed on these generated time series.

5. Evaluation of the reduced order model for extreme event statistics

5.1. Dimensionality of the wave episode space

We first choose the dimensionality of the wave episode space. During the construction of the wave episode region, the two most significant choices are the values of T, the length of the wave episode region, and n, the number of activated (nonzero) KL modes. By varying T, we change the size of the KL covariance matrix, and by extension both the shape of the KL modes and the associated eigenvalues. By varying n, we change the number of retained KL modes, which affects the energy content of the wave episode region as well as the dimensionality of the wave episode space. We choose as typical values $T = 40$ or $T = 80$, and $n = 4$ or $n = 6$. These choices guarantee sufficiently long wave episodes so that these exceed the memory of the system. Moreover the number of retained modes is large enough for each case of T, so that the energy of the random signal is captured.

We note that, when referring to Fig. 2, a truncation order of $n = 4$ or $n = 6$ may capture as little as 60% of the wave energy. However, we are interested in capturing the energy of the VBM signal, for which $n = 4$ seems to be generally sufficient. This paradox is addressed by Sapsis et al. (2020, 2021), where even sinusoidal waves (roughly equivalent to $n = 2$) were able to capture most of the energy of the VBM signal, and is probably due to weak coupling between high frequency wave components at the vessel-hull VBM.

5.2. Comparison between the random phase model and the random wave episodes representation

Our first objective is to establish that the KL parametrization of wave episodes combined with spectrum-consistent stochastic-preludes is a valid approach, i.e. that it agrees with the statistics obtained through the random phase model.

In Fig. 2, we present the KL eigenspectrum for the wave episodes with different values of T. We note the steady power law decay in mode energy, which is faster for smaller values of T. Moving our attention to the response, in Fig. 7, we present the corresponding KL eigenspectrum decay for the *VBM*, associated with different truncation orders of the KL series describing the wave episodes, and for $T = 80$. We note two properties: First, the VBM eigenspectrum has a fast initial decay, compared the wave episodes spectrum. We associate this behavior with the distribution of damping across different modes, as well as possible nonlinear dynamics that transfer energy from low order modes to higher modes. While KL modes do not have a simple characterization in terms of frequency or wavenumber like Fourier modes, it is generally true that higher order KL modes have higher frequency and wavenumber content. Second, the rate of decay of the VBM eigenspectrum depends on the order of truncation of the KL expansion for the wave episodes. In particular, the VBM eigenspectrum decays faster when there are fewer activated wave elevation modes, that is, when the wave episode is more regular. We can interpret this in a dynamical sense as showing that restricted wave episodes are prevented from exciting higher response modes directly.

Next, we examine how the ship dynamics respond to different stochastic preludes. In Fig. 8 we show in the first column three different wave episodes (one for each row) together with various stochastic preludes. In the second column we present the corresponding VBM time series realizations, and finally the mean and one σ spread of the

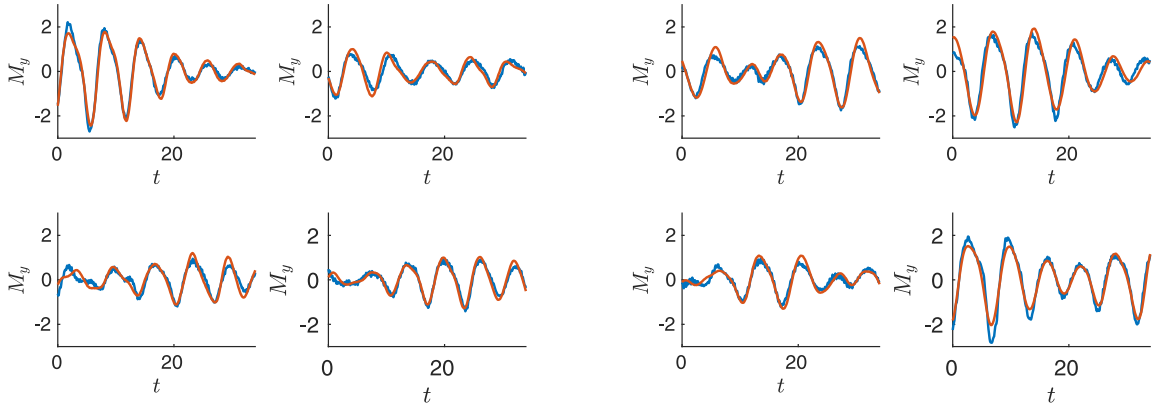

Fig. 13. Sample time series of the LAMP output (blue) and the Gaussian process reconstruction (red) for VBM and $n_{out} = 12$. Left four: $T = 80$, $n = 4$. Right four: $T = 80$, $n = 6$. Vertical axis expressed in dimensionless units; horizontal axis is scaled to the duration of the Lagrangian fixed region T_L. LAMP testing data unseen by model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

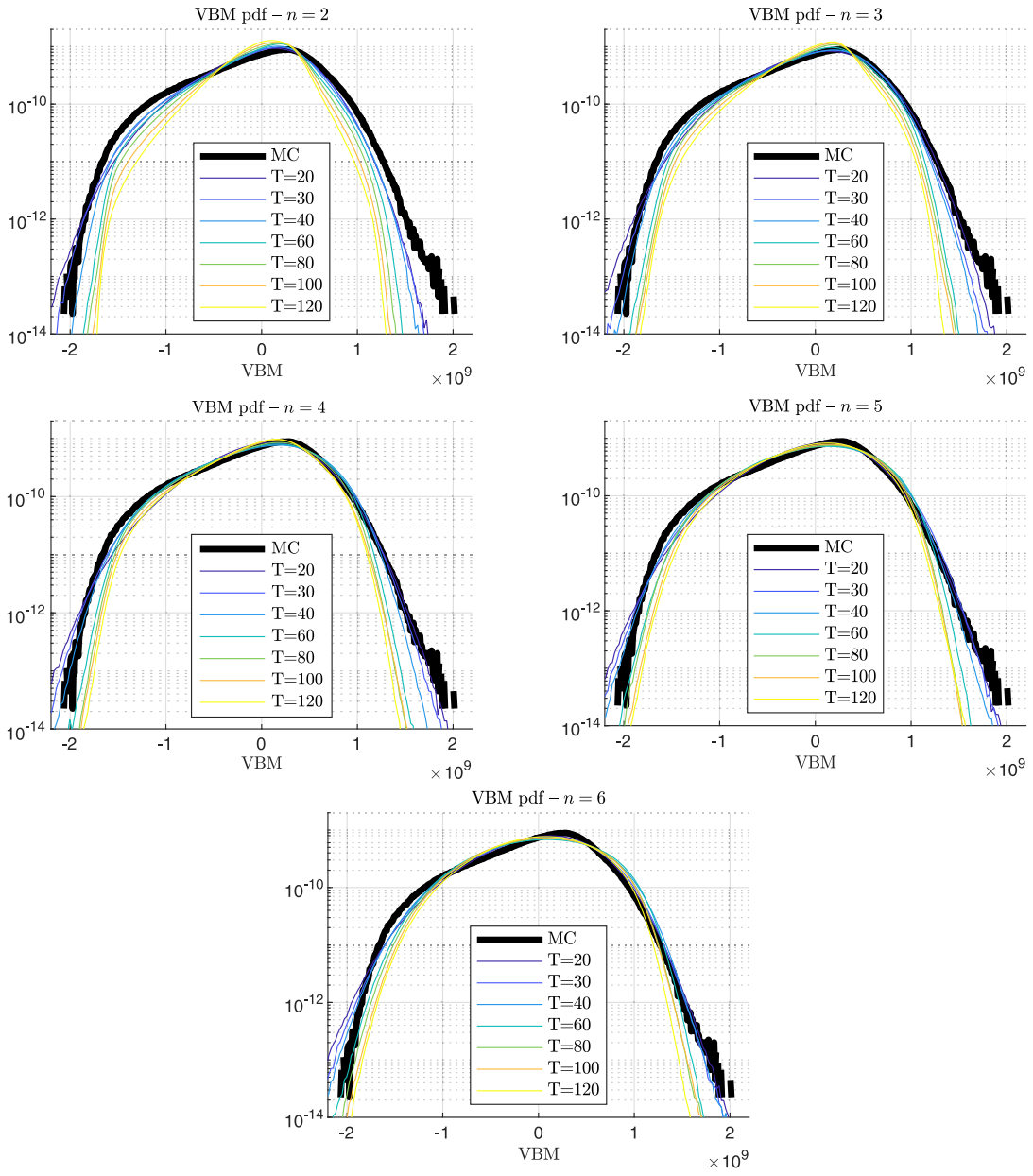

Fig. 14. Reconstructed VBM plots for varying wave episode length, T (colors), and wave episode KL modes, n (plots).

VBM ensemble (third column). For this numerical experiment, we have $T = 40$ and $n = 25$ (more than 99% retained wave elevation variance). In the first column, we can see dramatically how the wave episodes all come together at $t = 0$, recapitulating the ensemble variance displayed in Fig. 4. In the second column, we see both that the convergence is delayed (because of the Eulerian/Lagrangian shift), and spread out (because of the memory effect). Nonetheless, by consulting the third column we can see that the ensemble spread of the VBM due to different stochastic preludes does indeed decrease with time. Because the vessel response is largely (but not entirely!) independent from the stochastic prelude during data collection on the Lagrangian fixed region, we will generally perform simulations with only one stochastic prelude for each wave episode. This is sufficient since in the surrogate construction we will employ responses by numerous wave episodes and this will give information for the induced uncertainty. The single stochastic prelude will be selected randomly, according to the extrapolation procedure described above in Section 2.4.1.

Finally, we directly compare the Monte-Carlo estimates of the VBM pdf from long-time simulations using the random phase model to short-time wave episode based simulations. The random phase model Monte-Carlo simulation consists of 3640 separate simulations, each of which lasted a duration of $T = 3000$ seconds. The sea surface elevation was initialized directly with random phase model sample from the JONSWAP spectrum based on a discretization from $\omega_{\min} = 0.31$ s^{-1} to $\omega_{\max} = 10.2$ s^{-1} with $n_{LAMP} = 1024$ distinct sinusoids. A short duration "ramp-up time" process is used to avoid initial transients, but unlike the wave episode simulations does not include a stochastic prelude. Subsequently, we histogram the vessel VBM and produce the long-time steady-state Monte-Carlo estimate.

We note that, to construct the long-time steady state statistics, we assume ergodicity, at least after discarding a short period of initial transients. The short wave episode simulations similarly have a discarded initial transient period – the stochastic prelude – but they are intrinsically defined on the finite interval of length T, and therefor

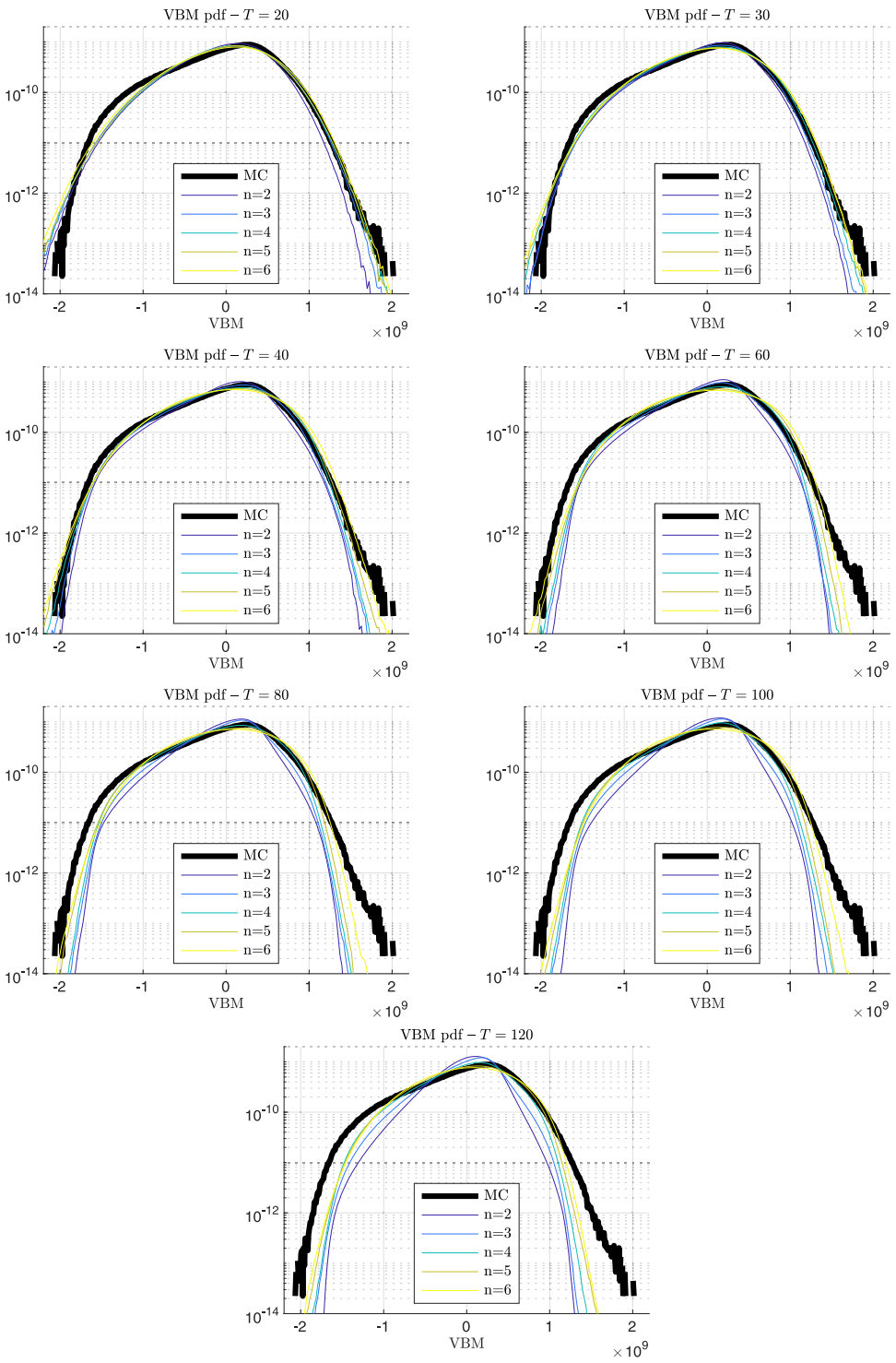

Fig. 15. Reconstructed VBM plots for varying wave episode KL modes, n (colors), and wave episode length, T (plots).

do not have a simple ergodic equivalence. Nonetheless, the ergodic behavior is the motivation for the program of estimating long-time steady state statistics using the statistics for many short intervals.

For the wave episodes based approach, we calculate the KL basis for the same JONSWAP spectrum, using $T = 60$. We sample 24000 wave episodes, each of which is generated with $n = 30$ activated (nonzero) KL modes. For this sampling step, we draw $\boldsymbol{\alpha}$ from a multivariate Gaussian distribution with variance given by the KL λ_i. For this demonstration, we choose n large enough to retain 99% of the wave elevation energy, which ensures that the wave episodes are high fidelity representations of the JONSWAP spectrum. Finally, for each generated wave episode we simulate the ship dynamics and extract the VBM. This method of producing the VBM is short-time KL Monte-Carlo.

In Fig. 9, we display the pdf obtained by these two Monte-Carlo estimates for the vessel VBM: the wave episodes based Monte-Carlo in blue, and the random-phase-model-based Monte-Carlo in red. We note that the two pdf estimates have very close agreement out to the tails, where the wave episodes based Monte-Carlo slightly underestimates the tail decay. This small discrepancy can be attributed to the finite

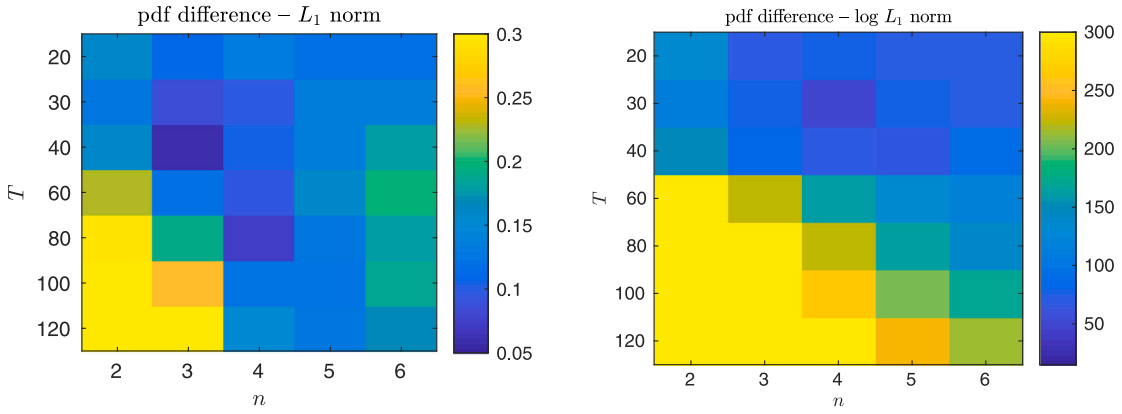

Fig. 16. Pdf error estimates between reconstructed VBM and wave episode Monte-Carlo for different choices of T and n; Left: ϵ_1 (l_1 norm); Right: ϵ_2 (l_1 norm of the difference of logs).

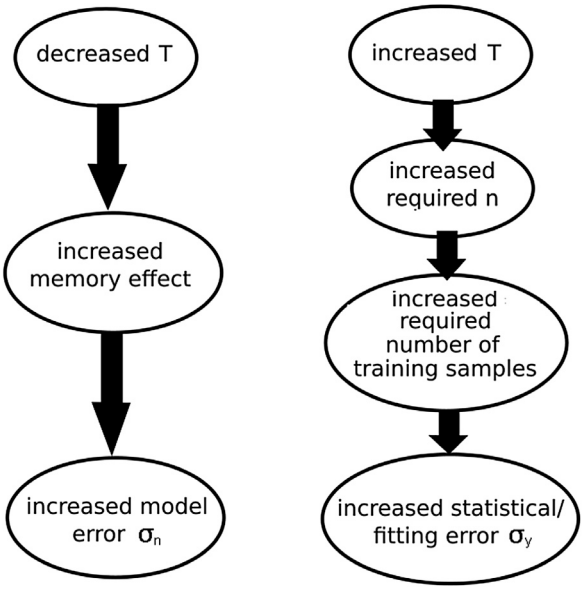

Fig. 17. Flowchart representing the balance between T and n and their effect on error in statistical reconstruction.

truncation order of the wave episode approach which neglects much higher order modes.

5.3. Surrogate modeling of output KL coefficients using GPR

GPR produces a surrogate model, an inferred function, that maps the wave episode characteristics, $\boldsymbol{\alpha}$, to the conditional expected mean of the output $\hat{q}$, written $\mu_i(\boldsymbol{\alpha}) = \mathbb{E}[q_i|\boldsymbol{\alpha}]$. During the learning process of the GPR surrogate we also obtain the noise parameter, σ_n, associated with the uncertainty induced by the stochastic prelude of each wave episode. We train the Gaussian process surrogate on $n_s = 625$ uniformly distributed samples.

In Fig. 10, we show four different projections of the Gaussian process surrogate (each with a different choice of $\alpha_i \neq 0$ and $\alpha_j \neq 0$) for each of the first four output modes (q_1 through q_4). For this demonstration, we have chosen $n = 4$ and $T = 60$. We note that, in broad strokes, each q_i is approximately linearly dependent on some of the α_i, and roughly independent of others. This rough correspondence, however, is not exact due to the nonlinearities in the surrogate model. In particular, we notice that the shape of the surrogate (corresponding to the estimate of $\mathbb{E}[q_i|\boldsymbol{\alpha}]$) is frequently different near the origin (small $|\boldsymbol{\alpha}|$ and low-amplitude wave episodes) compared to the edge of the domain (large $|\boldsymbol{\alpha}|$ and high-amplitude wave episodes). This is an important confirmation that learning the typical response for marine vessel hydrodynamics requires different data sets than learning extreme responses.

Next, we leverage the computed surrogates for two purposes. In Section 5.4, our primary thread is to use these models to parsimoniously estimate the output statistics from the input statistics. In Section 5.5, we take a detour to demonstrate the reconstruction of *particular* output time series.

5.4. Statistics for the output KL coefficients

In this section, we examine the GPR for each output mode. While our goal is the recovery of the VBM pdf, as an intermediate quantity we calculate probability distributions for each output KL mode during reconstruction. In Fig. 11, we show the pdf for each output mode computed using the GPR surrogate (blue curves), compared to the short-term KL Monte-Carlo histogram (black curves), obtained by directly computing the ship response through LAMP over a large number of wave episodes. We note that there is some variability in the shapes of the different mode pdfs, but that the Gaussian process surrogate recovers the shape very well for each of the modes up to $n_{out} = 8$. For the higher modes, the pdf shape recovery begins to break down. This is likely due to the fact that more training data is needed for these higher order modes. However, as shown in Fig. 7 and later in Section 6.3, the contribution of the higher output modes to the reconstructed VBM is negligible.

While the KL theorem guarantees that the projection coefficients of a random process are linearly uncorrelated, zero correlation implies independence only for Gaussian processes — and this is not the case here. In Fig. 12, we examine the Pearson correlation coefficient of the Gaussian process surrogate *residuals* (pointwise errors). We see that the residual error of all mode pairs is close to zero. This implies that, for this process, a vectorized Gaussian process model would not provide any significant improvement compared to the adopted approach to model each output mode as a separate GPR. Indeed, repeating the presented analysis using vector kernel techniques did not improve the obtained results.

5.5. Reconstruction of VBM in the time domain

Here we focus on the reconstruction of the *time series* of the vessel VBM for any wave episode vector, $\boldsymbol{\alpha}$. We rely on the VBM expansion in Eq. (38) in terms of the coefficients $\mathbf{q}(\boldsymbol{\alpha})$, which have already been determined through the GPR surrogate. To obtain samples of the

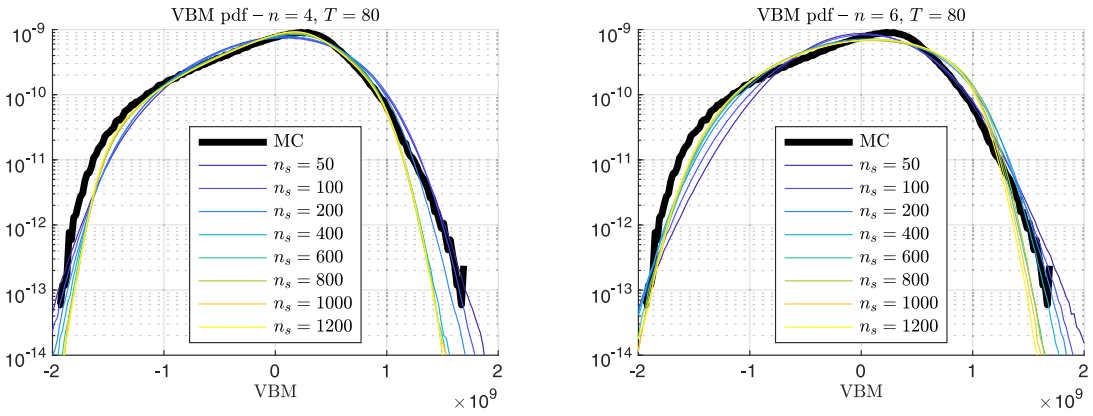

Fig. 18. VBM reconstruction for different sizes of training sets. Training points are uniformly chosen on a box with radius parameter $z^* = 4.5$; Left: $T = 80$, $n = 4$; Right: $T = 80$, $n = 6$.

coefficients we have:

$$\hat{q}_i(\boldsymbol{\alpha}) = \mu_i(\boldsymbol{\alpha}) + Z_i \sigma_i(\boldsymbol{\alpha}), \quad i = 1, \dots, n_{out} = 12, \quad \text{where} \quad Z_i \sim \mathcal{N}(0,1). \tag{39}$$

In Fig. 13, we compare the true VBM output from LAMP (in blue) with the reconstructed VBM output (in red), based only on the wave episode parameter vector $\boldsymbol{\alpha}$. We consider two wave episode types with $T = 80$ and also $n = 4$ (left four subplots) and $n = 6$ (right four subplots). We note that there is good qualitative agreement. This match is poorer typically at the beginning of the wave episode ($t \approx 0$), where the memory effects from the stochastic prelude are more pronounced. Finally, we note that the reconstructed time series does not recover any of the high frequency components of the VBM that LAMP calculates, which is no surprise given that we are using a relatively small number of KL modes to represent the VBM. Based on this construction of the VBM in the time domain, we can obtain any desired statistics using Monte-Carlo on the GPR surrogates, such as peaks, upcrossings, or value statistics.

5.6. Computational costs

Since one motivation for replacing long-time steady state simulations is the computational cost, in this section we briefly describe the resources required for computing the VBM statistics.

Each 3000s LAMP simulation required approximately 1 hour of computational time on a single CPU core. For our steady state pdf, we computed 3640 distinct simulations across 16 cores, for a total simulation time of about 150 CPU-days, or just under 10 calendar days.

Conversely, each short wave episode simulation took approximately 5 minutes of computational time on a single CPU core. For each of our surrogate reconstructions, we used $n_s = 625$ wave episodes as a training set, Section 6.2 excepted. This summed to approximately 50 CPU hours, or just over 3 calendar hours when spread across 16 cores. This cost was repeated for each (T, n) pair we investigated.

One difficulty with directly comparing these time costs is that Monte-Carlo statistical uncertainty will always improve like $\sim n^{-\frac{1}{2}}$ with more data, while error associated with surrogate modeling is split between model error and statistical error. Nonetheless, we feel both that this order of magnitude of Monte-Carlo samples for the steady state is appropriate to resolve the pdf tails, and that $n_s = 625$ is a conservative requirement for surrogate construction, at least for $n \leq 4$ nonzero wave modes.

We note that a significant portion of the CPU cost for the wave episode approach is bound up in initializing the simulation and damping out transients, perhaps up to 60% of the cost of the short simulations. This initialization corresponds to only a negligible fraction ($\approx 2\%$) of the long time simulations. It is therefor likely that more optimization of simulation initialization would tilt the computational balance further in favor of wave episodes.

6. Statistics quality with respect to reduced order model parameters

We now focus on measuring the performance of the developed approach on quantifying response statistics for the value of the VBM time series. Specifically, we consider as ground truth the pdf of the VBM time series, f_{MC}, as obtained from a direct Monte-Carlo simulation using the random phase model (procedure described supra in Section 5.2). On the other hand, the approximate pdf, f_{GPR}, was computed using a large number of samples for $\boldsymbol{\alpha}$ and subsequently computing, $\hat{q}_i(\boldsymbol{\alpha}), i = 1, \dots, n_{out}$ and samples for the VBM time series with $n_{out} = 12$, except if otherwise noted.

To measure the overall fit, we compute the l_1 norm of the pdf difference on the interval $[m_l, m_u] = [-1.9, 1.5] \times 10^9$. To measure tail fit in particular, we compute the l_1 norm of the difference of pdf logs on the same interval. That is to say, we consider the two error metrics

$$\epsilon_1 = \int_{m_l}^{m_u} |f_{MC}(s) - f_{GPR}(s)| ds \tag{40}$$

$$\epsilon_2 = \int_{m_l}^{m_u} |\log f_{MC}(s) - \log f_{GPR}(s)| ds \tag{41}$$

We chose the particular finite endpoints in order to avoid calculations where the Monte-Carlo pdf is poorly resolved, in particular to avoid the numeric ill conditioning involved with the logarithm of zero. When we vary the precise value of the endpoints, we find that the absolute value of the metrics (especially ϵ_2) changed, but that the trends remain the same.

6.1. Effect of wave episode length and retained wave episode KL modes

Figs. 14 and 15 show the reconstructed pdfs for various combinations of wave episode length T and number of activated input modes n. We note that as T increases, the minimum wave episode KL modes, n, necessary for an adequate reconstruction increases. This is not a surprise since longer waver episodes have higher complexity. However, as the required number of modes increases the dimensionality of the Gaussian process is higher and, therefore, for the same number of samples, the GPR error increases. This leads eventually to larger pdf mismatch for long wave episodes.

On the other hand, if the wave episode length is too small, then the system has not enough time to dissipate the effect of the stochastic prelude. Therefore, the VBM time series is characterized by irreducible uncertainty (Fig. 8). This is reflected in the learned noise coefficient of

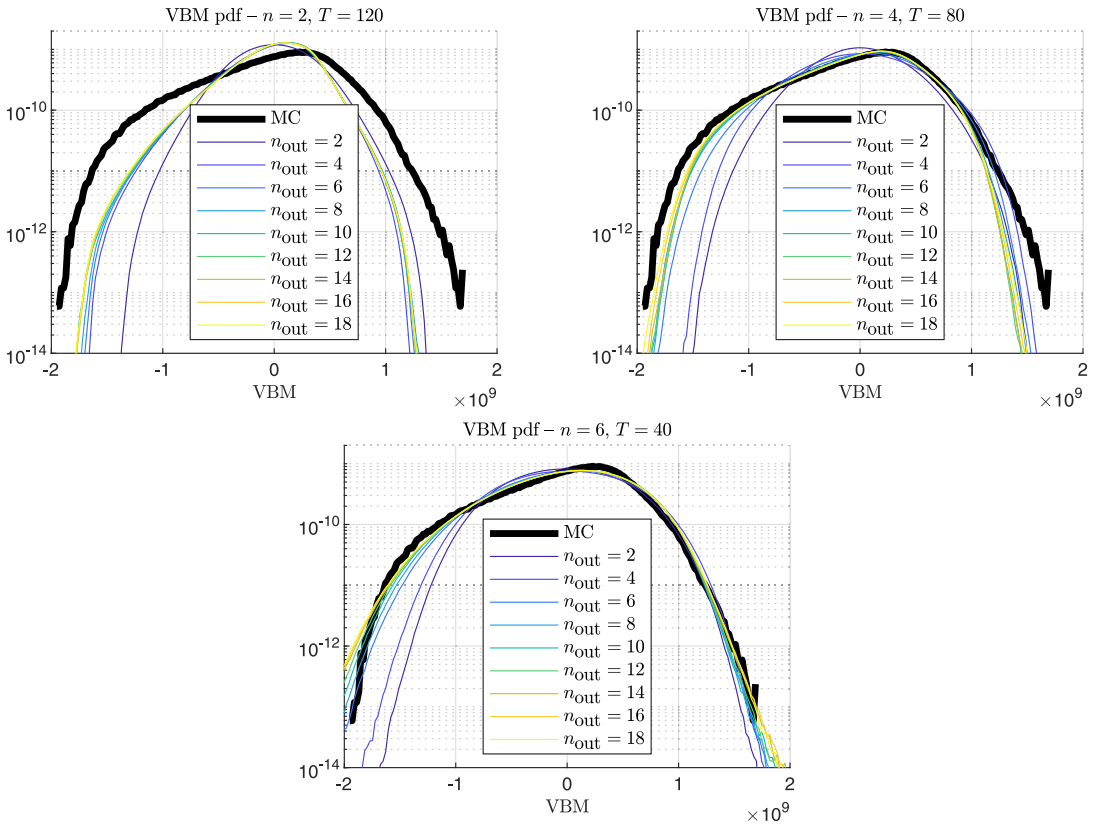

Fig. 19. Comparison of reconstructed VBM pdf for different choices of n_{out}, the number of retained output modes; (a) $n = 2$ activated wave episode KL modes and wave episode length, $T = 120$; (b) $n = 4$ activated wave episode KL modes and $T = 80$. (c) $n = 6$ wave episode KL modes and $T = 40$.

the GPR, $\sigma_{n,i}, i = 1, \dots, n_{out}$, which are, inevitably, increasing. The effect of higher noise coefficients leads to the convergence of the produced pdf to a Gaussian, i.e. to the 'smoothing' of the produced pdf.

For this reason there is an optimal combination of wave episode parameters, T and n. T should be sufficient to resolve the transient features but no longer, and n should be appropriate to capture the complexity of the wave episodes, but not larger than that as this will make the GPR regression more challenging than needed (Fig. 17). This optimal set of parameters can be seen in Fig. 16 where the two error metrics (l_1 and log l_1) are presented in terms of T and n. Considering all of these factors, we expect the optimal number of activated input modes to be $n = 3$, and the optimal duration of the parametrically fixed region to be $T = 40$.

Finally, we take a moment to describe the way that the reconstructed statistics do and do not match the steady state. At the mode of the distribution, the recovered pdf frequently underestimates the pronounced asymmetric peak, especially for $n \geq 5$. This is likely due the decreasing performance of GPR with higher dimensions — as the surrogate attributes more features to 'intrinsic noise,' the resulting statistics become more Gaussian. The other mismatch is the position of the tails. For $n \leq 2$, and especially for large T, too much wave energy is lost in the KL truncation to accurately model the large VBM peaks.

6.2. Size of the training set

We discuss the convergence of the reconstructed pdf with respect to the training set size. As described in Section 4.1, the training set is a collection of n_s input–output pairs corresponding to the parametrization $\boldsymbol{\alpha}$ and the VBM time series KL coefficients $q_i(\boldsymbol{\alpha}), i = 1, \dots, n_{out}$. We select the vector $\boldsymbol{\alpha}$ by Latin Hypercube Sampling (Olsson et al., 2003). Unlike simple Monte-Carlo samples, Latin Hypercube sampling requires an explicit box-shaped domain. We choose a n dimensional hyperbox with side length given by

$$d_i = 2z^* \sqrt{\lambda_i}, i = 1, \dots, n, \tag{42}$$

where λ_i is the KL eigenvalue corresponding to dimension i, and z^* is a radius parameter, akin the z-score of the standard Gaussian distribution, that specifies how many standard deviations away from the mean we will include. This analogy is justified, because the distribution of KL coefficients for Gaussian processes is jointly Gaussian with diagonal covariances given by λ_i. In all presented results, we use $z^* = 4.5$ to include extreme wave episodes. When we try our procedure with smaller $z^* = 3$, we found slightly better convergence near the distribution peak and moderately worse convergence near the tails.

In Fig. 18, we show the recovered VBM pdf for different values of n_s, and for two different wave episode parameters: $T = 80$, $n = 4$ and $T = 80$, $n = 6$. We note that for the $n = 4$ case, which we consider to result in good performance, the pdf converges very quickly, between 100 and 400 samples, depending on how far down the tails we consider. For the case $n = 6$, which we consider high dimensional, the convergence is somewhat slower. In particular, the characteristic shoulder asymmetry does not appear until around 200 samples, and the central peak shape requires more samples to converge to the ground truth.

For the results presented in the rest of this paper, we use $n_s = 625$. This choice balances computation time (both in data collection and GP fitting), with reconstruction fidelity. However, since our goal is to reconstruct the VBM statistics with minimal data, we consider the slow sample convergence of the high dimensional scenarios ($n = 5, 6$) to be part of the trade-off between T and n discussed supra in Section 6.1.

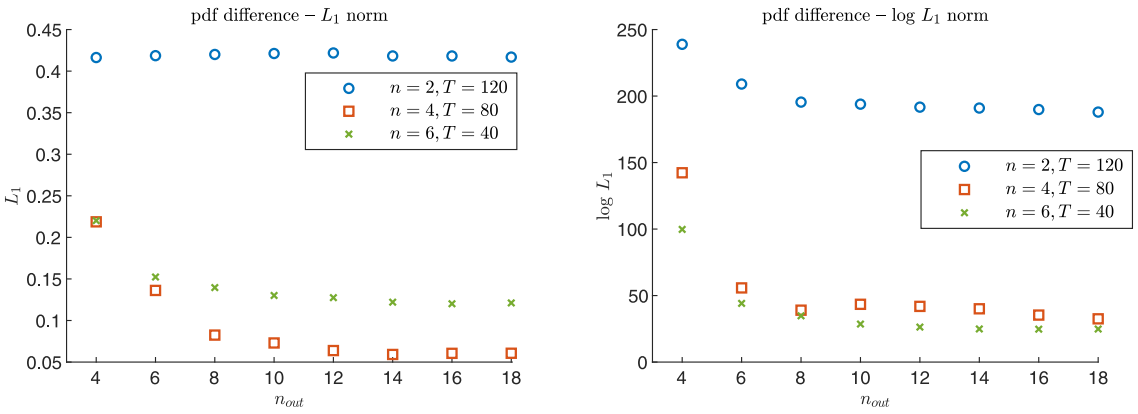

Fig. 20. Pdf error estimates between reconstructed VBM and Monte-Carlo for different choices of n_{out}, the number of retained output modes. Left: ϵ_1 (l_1 norm). Right: ϵ_2 (l_1 norm of the difference of logs).

6.3. Number of retained VBM KL modes

Here we examine the dependence of the reconstructed pdf on the number of output modes retained. Fig. 19 shows the reconstructed pdf for three choices of T and n and for various choices of n_{out} ranging from 2 to 18. We note fast convergence in the shape of pdf, consistent with the fast decay of the output eigenspectrum as shown in Fig. 7. Values of $n_{out} \geqslant 8$ appear to have a minor impact in increasing the tail-variance of the recovered pdf. In Fig. 20 we quantify the fit between the recovered pdf and the MC pdf using the two error metrics previously discussed in Eq. (41). We conclude that the reconstructed pdfs converge with $\approx$ 6–8 output modes.

7. Conclusions

We have presented a reduced-order probabilistic framework for parsimoniously modeling the non-Gaussian statistics of ship loads due to nonlinear interactions with irregular waves. The first ingredient of the framework is the formulation of random wave episodes of finite-duration and finite-dimensionality using the KL decomposition. These wave episodes are 'equipped' with a properly designed stochastic prelude that brings the ship to a statistical steady state, before it encounters the prescribed wave episode. A similar reduced-order representation is used for the VBM time series, i.e. the quantity of interest.

We employed these low-dimensional representations with the hydrodynamic and structural numerical code LAMP for ship dynamical responses. Using the resulted input–output vectors, we trained a GPR scheme, which acts as a surrogate model for the wave-ship interaction problem. We thoroughly discussed the choice of wave episodes parameters T and n and what is their role on the accurate recovery of the VBM pdfs even for strongly non-Gaussian regimes. A detailed set of numerical simulations confirms that the developed framework requires a fraction of the data and computational time compared with traditional Monte-Carlo approaches, without major sacrifices in the fidelity of the pdf tails (rare events). In addition, it allows us to trivially compute the desired statistics for different sea spectra, i.e. without the need to obtain new training data for a new surrogate model.

This approach is particularly well suited to physical tow tank experiments, where data constraints are high and the physical realizability of wave episodes is vital. However, the framework is applicable to other marine structure problems and more generally dynamical systems with rapid decorrelation, where the goal is the statistical response to stochastic excitations with few samples.

Another exciting application of this approach is its combination with active learning or optimal experimental design methods, which will require an even smaller set of training wave episodes or experiments. Additionally, ideas related to the use of models with variable fidelity fit naturally to the presented framework.

CRediT authorship contribution statement

Stephen Guth: Conceptualization, Methodology, Writing – original draft, Implementation of algorithms and direct simulations, Selection of test-cases for validation of trained models, Analysis and interpretation of the results, Editing of the manuscript. **Themistoklis P. Sapsis:** Conceptualization, Methodology, Writing – original draft, Supervision, Funding acquisition, Selection of test-cases for validation of trained models, Analysis and interpretation of the results, Editing of the manuscript.

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.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge support from the Office of Naval Research, USA under the program of Dr. Woei-Min Lin (Grant No. N00014-20-1-2366). We are also thankful to Dr. Vadim Belenky and Dr. Vladas Pipiras for several stimulating discussions, as well as to Mr. Kenneth Weems for his support with LAMP.

References

Álvarez, M.A., Rosasco, L., Lawrence, N.D., 2012. Kernels for vector-valued functions: A review. Found. Trends Mach. Learn. 4 (3), 195–266.

Anastopoulos, P.A., Spyrou, K.J., 2016. Ship dynamic stability assessment based on realistic wave group excitations. Ocean Eng. 120, 256–263.

Anastopoulos, P.A., Spyrou, K.J., 2019. Evaluation of the critical wave groups method in calculating the probability of ship capsize in beam seas. Ocean Eng. 187, 106213.

Anastopoulos, P.A., Spyrou, K.J., Bassler, C.C., Belenky, V., 2016. Towards an improved critical wave groups method for the probabilistic assessment of large ship motions in irregular seas. Probab. Eng. Mech. 44, 18–27, Special Issue Based on Papers Presented at the 7th International Conference on Computational Stochastic Mechanics (CSM7).

Belenky, V., Glotzer, D., Pipiras, V., Sapsis, T., 2016. On the tail of nonlinear roll motions. In: Proceedings of the 15th International Conference on Ship Stability Workshop.

Belenky, V., Glotzer, D., Pipiras, V., Sapsis, T.P., 2019. Distribution tail structure and extreme value analysis of constrained piecewise linear oscillators. Probab. Eng. Mech. 57, 1–13.

Belenky, V., Weems, K., Bassler, C., Dipper, M., Campbell, B., Spyrou, K., 2012. Approaches to rare events in stochastic dynamics of ships. Probab. Eng. Mech. 28, 30–38.

Belenky, V., Weems, K., Pipiras, V., Glotzer, D., Sapsis, T., 2018. Tail structure of roll and metric of capsizing in irregular waves. In: 32nd Symposium on Naval Hydrodynamics.

Belenky, V., Weems, K., Sapsis, T., Pipiras, V., 2021. Influence of deck submergence events on extreme properties of wave-induced vertical bending moment. In: Proceedings of the 1st International Conference on the Stability and Safety of Ships and Ocean Vehicles.

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, 564.

Boccotti, P., 1989. On mechanics of irregular gravity waves. Atti Accad. Naz. Lincei 19, 110–170.

Campbell, B., Belenky, V., Pipiras, V., 2016. Application of the envelope peaks over threshold (EPOT) method for probabilistic assessment of dynamic stability. Ocean Eng. 120, 298–304.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statist. Sci. 10 (3), 273–304.

Chasparis, F., Modarres-Sadeghi, Y., Hover, F.S., Triantafyllou, M.S., Tognarelli, M., Beynet, P., 2009. Lock-in, transient and chaotic response in riser VIV. In: Polar and Arctic Sciences and Technology; CFD and VIV of International Conference on Offshore Mechanics and Arctic Engineering, vol. 5, pp. 479–485.

Cousins, W., Sapsis, T.P., 2016. Reduced order precursors of rare events in unidirectional nonlinear water waves. J. Fluid Mech. 790, 368–388.

Farazmand, M., Sapsis, T.P., 2017. Reduced-order prediction of rogue waves in two-dimensional deep-water waves. J. Comput. Phys. 340, 418–434.

Fedele, F., Tayfun, M.A., 2009. On nonlinear wave groups and crest statistics. J. Fluid Mech. 620, 221–239.

Forrester, A., Sobester, A., Keane, A., 2008. Exploring and exploiting a surrogate. In: Engineering Design Via Surrogate Modelling. John Wiley & Sons, Ltd, pp. 77–107.

Forristall, G.Z., 2000. Wave crest distributions: Observations and second-order theory. Am. Meteorol. Soc. 30, 1931–1943.

Hasselmann, K., Barnett, T., Bouws, E., Carlson, H., Cartwright, D., Enke, K., Ewing, J., Gienapp, H., Hasselmann, D., Kruseman, P., Meerburg, A., Muller, P., Olbers, D., Richter, K., Sell, W., Walden, H., 1973. Measurements of wind-wave growth and swell decay during the joint North sea wave project (JONSWAP). Deut. Hydrogr. Z. 8, 1–95.

Huan, X., Marzouk, Y.M., 2013. Simulation-based optimal Bayesian experimental design for nonlinear systems. J. Comput. Phys. 232, 288–317.

Karhunen, K., 1947. Über lineare methoden in der wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fennicae Ser. A. I. Math.-Phys. 1947 (37), 79.

Khan, R.A., Ahmad, S., 2007. Dynamic response and fatigue reliability analysis of marine riser under random loads. In: Structures, Safety and Reliability; Petroleum Technology Symposium of International Conference on Offshore Mechanics and Arctic Engineering, vol. 2, pp. 183–191.

Kyul Joo, H., Mohamad, M.A., Sapsis, T.P., 2018. Heavy-tailed response of structural systems subjected to stochastic excitation containing extreme forcing events. J. Comput. Nonlinear Dyn. 13 (9), 090914.

Lin, W.-M., Collette, M., Lavis, D.R., Jessup, S.D., Kuhn, J., 2007a. Recent hydrodynamic tool development and validation for motions and slam loads on ocean-going high-speed vessels.

Lin, W.-M., Zhang, S., Weems, K.M., 2010. Numerical simulations of surface effect ship in waves. In: Proceedings of the 2010 Conference on Grand Challenges in Modeling & Simulation. GCMS '10, Society for Modeling & Simulation International, Vista, CA, pp. 414–421.

Lin, W.-M., Zhang, S., Weems, K., Jones, P., Meinhold, M., Metcalf, B., Powers, A.M., 2007b. Numerical simulation and validation study of wetdeck slamming on high speed catamaran. International Conference on Numerical Ship Hydrodynamics [9th].

Loève, M., 1948. Processus Stochastiques et Mouvement Brownien. Gauthier-Villars,

Longuet-Higgens, M.S., 1957. The statistical analysis of a random, moving surface. Phil. Trans. R. Soc. A 249 (966), 321–387.

Longuet-Higgins, M.S., 1975. On the joint distribution of the periods and amplitudes of sea waves. J. Geophys. Res. 80 (18), 2688–2694.

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.

Naess, A., Moan, T., 2013. Stochastic Dynamics of Marine Structures. Cambridge University Press.

Olsson, A., Sandberg, G., Dahlblom, O., 2003. On Latin hypercube sampling for structural reliability analysis. Struct. Saf. 25 (1), 47–68.

Phillips, O.M., Gu, D., Donelan, M., 1992. Expected structure of extreme waves in a Gaussian sea. Part I: Theory and SWADE buoy measurements. J. Phys. Oceanogr. 23, 992–1000.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. In: Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, USA, p. 248.

Rudy, S., Fan, D., Ferrandis, J.d.A., Sapsis, T., Triantafyllou, M.S., 2021. Learning optimal parametric hydrodynamic database for vortex-induced crossflow vibration prediction of both freely-mounted rigid and flexible cylinders. In: International Ocean and Polar Engineering Conference, vol. All Days, ISOPE-I-21-3214.

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

Sapsis, T.P., 2021. Statistics of extreme events in fluid flows and waves. Annu. Rev. Fluid Mech. 53 (1), 85–111.

Sapsis, T., Belenky, V., Weems, K., Pipiras, V., 2021. Extreme properties of impact-induced vertical bending moments. In: Proceedings of the 1st International Conference on the Stability and Safety of Ships and Ocean Vehicles.

Sapsis, T., Pipiras, V., Weems, K., Belenky, V., 2020. On extreme value properties of vertical bending moment. In: Proceedings of the 33rd Symposium on Naval Hydrodynamics Osaka, Japan, Virtual.

Sclavounos, P.D., 2012. Karhunen–Loève representation of stochastic ocean waves. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 468 (2145), 2574–2594.

Serebrinsky, S., Ortiz, M., 2005. A hysteretic cohesive-law model of fatigue-crack nucleation. Scr. Mater. 53 (1), 1193–1196.

Sharma, J.N., Dean, R.G., 1979. Development and evalutation of a procedure for simulating a random directional second order sea surface and associated wave forces. Oceon Eng. Rep.

Shin, Y.S., Belenky, V., Lin, W.M., Weems, K.M., Engle, A.H., McTaggart, K., Falzarano, J., Hutchison, B.L., Gerigk, M., Grochowalski, S., 2003. Nonlinear time domain simulation technology for seakeeping and wave-load analysis for modern ship design. Trans. Soc. Nav. Archit. Mar. Eng. 111, 557–583.

Stevens, K.W., 2018. Adaptive Sequential Sampling for Extreme Event Statistics in Ship Design (Master's thesis). Massachusetts Institute of Technology.

Tayfun, M.A., 1980. Narrow band nonlinear sea waves. J. Geophys. Res. 85 (C3), 1548–1552.

Vu, K.K., D'Ambrosio, C., Hamadi, Y., Liberti, L., 2017. Surrogate-based methods for black-box optimization. Int. Trans. Oper. Res. 24 (3), 393–424.

Weems, K., Belenky, V., Campbell, B., Pipiras, V., Sapsis, T., 2019. Envelope peaks over threshold (EPOT) application and verification. In: Proceedings of the 17th International Ship Stability Workshop.

Wiggins, S., 2003. Introduction to Applied Nonlinear Dynamical Systems and Chaos, vol. 4.