Developing closed-form equations of maximum drag and moment on rigid vegetation stems in fully nonlinear waves

Ling Zhu[1] and Qin Chen[2]

[1]Senior Research Scientist, Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA. Email: l.zhu@northeastern.edu

[2]Professor, Civil and Environmental Engineering, Marine and Environmental Sciences, Northeastern University, Boston, MA 02115, USA. Email: q.chen@northeastern.edu

ABSTRACT

Coastal wetlands act as natural buffers against wave energy and storm surges. In the course of energy dissipation, vegetation stems are exposed to wave action, which may lead to stem breakage. An integral component of wave attenuation modeling involves quantifying the extent of damaged vegetation, which relies on determining the maximum drag force ($F_{D\,\max}$) and maximum moment of drag ($M_{D\,\max}$) experienced by vegetation stems. Existing closed-form theoretical equations for $M_{D\,\max}$ and $F_{D\,\max}$ are only valid for linear and weakly nonlinear deep water waves. To address this limitation, this study first establishes an extensive synthetic dataset encompassing 256,450 wave and vegetation scenarios. Their corresponding wave crests, wave troughs, $M_{D\,\max}$, and $F_{D\,\max}$, which compose the dataset, are numerically computed through an efficient algorithm capable of fast computing fully nonlinear surface gravity waves in arbitrary depth. Seven dominant wave and vegetation related dimensionless parameters that impact $M_{D\,\max}$ and $F_{D\,\max}$ are discerned and incorporated as input feature parameters into an innovative sparse regression algorithm to reveal the underlying nonlinear relationships between $M_{D\,\max}$, $F_{D\,\max}$ and the input features. Sparse regression is a subfield of machine learning that primarily focuses on identifying a subset of relevant feature functions from a feature function library. Leveraging this synthetic dataset and the power of

sparse regression, concise yet accurate closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$ are developed. The discovered equations exhibit good accuracy compared with the ground truth in the synthetic dataset, with a maximum relative error below 6.6% and a mean relative error below 1.4%. Practical applications of these equations involve assessment of the extent of damaged vegetation under wave impact and estimation of $M_{D\,\max}$ and $F_{D\,\max}$ on cylindrical structures.

Keywords: sparse regression, equation discovery, maximum drag/bending moment, vegetation stem breakage, nonlinear wave theory

INTRODUCTION

Coastal salt marshes play a crucial role in dissipating wave energy, stabilizing sediment along the shoreline, mitigating the impacts of coastal flooding, and providing vital habitats for plants and animal species. Designing and implementing nature-based solutions for the protection and restoration of coastal salt marshes are imperative to the preservation of coastal ecosystems and the advancement of sustainable shoreline management practices. Coastal salt marshes are subjected to drag force and bending moment induced by waves. Determining the maximum wave-induced drag force ($F_{D\,\max}$) on salt marsh stems and maximum moment of drag force about the stem base ($M_{D\,\max}$) is crucial to evaluate the susceptibility of salt marshes to stem breakage and to quantify the effectiveness of salt marshes in dissipating wave energy in a high energy wave environment.

Closed-form analytical approximations of $F_{D\,\max}$ or $M_{D\,\max}$ are desired in engineering application. Salt marshes are commonly treated as rigid cylinders. Researchers developed simple closed-form analytical equations for $F_{D\,\max}$ or $M_{D\,\max}$ based on linear wave theory (LWT) (e.g., Dalrymple et al. 1984; Vuik et al. 2018) and Stokes 2nd order wave theory (STK2) (Zhu and Chen 2019). However, these wave theories, even higher order Stokes wave theories (e.g., Stoke's 5th-order wave theory by Fenton 1985), have limitations when applied to highly nonlinear waves in intermediate and shallow water regions, rendering the established closed-form analytical equations or lookup tables unsuitable for salt marshes in these regions. Few attempts have been made to calculate $F_{D\,\max}$ or $M_{D\,\max}$ from wave theories applicable to shallow water waves.

Fenton (1988) proposed an algorithm for calculating steady surface waves in deep water or water of finite depth based on stream function wave theory (SFWT). However, numerical approximations from Fenton's algorithm do not converge or converges to ghost solutions with spurious oscillations for waves with

$kh < \pi/15$ (Clamond and Dutykh 2018), where kh is the dimensionless wavelength parameter that measures the wave dispersion. Recently, Clamond and Dutykh (2018) proposed an efficient algorithm, with time complexity $O(N \log N)$ (N is the number of Fourier modes), for fast computation of steady surface gravity waves in arbitrary depth (i.e., Stokes, cnoidal, and solitary waves) with wave steepness up to approximately 99% of the maximum steepness for all wavelengths. Clamond and Dutykh's algorithm, denoted as SSGW (steady surface gravity waves), numerically solves the modified Babenko equation (Babenko 1987) via the classical Petviashvili method (Petviashvili 1976). Neither Fenton's algorithm nor SSGW provides explicit equations of horizontal velocity (u) and surface elevation (η), and thus, it is not possible to theoretically derive closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$ from these two algorithms.

Instead of deriving theoretical equations, an alternative approach is to formulate semi-theoretical equations by leveraging machine learning techniques. Recent advances in machine learning, bolstered by the increasing computational capabilities, facilitate the development of data-driven models capable of harnessing extensive data to make predictions based on input features. Among them, neural network models give only implicit relationship between input and output variables. In practical applications, engineers experienced with the utilization of empirical formulas may prefer an explicit calculation method. Compared to neural network models, equation discovery techniques provide explicit and interpretable mathematical formulas to describe the underlying dependencies between variables in a dataset. Equation discovery techniques have been used to uncover governing partial differential equations of nonlinear dynamical systems from noisy observation data (e.g., Wang et al. 2011; Raissi et al. 2018; Chen et al. 2021; Wang et al. 2021). Two popular methods for equation discovery are the genetic algorithm (e.g., Bongard and Lipson 2007; Schmidt and Lipson 2009; Pourzangbar 2012; Bonakdar et al. 2015; Pourzangbar et al. 2017a; Pourzangbar et al. 2017b; Formentin and Zanuttigh 2019; Lee and Suh 2019; Udrescu and Tegmark 2020; Dalinghaus et al. 2023) and the sparse regression (e.g., Brunton et al. 2016; Lee et al. 2022). The genetic algorithm is expressive and versatile but does not scale well to large systems and may be prone to overfitting (Brunton et al. 2016). Sparse regression is considered more efficient or manageable. Brunton et al. (2016) formulated system identification as sparse regression problems and developed an innovative framework, SINDy (Sparse Identification of Nonlinear Dynamics). SINDy leverages sparsity-promoting techniques to find out the fewest active terms from a space of nonlinear candidate functions to accurately represent the data. The resulting parsimonious models balance accuracy with model complexity while avoiding overfitting the model to data.

SINDy has been successfully implemented to identify nonlinear dynamical systems in various domains such as fluid dynamics (Loiseau and Brunton 2018), structural engineering (Li et al. 2019), and chemical systems (Hoffmann et al. 2019). To the best of the authors' knowledge, sparse regression has not been previously applied in the field of coastal and ocean engineering.

The objectives of this study are to: (1) create an extensive synthetic dataset encompassing a wide range of wave conditions, vegetation submergence, and the corresponding wave crests ($\eta_{\max}$), wave troughs ($\eta_{\min}$), $M_{D\,\max}$, and $F_{D\,\max}$ based on steady surface waves calculated by using the SSGW algorithm; (2) leverage the SINDy framework to formulate closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$ based on the synthetic dataset; and (3) apply the discovered equations to assess the extent of damaged vegetation under wave impact and estimate $M_{D\,\max}$ and $F_{D\,\max}$ on cylindrical structures. The created dataset includes 256,450 combinations of wave height, wave period, water depth, vegetation submergence (vegetation height to water depth ratio), and numerical approximations of $M_{D\,\max}$ and $F_{D\,\max}$. The core processes of equation discovery involve identifying input feature parameters and utilizing the sparse regression algorithm to achieve a balance between the sparsity and accuracy of the equations. Given the established exact theoretical equations for $M_{D\,\max}$ and $F_{D\,\max}$ from LWT and STK2, along with those for η and u from SFWT, we discern the dominant wave and vegetation related dimensionless parameters that impact $M_{D\,\max}$ and $F_{D\,\max}$, and incorporate them as input features into the SINDy framework to reveal the underlying nonlinear relationships between $M_{D\,\max}$, $F_{D\,\max}$ and input features.

This paper is structured as follows: in the data and methods section, we outline the procedure for creating the synthetic dataset, introduce the established theoretical equations for $M_{D\,\max}$ and $F_{D\,\max}$ from LWT and STK2, and identify the input features. Additionally, we briefly introduce the SINDy framework, focusing on the feature library and hyper-parameter. In the results section, we demonstrate SINDy's ability to recover theoretical LWT-based equations for $M_{D\,\max}$ and $F_{D\,\max}$ for linear waves, and present the discovered equations for fully nonlinear waves from shallow to deep waters, along with discussions on generalization and accuracy. Two case studies are illustrated in the application section. Concluding remarks and future research topics are presented in the final section.

DATA AND METHODS

Building a synthetic dataset

The drag force dominates over inertia force because $b_v \ll H$ (Journée and Massie 2001). The drag force per unit length of the vegetation stem f_D can be approximated as a quadratic function of the time-dependent vertically-varying horizontal velocities $u(t, z)$ (Morison et al. 1950) as $f_D(t, z) = \frac{1}{2}\rho C_D b_v u|u|$, where t is the time, z is the vertical axis, ρ is the water density, C_D is the drag coefficient, and b_v is the stem width. The total drag force acting on the vegetation stem $F_D(t)$ and the moment of drag force about the sea floor $M_D(t)$ are determined as:

$$F_D(t) = \int_{-h}^{\min(\eta,-h+h_v)} f_D(t, z)dz \quad \text{and} \quad M_D(t) = \int_{-h}^{\min(\eta,-h+h_v)} f_D(t, z) \cdot (z + h)dz \tag{1}$$

where h is the water depth, η is the surface elevation, and h_v is the vegetation stem height. The integrations are from the sea floor ($z = -h$) to either the vegetation stem top ($z = -h + h_v$) for submerged vegetation or the free surface η for emergent vegetation (see Fig. 1a). The maximum $F_D(t)$ and $M_D(t)$ within a wave cycle, denoted as $F_{D\,\max}$ and $M_{D\,\max}$, occur at the passage of the wave crest $\eta_{\max}$, and are computed as:

$$F_{D\,\max} = \int_{-h}^{\min(\eta_{\max},-h+h_v)} \frac{1}{2}\rho C_D b_v u_{\max}|u_{\max}|dz \tag{2}$$

$$M_{D\,\max} = \int_{-h}^{\min(\eta_{\max},-h+h_v)} \frac{1}{2}\rho C_D b_v u_{\max}|u_{\max}| \cdot (z + h)dz \tag{3}$$

where $u_{\max}(z)$ is the maximum horizontal velocity within the wave cycle. Given that C_D and b_v are usually considered constants along the stem in numerical models (e.g., Mendez and Losada 2004; Anderson and Smith 2014; Luhar and Nepf 2016; Zhu et al. 2023) and are factored out of the integrations, we omit $\frac{1}{2}\rho C_D b_v$ from the dataset but can reintroduce it during applications.

The first step towards generating a dataset with general applicability to non-breaking waves and vegetation in natural environments is to determine an adequate quantity of representative, impartial scenarios of waves and vegetation. Waves are characterized with dimensionless parameters. Following the wave classification presented in LeMehaute (1976), wave characteristics are determined based on conditions as below:

(i) $-4 \leq \log_{10}(h/gT^2) \leq -0.1$,

(ii) $-5 \leq \log_{10}(H/gT^2) \leq -1$,

(iii) $H/L < 0.14 \tanh kh$ (Kamphuis 1991) to meet the non-breaking wave criteria,

(iv) $kh \leq 2\pi$ because coastal salt marshes rarely experience very deep water waves,

(v) $kH/2$ is no greater than the highest computable $kH/2$ by SSGW (see Fig. 2).

where g is the gravitational acceleration, T is the wave period, H is the wave height, L is the wavelength, and $k = 2\pi/L$ is the wavenumber. The ranges of $\log_{10}(h/gT^2)$ and $\log_{10}(H/gT^2)$ are discretized with increments of 0.2, and 0.17, respectively, to balance the diversity of wave conditions in the dataset and the computational cost. In the dataset, we let $0.5 \leq h \leq 8.0$ m. However, later in the generalization of the discovered equations, we will demonstrate that the discovered equations are applicable to water depth beyond this scope. When the values of h, $\log_{10}(h/gT^2)$ and $\log_{10}(H/gT^2)$ are provided, T and H can be easily computed, and L and k are pragmatically determined from the dispersion relationship based on LWT: $\omega^2 = gk \tanh(kh)$, where $\omega = 2\pi/T$ is the wave angular frequency. This dispersion relationship is limited to LWT and STK2; however, it provides a more pragmatic option for users of the equations developed in this study. Therefore, we use k from linear dispersion relationship to develop the dataset and discover equations. The maximum H/h among all wave scenarios is 0.837, close to the breaking index of $(H/h)_{\max} = 0.826$ proposed in Longuet-Higgins (1974). All wave scenarios in the dataset are depicted in Fig. 3. The x-axis represents the wave dispersion and the y-axis represents the wave nonlinearity.

The computations of $M_{D\max}$ and $F_{D\max}$ by definition involve integrations from the sea floor ($z = -h$) to the uppermost wetted segment along the vegetation stem (Eqs. 2 - 3). The vegetation stem should exceed $h+\eta_{\max}$ to be fully emergent. Fig. 3 presents the variations of $\eta_{\max}/H$, computed through SSGW. The $\eta_{\max}/H$ ratios distinct across different wave categories, with approximate values of 0.5 for linear waves, up to 0.7 for high-order Stokes waves, and up to 0.9859 for cnoidal waves. Given that $\left(\frac{h_v}{h}\right)_{\max} \leq \left(\frac{\eta_{\max}}{H}\right)_{\max} \cdot \left(\frac{H}{h}\right)_{\max}$, we encompass a comprehensive spectrum of submergence scenarios h_v/h from 0 to 1.85 in the dataset, and discretize this range with an increment of 0.05. The novel synthetic dataset encompasses a total of 256,450 combinations of wave conditions and vegetation submergence.

For each combination of wave conditions (h, H, and T) and vegetation submergence (h_v/h) in the dataset, we apply SSGW to get numerical approximations of $\eta_{\max}$, $\eta_{\min}$, and $u_{\max}(z)$. As the wave steepness or the wave wavelength increases, SSGW requires a rapidly increasing number of Fourier modes (N) to achieve spectral accuracy. For instance, for an extreme wave with $kh = 0.0885, H/h = 0.82$, the required N for full spectral accuracy is 2^{17} (Clamond and Dutykh 2018). In this study, we adopted $N = 2^{17}$ and numerical tolerance of 10^{-10} for all waves in the dataset. The time complexity of SSGW algorithm is $O(N \log N)$. Given the swift vertical variations in $u_{\max}(z)$ near the free surface of highly nonlinear waves, a 50-point Gauss-Legendre quadrature is employed for the numerical integrations in the calculations of $M_{D\,\max}$ and $F_{D\,\max}$ (Eqs. 2 and 3). When necessary, the amount of Gauss nodes is reduced to ensure that the free surface (the first Gauss node) maintains an adequate separation from the next Gauss node (Clamond and Dutykh 2018). The $M_{D\,\max}$ and $F_{D\,\max}$ from SSGW, denoted as $M_{D\,\max,\mathrm{SSGW}}$ and $F_{D\,\max,\mathrm{SSGW}}$, serve as the ground truth for assessing the accuracy of the discovered closed-form equations.

To validate the implementation of SSGW, we compare the total horizontal force (F_T) calculated by SSGW against laboratory measurements obtained from Luhar and Nepf (2016). Their experiment with 5 cm long HDPE (high density polyethylene plastic) blades, $T = 2$ s, $H = 7.8$ cm, and $h = 0.3$ m is carefully selected for validation because the mimic vegetation in this particular experiment is essentially rigid (see Fig. 6 in Luhar and Nepf 2016). The mimic vegetation width b_v is 0.02 m and thickness is 0.4 mm. The total horizontal force is the sum of drag force and inertia force (see Eq. 22 in Appendix I). The inertia coefficient $C_M = 1.0$ is chosen following Luhar and Nepf (2016), whereas the drag coefficient $C_D = 3.6$ is determined from the empirical relationship $C_D = \max(10KC^{-1/3}, 1.95)$ in Luhar and Nepf (2016), where $KC = U_w T/b_v$ is the Keulegan-Carpenter number (U_w is the orbital wave velocity).

Discovering closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$ using SINDy

Existing theoretical closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$

Following Dalrymple et al. (1984), we can insert $u_{\max} = \frac{\omega H}{2\sinh kh} \cdot \cosh k(z+h)$ from LWT into Eqs. (2) and (3), and perform integrations from the sea floor to $\min(0, -h+h_v)$ rather than to $\min(\eta_{\max}, -h+h_v)$ to get theoretical closed-form equations as below:

$$M_{D\,\max,\mathrm{LWT}} = \frac{\omega^2 H^2}{32k^2 \sinh^2 kh}\left[2kh_v \sinh 2kh_v - \cosh 2kh_v + 1 + 2(kh_v)^2\right] \quad (4)$$

$$F_{D\,\max,\mathrm{LWT}} = \frac{\omega^2 H^2}{16k \sinh^2 kh}\left[2kh_v + \sinh 2kh_v\right] \quad (5)$$

where $\omega = 2\pi/T$ is the wave angular frequency. h_v in Eqs. (4) and (5) should be $\min(h_v, h)$. The assumptions made in the derivation of Eqs. (4)-(5) are: (i) waves can be described by LWT, and (ii) wave heights are so small that $\eta_{\max}$ can be replaced with the still water level ($z = 0$ m) in the vertical integrations. Fig. 1b presents

$u_{\max}$ of a nonlinear wave in shallow water (fall in the cnoidal wave region) computed by LWT, Fenton's algorithm, and SSGW. For nonlinear waves in shallow water, the wave shape is distorted (peaky crest and flat trough). Fig. 1b clearly indicates that LWT is unsuitable for cnoidal waves, and $u_{\max}$ above the still water level makes a substantial contribution to $M_{D\max}$ and $F_{D\max}$. Fig. 4a-b illustrates $M_{D\max,\mathrm{SSGW}}/M_{D\max,\mathrm{LWT}}$ and $F_{D\max,\mathrm{SSGW}}/F_{D\max,\mathrm{LWT}}$ for an emergent vegetation. It clearly shows that the LWT-based theoretical equations are not applicable to vegetation in shallow water waves. The moment ratio can be as much as 7 in shallow water waves with large relative wave height (H/h).

Zhu et al. (2019) inserted $u = \frac{\omega H}{2\sinh kh}\cosh k(z+h)\cos\theta + \frac{3\omega H}{16\sinh^4 kh} kH \cosh 2k(z+h)\cos 2\theta$ from STK2 into Eq. (1) and proposed a STK2-based closed-form equations for $M_{D\max}$:

$$M_{D\max,\mathrm{STK2}} = \frac{\omega^2 H^2}{32k^2\sinh^2 kh}\left[M_1 + \frac{kH}{\sinh^3 kh}M_2 + \frac{(kH)^2}{\sinh^6 kh}M_3\right] \tag{6}$$

where

$$\begin{aligned} M_1 &= 2kh_v \sinh 2kh_v - \cosh 2kh_v + 1 + 2(kh_v)^2 \\ M_2 &= 5 - 3\cosh kh_v - 2\cosh^3 kh_v + 3kh_v \sinh kh_v + 6kh_v \sinh kh_v \cosh^2 kh_v \\ M_3 &= 4kh_v \sinh 4kh_v - \cosh 4kh_v + 1 + 8(kh_v)^2 \end{aligned} \tag{7}$$

STK2 is generally applicable to waves with Ursell number (Ur, defined as $\mathrm{Ur} = \frac{HL^2}{h^3}$) less than 26 (Isobe et al. 1982), where the dimensionless wave parameter Ur indicates the wave nonlinearity with respect to wave dispersion. For shallow water waves with finite amplitude (Ur > 26), STK2 introduces a secondary crest at the wave trough. We should seek nonlinear wave theory and numerical algorithms, such as SSGW, for computing free surface waves and $M_{D\max}$ and $F_{D\max}$ in arbitrary water depth.

Identification of major dimensionless parameters controlling $M_{D\max}$ and $F_{D\max}$

A wave can be characterized by dimensionless parameters that represent wave dispersion (kh) and wave nonlinearity. In deep water and water with finite depth, a wave is usually characterized by kh (wave dispersion) and kH (wave nonlinearity), while in shallow water, the relative wave height H/h is used in lieu of kH. Beji (1995) proposed a nonlinearity parameter $ka/\tanh(kh)$ (a is the wave amplitude) that is valid for both deep and shallow water waves. Nevertheless, these dimensionless parameters are not sufficient for formulating $M_{D\,\max}$ and $F_{D\,\max}$ in scenarios involving general wave conditions.

Fenton's algorithm, based on SFWT, and SSGW exhibit a close alignment, with an approximate agreement of up to six decimal places for deep water waves (Clamond and Dutykh 2018), and small discrepancies for finite amplitude shallow water waves (see Fig. 4c-d). Therefore, the theoretical equations for $\eta_{\max}$ and $u_{\max}$ from SFWT as below (Fenton 1988) can offer valuable insights into the selection of input features:

$$\eta_{\max,\mathrm{SFWT}} = \sum_{j=1}^{N} A_j \quad \text{and} \quad u_{\max,\mathrm{SFWT}} = \sum_{j=1}^{N} jB_j \frac{\cosh jk(z+h)}{\cosh jkh} \tag{8}$$

where $\theta = \omega t + kx$ is the phase angle, and N represents the number of Fourier modes. Fenton's algorithm solves for the coefficients A_j and B_j ($j = 1, \ldots, N$) from a system of $2N + 10$ nonlinear equations including free surface boundary condition, kinematic boundary condition, and stream function equation. It is important to note that the equations for $u_{\max,\mathrm{SFWT}}$ in Eq. (8) only holds for scenarios where the Eulerian time-mean velocity C_E is equal to 0. Inserting Eq. (8) into Eqs. (2) and (3) yields

$$M_{D\,\max,\mathrm{SFWT}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{ijB_iB_j}{\cosh ikh \cosh jkh} \int_{-h}^{\min(\eta_{\max},-h+h_v)} (z+h)\cosh ik(z+h)\cosh jk(z+h)dz \tag{9}$$

$$F_{D\,\max,\mathrm{SFWT}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{ijB_iB_j}{\cosh ikh \cosh jkh} \int_{-h}^{\min(\eta_{\max},-h+h_v)} \cosh ik(z+h)\cosh jk(z+h)dz \tag{10}$$

Using hyperbolic trigonometry identities, the integral in Eq. (9) can be explicitly expressed as the product of $1/k^2$ and a polynomial function of

$$\begin{cases} \sinh(i \pm j)kh_v,\ \cosh(i \pm j)kh_v,\ \text{and } kh_v & \text{if } h_v < h + \eta_{\max} \\ \sinh(i \pm j)k(h+\eta_{\max}),\ \cosh(i \pm j)k(h+\eta_{\max}),\ kh,\ \text{and } k\eta_{\max} & \text{if } h_v \geq h + \eta_{\max}, \end{cases} \tag{11}$$

and the integral in Eq. (10) can be explicitly expressed as the product of $1/k^2$ and a polynomial function of

$$\begin{cases} \sinh(i \pm j)kh_v,\ \text{and } kh_v & \text{if } h_v < h + \eta_{\max} \\ \sinh(i \pm j)k(h+\eta_{\max}),\ kh,\ \text{and } k\eta_{\max} & \text{if } h_v \geq h + \eta_{\max}. \end{cases} \tag{12}$$

Note that "$i \pm j$" inside the hyperbolic functions in (11) and (12) can be further dropped by using hyperbolic trigonometry identities. The coefficients $\frac{ijB_iB_j}{\cosh ikh \cosh jkh}$ in Eqs. (9) and (10) do not have explicit forms;

nonetheless, guided by our knowledge on the theoretical equations for $M_{D\,\max,\mathrm{STK2}}$ (Eq. 6), it can be inferred that these coefficients are dependent on $\frac{\omega^2 H^2}{\sinh^2 kh}$ and $\frac{kH}{\sinh^3 kh}$.

A conjecture regarding the closed-form equations for $M_{D\,\max}$ and $F_{D\,\max}$ would be:

$$M_{D\,\max} = \frac{\omega^2 H^2}{32k^2 \sinh^2 kh}\Bigg[\sum\Big(\frac{kH}{\sinh^3 kh}\Big)^p \cdot f_1\big(k\eta_{\max},\ kh,\ kh_v,\ \sinh k(h+\eta_{\max}),\ \cosh k(h+\eta_{\max}),\ \sinh kh_v,\ \cosh kh_v\big)\Bigg] \quad (13)$$

$$F_{D\,\max} = \frac{\omega^2 H^2}{16k \sinh^2 kh}\Bigg[\sum\Big(\frac{kH}{\sinh^3 kh}\Big)^q \cdot f_2\big(k\eta_{\max},\ kh,\ kh_v,\ \sinh k(h+\eta_{\max}),\ \cosh k(h+\eta_{\max}),\ \sinh kh_v,\ \cosh kh_v\big)\Bigg] \quad (14)$$

where p and q are integers, and f_1 and f_2 represent polynomial functions. This conjecture is based on the equations for $\eta_{\max}$ and $u_{\max}$ from SFWT, which has discrepancies with SSGW's solutions for very steep waves. We will discuss the feasibility of input features identified by this method, and the potential errors in the results section.

The dimensionless parameters in Eqs. (13) and (14) are potential feature parameters to be used in SINDy. However, It is not feasible to include $\eta_{\max}$ in the feature parameters because it is unknown *a priori*. From Fig. 3, it is observed that waves in deep and intermediate water depth with $\mathrm{Ur} \leq 40$ exhibit a slight skewness with $\eta_{\max} = 0.5H \sim 0.7H$, whereas shallow water waves with $\mathrm{Ur} > 40$ are strongly skewed with $\eta_{\max} = 0.7H \sim 0.9859H$. Hence, we divide our synthetic dataset into four distinct subsets following the criteria in Table 1, and use the mean of $\eta_{\max}$ of the subset in the feature parameters. Moreover, given that $\eta_{\max}$ is no greater than $0.7H$ for waves with $\mathrm{Ur} \leq 40$, the height of wetted stem should be no greater than $h + 0.7H$ in subset II. We enforce $kh_v \approx \min(kh_v, kh + 0.7kH)$ in subset II. Similarly, we enforce $kh_v \approx \min(kh_v, kh + 0.9859kH)$ in subset IV.

Eventually, the feature parameters to be used in SINDy for general wave conditions are identified as:

$$x_1 = \frac{kH}{\sinh^3 kh},\ x_2 = k(h+\gamma H),\ x_3 = kh, x_4 = \sinh k(h+\gamma H),$$
$$x_5 = \cosh k(h+\gamma H),\ x_6 = \sinh kh_v,\ x_7 = \cosh kh_v \quad (15)$$

and denoted as $\mathbf{X} = [x_1, x_2, \ldots, x_7]$, in which, $\gamma = 0.53$ and 0.86 for Stokes and cnoidal waves, respectively. Among these feature parameters, x_1 approaches Ur in shallow water, kH is a measure of wave steepness, x_3 is a measure of wave dispersion, and kh_v is the vegetation height to wavelength ratio. In our synthetic dataset, kh varies from 0.06 to 4.7, and wave steepness kH ranges from 4×10^{-4} to 0.85. To achieve better regression results, $M_{D\,\max}$ and $F_{D\,\max}$ are normalized as

$$M^*_{D\,\max} = M_{D\,\max} \Big/ \left[\frac{\omega^2 H^2}{32k^2 \sinh^2 kh \sinh^2 kh_v} \right], \tag{16}$$

$$F^*_{D\,\max} = F_{D\,\max} \Big/ \left[\frac{\omega^2 H^2}{16k \sinh^2 kh \sinh kh_v} \right]. \tag{17}$$

Equation discovery with SINDy

$\mathbf{X}^p$ (p is an integer) denotes the p-th order polynomial. For instance, when $p = 2$,

$$\mathbf{X}^2 = \left[x_1^2,\ x_1x_2,\ \cdots,\ x_1x_7,\ x_2^2,\ x_2x_3,\ \cdots,\ x_2x_7,\ x_3^2,\ \cdots,\ x_7^2\right]$$

With a polynomial order K, the feature parameters $\mathbf{X}$ can yield a combination of polynomial terms $\Theta(\mathbf{X}) = \left[1, \mathbf{X}, \mathbf{X}^2, \cdots, \mathbf{X}^K\right]$, referred to as the feature library. The feature library constitutes a space of polynomial functions, with each function serving as a candidate term in the equation to be discovered. In this study, we aim to express the target $\mathbf{y}$ (i.e., $M^*_{D\,\max}$ and $F^*_{D\,\max}$) as a sparse linear combination of terms in $\Theta(\mathbf{X})$:

$$\mathbf{y} = \Theta(\mathbf{X})\xi \tag{18}$$

$\Theta(\mathbf{X})$ is a $n \times m$ matrix, where n is the number of training scenarios in the synthetic dataset, and m is the number of polynomial functions in the feature library. The target $\mathbf{y}$ is a vector of dimension n, and coefficients ξ is a vector of dimension m. The $n \gg m$ in the training data, making Eq. (18) an overdetermined system. The equation sparsity is measured by the count of polynomial terms present in the discovered equation, which is equivalent to the count of non-zero entries in ξ, denoted as $\|\xi\|_0$.

The standard regression method, specifically the least squares regression, yields a solution for ξ with many non-zero entries, indicating that many candidate functions in the feature library contribute to $\mathbf{y}$. Contrastingly, SINDy employs a sequential thresholded least-squares (STLS) algorithm to recursively determine the sparse

ξ with a cutoff value λ (see Algorithm 1 in Appendix IV). This algorithm is computationally efficient, rapidly converges to a sparse solution for ξ in a few iterations. If coefficient ξ_i of the i-th candidate function in the feature library is zero, it means that candidate function does not contribute to $\mathbf{y}$. This algorithm is robust with only one tuning parameter λ. To discover the equations with a balance of optimal complexity and accuracy, we employed the Pareto front analysis (Smits and Kotanchek 2005), which represents a set of solutions that achieve the best trade-offs between model accuracy and sparsity, enabling the selection of an optimal model based on the desired level of simplicity and predictive performance.

To exemplify the applicability of SINDy, the algorithm is initially employed to recover Eqs. (4) and (5) for small amplitude waves in deep water. The feature parameters $\mathbf{X}$ in Eq. (15) is not feasible because Eqs. (4) and (5) are obtained by integrating from sea floor to $\min(0, -h + h_v)$ rather than to $\min(\eta_{\max}, -h + h_v)$, thus $\eta_{\max}$ should not be included in the feature parameters. We select $\mathbf{X} = [kh_v,\ \sinh 2kh_v,\ \cosh 2kh_v]$ and construct a feature library $\Theta(\mathbf{X})$ that includes quadratic nonlinearity of input features, i.e., $K = 2$. Scenarios with $M_{D\,\max,\mathrm{SSGW}}$ close to $M_{D\,\max,\mathrm{LWT}}$ are selected from the synthetic dataset for recovering Eqs. (4) and (5). Fig. 5 schematizes how sparse coefficients are identified in a space of polynomial functions.

For the equation discovery of general wave conditions, the feature parameters in Eq. (15) is used. We reserve scenarios with $0.5 \leq h \leq 6.75$ m in the synthetic dataset, totaling 215,280 scenarios, for equation discovery and testing purposes. Specifically, 70% of these scenarios are randomly selected as training data for equation discovery, while 30% are set aside for testing. A grid search for λ from 0 to 10 with the increment of 0.0025 is performed to obtain the optimal λ that balances equation sparsity and accuracy. Moreover, we reserve scenarios with $7 \leq h \leq 8$ m in the synthetic dataset, totaling 41,170 scenarios, for further testing the model's generalization.

The accuracy of the discovered equations from SINDy is evaluated using two metrics: the coefficient of determination R^2 and the relative error ϵ, defined as follows:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \quad \text{and} \quad \epsilon = \frac{|T - P|}{T} \tag{19}$$

where the sum of squares of residual RSS is computed as $\mathrm{RSS} = \sum_{i=1}^{n} (T_i - P_i)^2$. The total sum of squares TSS is computed as $\mathrm{TSS} = \sum_{i=1}^{n} \left(T_i - \overline{T}\right)^2$. P and T are the predictions and true values, respectively. $R^2 = 1$ indicates perfect agreement between the predictions and true values. $R^2 = 0$ indicates that the model results are as good as random guesses around the mean of the true values.

RESULTS

Validation of SSGW implementation

Fig. 6 shows the temporal variations of the horizontal force. The predominance of drag force over inertia force is attributed to $b_v \ll H$ (Journée and Massie 2001). Thus, the total horizontal force (F_T) is in phase with the drag force and $F_{T\,\max} \approx F_{D\,\max}$ in this experimental case. The numerically computed F_T, using u and η from SSGW, compares reasonably well with laboratory measurements throughout the wave cycle. This validation affirms that F_T from SSGW, together with values of C_D and C_M based on empirical formulas and previous literature, provide a reliable estimation of the horizontal force acting on the mimic vegetation. The modeled and measured F_T have R^2 of 0.92 and normalized root-mean-square error of 11% (normalized by the difference between the maximum and minimum measurements). The measured $F_{T\,\max}$ in the wave cycle is 0.084 N, whereas the computed $F_{T\,\max}$ is 0.093 N. There are around 0.0165 ~ 0.0255 N discrepancies between the computed and measured F_T at the trough around 0.5 ~ 1.7 s. Such discrepancies may stem from factors including (1) higher swaying velocities of the mimic vegetation at the trough, (2) measurement uncertainty, such as the 10% accuracy of the load cell, and (3) errors introduced by the Morison equation, which is fundamentally a parametric formula.

Recovering theoretical equations in Eqs. (4) and (5) for linear waves

Fig. 4a-b shows contour lines of 30% discrepancies between $M_{D\,\max,\mathrm{SSGW}}$ in the synthetic dataset and $M_{D\,\max,\mathrm{LWT}}$ from the LWT-based theoretical equations. For waves with $M_{D\,\max,\mathrm{SSGW}}/M_{D\,\max,\mathrm{LWT}} \leq 30\%$, the application of SINDy with $\lambda = 0.5$ yields $M_{D\,\max}$ and $F_{D\,\max}$ equations closely resembling those in Eqs. (4) and (5). The relevant feature library and the corresponding coefficients can be found in Table 2. We can also apply the least squares regression (LSQR) method (Barrett et al. 1994) to solve the linear system $\mathbf{y} = \Theta(\mathbf{X})\xi$ for the coefficients ξ. With tolerance of 10^{-15}, the LSQR method leads to a set of dense ξ that does not align with the theoretical solutions (see Table 2). SINDy aims to identify a subset of relevant feature functions (i.e., the feature functions whose coefficients are not zero) from the feature library, whereas the LSQR method attempts to find coefficients for all feature functions to best fit the data and does not perform feature selection. Therefore, SINDy identifies that $F_{D\,\max}$ is a function of 'kh_v' and '$\sinh 2kh_v$' and filters out irrelevant feature functions, whereas the LSQR method keeps all feature functions in the library.

Discovering closed-form equations for nonlinear waves

In the case of general wave conditions of various nonlinearity up to 99% of the wave breaking threshold, the feature parameters in Eq. (15) are employed in the SINDy framework to discover equations for both $M^*_{D\,\max,\mathrm{SFWT}}$ and $F^*_{D\,\max,\mathrm{SFWT}}$ within each subset of the synthetic dataset. A higher polynomial order K in the feature library does not necessary lead to greater accuracy, as it may render $\Theta(\mathbf{X})$ more likely to be ill-conditioned for least squares regression. Here, the feature library with up to 4th-order polynomials (i.e., $K = 4$) is found to provide the best-fit ξ. The library $\Theta(\mathbf{X})$ consists of a total of 330 candidate polynomial functions (see *Supplemental Material* for the complete list). The hyper-parameter λ can be fine-tuned to achieve a balance between equation accuracy, quantified by max ϵ, and equation sparsity, indicated by $\|\xi\|_0$. A smaller λ typically results in more terms in the equation and higher accuracy, while a larger λ tends to yield fewer terms in the equation and lower accuracy. Fig. 7 shows the variations of equation accuracy with the equation sparsity for each subset. Table 3 summarizes the optimal values of the hyper-parameter λ, ensuring that max ϵ remains below 7%, and the count of terms in the discovered equations. The discovered equations exhibit sparsity, containing approximately 34 to 64 terms. A complete list of ξ for four subsets is available in *Supplementary Materials*. The computation for $M_{D\,\max}$ and $F_{D\,\max}$ using the discovered equations has time complexity $O(\|\xi\|_0)$, where $\|\xi\|_0 \leq 64$. In comparison, the time complexity of SSGW is $O(N \log N)$ with $N = 2^{17}$.

Fig. 8 compares the predicted $M_{D\,\max}$ and $F_{D\,\max}$ from the discovered closed-form equations with exact $M_{D\,\max,\mathrm{SSGW}}$ and $F_{D\,\max,\mathrm{SSGW}}$. Almost perfect agreement is achieved with $R^2 \approx 1.0$ and max ϵ <6.5% for both training and testing data (Table 3). The discovered equations can also accurately predict $M_{D\,\max}$ and $F_{D\,\max}$ for the reserved scenarios (i.e., $h = 7 \sim 8$ m) with $R^2 \approx 1.0$, max $\epsilon = 6.7\%$ and mean $\epsilon = 0.5\%$. The up to 6.7% error is partially due to the fact that the input features are identified based on SFWT, which has discrepancies with SSGW's solutions for shallow water waves with $kh < \pi/15$ and for waves with large steepness (see Figs. 4c-d).

The discovered equations can be generalized to different water depth, beyond the scope of $0.5 \leq h \leq 8.0$ m, because waves are characterized by dimensionless parameters, such as kh and H/h. The discovered closed-form equations are applicable as long as waves satisfy: (i) $kh \leq 2\pi$; and (ii) $kH/2 \leq$ highest computable waves by SSGW (see Fig. 2).

APPLICATIONS

Evaluation of vegetation stem breakage

Coastal salt marshes experience wave-induced forces that can lead to stem deformation, buckling, and, in some cases, breakage. When the wave-induced bending stress (σ_{wave}) exceeds the allowable bending stress threshold for stems (σ_{veg}), as determined through three-point bending tests, stem breakage may occur (e.g., Heuner et al. 2015; Vuik et al. 2018). Wetlands with a significant proportion of broken stems would exhibit diminished wave damping capacity. Given that the vegetation stem width is typically significantly smaller than the wave heights in energetic wave conditions, the drag force dominates the total force (Journée and Massie 2001). Thus, σ_{wave} can be approximated as

$$\sigma_{\text{wave}} \approx \frac{M_{D\,\max} b_v}{2I} \tag{20}$$

where $I = \pi b_v^4/64$ is the second moment of area of a circular cylinder. Vuik et al. (2018) presented an explicit formula for the critical horizontal velocity (u_{crit}) for determining whether a vegetation stem can withstand wave loads without breakage under the following simplifications: (1) u is uniform along the vegetation stem; (2) u can be represented with LWT; and (3) wave crest is not involved in the vertical integration.

In contrast, the $M_{D\,\max}$ equation discovered in this study offers a better tool for general wave conditions, especially for nonlinear waves in shallow water depth. Eq. (20) becomes:

$$\sigma_{\text{wave}} \approx \frac{1}{2}\rho C_D \frac{\omega^2 H^2 M^*_{D\,\max}}{\pi b_v^2 k^2 \sinh^2 kh \sinh^2 kh_v} \tag{21}$$

The discovered equation for $M^*_{D\,\max}$ is a function of H, h, T, and h_v. In a random wave field, Vuik et al. (2018) proposed the use of $H_{1/10}$, which represents the mean of the highest 10% of waves, as a parameter to characterize the wave height that causes potential damage to vegetation. These top 10% of waves are more nonlinear, making them well-suited for the application of our discovered equations. For random waves with wave height distributions following the Rayleigh distribution, $H_{1/10} = 1.27H_{m0}$, where H_{m0} is the zero-moment wave height.

As a demonstration of this application, we consider a problem where we evaluate the percentage of broken vegetation during a tropical storm. The random wave condition is selected from field data during Tropical Storm Lee (Jadhav et al. 2013), with the following parameters: $H_{m0} = 0.7$ m, $T_p = 3.2$ s, $h = 1.04$ m. The vegetation species is *Spartina alterniflora*, with randomly distributed biophysical properties.

Based on *S. alterniflora* samples collected in upper Terrebonne Bay, Louisiana, we best-fit the probability distributions of h_v and b_v to log-normal distributions. These distributions exhibited a correlation coefficient of 0.2, with mean values of 15.8 cm for h_v and 7.9 mm for b_v. In the absence of specific field measurements for σ_{veg}, we assume a mean value of 6 MPa (the same order of magnitude of σ_{veg} as in Vuik et al. 2018), following a log-normal distribution. Following the approach outlined in Vuik et al. (2018), we conduct Monte-Carlo simulations to determine the percentage of broken vegetation stems. 2,000 correlated random samples for h_v, b_v, and σ_{veg} are drawn, all following lognormal distributions. The drag coefficient C_D is not included in the synthetic dataset and thus does not play a role in the developed equations. In real applications, C_D needs to be reintroduced. Here, we compute $C_D = 0.88$ from the unified drag coefficient formula in Zhu et al. (2023): $C_D = 0.57 + (1546/Re)^{1.11}$, where Re is the Reynolds number. For each set of these random samples, we compute σ_{wave} using Eq. (21) and compare it to σ_{veg}. The percentage of broken vegetation stems is considered the same as the probability of σ_{wave} exceeding σ_{veg}. Further details on the Monte Carlo simulations can be found in (Vuik et al. 2018).

In this particular wave condition, we estimate that 22% of vegetation stems get broken, signifying a 22% reduction in the vegetation-induced energy dissipation rate. When employing the LWT-based equation for $M^*_{D\,\max}$ in Eq. (21), 12% of vegetation stems is estimated to be broken. The use of the discovered equations allows us to obtain these results in approximately 3 minutes following the procedures in Appendix II, while the SSGW algorithm takes about 1.5 hours (on a 2017 MacBook Pro laptop computer equipped with 16GB memory and 2.3 GHz Intel Core i5 processor, with $N = 2^{13}$, tolerance of 10^{-12}). When applying numerical wave models, such as SWAN (Booij et al. 1999) and CSHORE (Johnson et al. 2012), to determine the wave attenuation by vegetation, a pragmatic treatment is to divide the vegetation field into segments in both cross-shore and long-shore directions, with segment length and width of $O(10\text{m})$. For each segment, we perform Monte Carlo simulations. Our discovered equations offer a low cost, convenient solution for estimating fractions of broken vegetation and the resulting wave attenuation.

Estimation of $F_{D\,\max}$ and $M_{D\,\max}$ on piles

The Coastal Engineering Manual (CEM, U.S. Army Corps of Engineers 2011) estimates $F_{D\,\max}$ and $M_{D\,\max}$ suing SFWT-based graphs developed by Dean (1974). These graphs depict variations in $\frac{F_{D\,\max}}{\frac{1}{2}\rho g C_D b_v H^2}$ or $\frac{M_{D\,\max}}{F_{D\,\max} b_v}$ with h/gT^2 and H/H_b, where H_b is the breaking wave height. These graphs are available for specific H/H_b values, namely, 0, 0.25, 0.5, 0.75, and 1. When H/H_b falls between these values, linear interpolations are necessary, which can introduce errors and complicate their use in numerical wave models. In contrast, the discovered equations offer a reliable and efficient alternative, providing results in milliseconds.

As a demonstration of this application, we consider problems where we calculate the drag force and bending moment of drag on an emergent, small-diameter cylindrical vertical pile. One problem is the example problem VI-7-19 from CEM, and another problem is the modified example problem 9.3.6 from Basco (2020). The cylinder width and wave conditions, together with kh and Ur, as calculated from LWT, are listed in Table 4. Both examples have cnoidal waves. Following the suggested values of $C_D = 0.7$ in CEM and Basco (2020), we calculate $F_{D\,\max}$ and $M_{D\,\max}$ using the discovered equations, LWT, Fenton's algorithm, and SSGW, as summarized in Table 4. The $F_{D\,\max}$ and $M_{D\,\max}$ from the discovered equations exhibit a maximum error of 6% when compared to the ground truth obtained from SSGW. This 6% differences are consistent with the accuracy of the discovery equations (see Table 3). Procedures to use the discovered equations with given H, h, T and h_v are provided in Appendix II.

CONCLUDING REMARKS

Vegetated ecosystems play a crucial role in coastal protection by effectively dissipating wave energy. However, associated with the energy dissipation, the vegetation stems are also exposed to wave forces, potentially leading to breakage when the wave-induced stress surpasses the allowable bending threshold. An important aspect of developing precise wave dissipation models lies in accurately considering the proportion of damaged vegetation within the vegetated area. This study demonstrates a promising application of sparse regression for predicting the maximum drag force ($F_{D\,\max}$) and maximum moment of drag ($M_{D\,\max}$) acting on vegetation stems in nonbreaking waves. Here, a synthetic dataset of $M_{D\,\max}$ and $F_{D\,\max}$ for 256,450 wave and vegetation scenarios is constructed. Inspired by the existing theoretical equations for η and u from stream function wave theory, seven dimensionless parameters characterizing wave dispersion, wave nonlinearity, and vegetation submergence are identified as feature parameters. Employing the sparse regression framework, Sparse Identification of Nonlinear Dynamics (SINDy), with one thoughtfully chosen hyper-parameter, yields the discovery of sparse yet precise closed-form equations for $M_{D\,\max}$, $F_{D\,\max}$ depending on these feature parameters. The discovered equations exhibit good accuracy, with a maximum relative error below 6.6% and a mean relative error below 1.4%. These discovered equations can be readily implemented into numerical wave models and structural analysis software, delivering results within milliseconds. The methodology elucidated in this study can also be adapted for discovering equations pertaining to other quantities of highly nonlinear waves, such as total wave force and moment, wave celerity, maximum crest elevation, and maximum bottom shear stresses for sediment transport modeling.

The discovered equations are not applicable to the most extreme waves, as SSGW is not designed for them. For a given water depth with kh up to 2π, the discovered equations works for waves with up to approximately 99% of the maximum steepness. Several algorithms have been developed for the computation of the almost highest gravity waves in finite water depth (e.g., Lu et al. 1987; Maklakov 2002). Among them, the algorithm by Maklakov (2002) can compute waves with wave steepness reaching 99.99997% of the limiting value in intermediate to deep water depth. These algorithms can be employed to fill the small data gap of the synthetic dataset in this study.

The computation of $M_{D\,\max}$ and $F_{D\,\max}$ relies on an appropriate bulk drag coefficient C_D. Current empirical C_D formulas are primarily derived from LWT, and thus, the wave nonlinearity is actually lumped into C_D. It brings inconsistency to use the LWT-based C_D in the closed-form equations that stem from nonlinear wave theory. As a future endeavor, it is worthwhile to apply SSGW to solve the energy balance equation, best fit C_D by matching the wave height reduction, and compare the fitted C_D with LWT-based empirical C_D. Zhu and Chen (2017) applied two nonlinear wave theories (i.e., Stokes second-order and cnoidal wave theories) to solve the energy balance equation for wave height. They found that wave heights from different wave theories exhibited disparities of less than 5% for emergent waves but extended to up to 25% for submerged vegetation. A hypothesis arises that, for emergent vegetation, the fitted C_D aligns closely with the LWT-based empirical C_D, while for submerged vegetation, the fitted C_D may significantly differ from LWT-based empirical C_D. Numerical experiments are desired to test this hypothesis in future research.

The discovered equations do not account for the reduction in hydrodynamic forces resulting from vegetation flexibility, or the reduction in orbital velocities within dense canopies. Consequently, the $M_{D\,\max}$ and $F_{D\,\max}$ from this study tend to overestimate real-world field conditions. Luhar and Nepf (2016) introduced the concept of effective vegetation height (h_e), which represents the height of a rigid, vertical vegetation that produces the same drag as a flexible vegetation of length h_v. They also conducted laboratory experiments to establish scaling laws for h_e. A similar concept of effective vegetation height could be introduced regarding the moment of drag. Future work could involve conducting laboratory experiments to explore scaling laws for this new effective vegetation height, and integrating it into the closed-form equations presented in this study for computing $M_{D\,\max}$ of flexible vegetation. Zhu and Chen (2017) found a reduction in in-canopy horizontal velocity compared to horizontal velocity from LWT, with a rate ranging from 6% for waves with Ur = 11 (Stokes waves) to 46% for waves with Ur = 142 (cnoidal waves). Quantifying the in-canopy velocity reduction rate relative to horizontal velocity calculated from SSGW is a crucial step towards improving the accuracy of $M_{D\,\max}$ and $F_{D\,\max}$ in future research.

For coastal structures like piles or coastal vegetation with large stem diameters like mangroves, inertia force may play an important role. The KC parameter serves as a determinant for the dominance of inertia or drag forces. For $KC < 3$, the inertia force is dominant and the drag force can be neglected; for $3 \leq KC < 45$, the full Morison equation should be employed; and for $KC \geq 45$, the drag force is dominant and the inertia can be neglected (Journée and Massie 2001; Zhu and Chen 2015). The methodology developed in this study can be applied to establish a dataset of the maximum inertia force ($F_{I\,\max}$) and maximum total force ($F_{T\,\max}$), along with their moments ($M_{I\,\max}$ and $M_{T\,\max}$). Closed-form equations for $F_{I\,\max}$ and $M_{I\,\max}$ can be obtained using sparse regression. However, finding closed-form equations for ($F_{T\,\max}$) and $M_{T\,\max}$ can be challenging. The total force is the sum of drag force and inertia force (Eq. 22), which both involve coefficients (C_D and C_M) to be determined based on wave and vegetation conditions. The ratio of C_D/C_M needs to be included in the dataset and the feature library of the regression model. Additionally, there is a phase difference in F_D and F_I, attributed to the phase difference in $u|u|$ and $\frac{du}{dt}$. These aspects are identified as potential future research topics.

DATA AVAILABILITY STATEMENT

All data, models, and code that support the findings of this study are available at https://github.com/lzhu5/EquationDiscovery_MDmax_FDmax.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Award No. 2139882/2052443. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U. S. Government. The authors thank Zhao Chen and Steve Brandt for their helpful suggestions and discussions in the early stages of this manuscript.

APPENDIX I. MORISON'S EQUATION FOR THE TOTAL HORIZONTAL FORCE

Morison et al. (1950) proposed that the total horizontal force F_T on a slender cylinder is the sum of the drag force F_D and inertia force F_I as below :

$$F_T = \underbrace{\frac{1}{2}\rho C_D b_v \int_{-h}^{\min(\eta,-h+h_v)} u|u| dz}_{\text{drag}, F_D} + \underbrace{\rho C_M \frac{\pi b_v^2}{4} \int_{-h}^{\min(\eta,-h+h_v)} \frac{du}{dt} dz}_{\text{inertia}, F_I} \tag{22}$$

APPENDIX II. PROCEDURES OF COMPUTING $M_{D\,\max}$ AND $F_{D\,\max}$

There are two options to use the discovered equations: (a) the discovered equations have been converted to symbolic expressions using *SymPy* (https://www.sympy.org/en/index.html), and these symbolic expressions can be directly imported and used in a Python script; (b) users can use the functions (in MATLAB and Python) provided in https://github.com/lzhu5/EquationDiscovery_MDmax_FDmax to obtain $M_{D\,\max}$ and $F_{D\,\max}$ directly with given wave conditions and cylinder height. For the second option, the following procedures are complied in the functions to compute $M_{D\,\max}$ and $F_{D\,\max}$:

- compute k from the linear dispersion relationship $\omega^2 = gk \tanh kh$, and compute $L = 2\pi/k$.
- compute $\mathrm{Ur} = \frac{HL^2}{h^3}$.
- crop $h_v = \min(h_v,\ h + 0.9859H)$ and compute h_v/h.
- determine the subset using Ur and h_v.
- choose the corresponding coefficients ξ in that subset, compute feature parameters $\mathbf{X}$ from Eq. (15), and compute values of functions in the feature library $\Theta(\mathbf{X})$. The coefficients ξ and codes to generate the feature library are available at https://github.com/lzhu5/EquationDiscovery_MDmax_FDmax.
- compute $M_{D\,\max}$ or $F_{D\,\max}$ as $\Theta(\mathbf{X})\xi$.

APPENDIX III. NOTATION

The following symbols are used in this paper:

C_D = drag coefficient;

C_M = inertia coefficient;

D = cylinder diameter (m);

$f_D(t, z)$ = drag force per unit length of cylinder (N/m);

$F_{D\,\max}$ = maximum drag force in a wave cycle, without the constant $\frac{1}{2}\rho C_D$(m^3/s^2);

$F^*_{D\,\max}$ = normalized $F_{D\,\max}$;

F_I = inertia force (N);

F_T = total force (N);

g = gravitational acceleration (m/s^2);

k = wave number (1/m);

K = maximum order of polynomials in the feature library;

L = wavelength (m);

LWT = linear wave theory;

N = number of Fourier modes;

h = water depth (m);

h_v = vegetation stem height (m);

H = wave height (m);

I = second moment of area of a vegetation stem (m^4);

$M_{D\,\max}$ = maximum bending moment in a wave cycle, without the constant $\frac{1}{2}\rho C_D$ (m^4/s^2);

$M^*_{D\,\max}$ = normalized $M_{D\,\max}$;

R^2 = coefficient of determination;

SFWT = stream function wave theory;

STK2 = Stokes 2nd-order wave theory;

T = wave period (s);

u = horizontal velocity (m/s);

$u_{\max}$ = maximum horizontal velocity in a wave cycle (m/s);

Ur = Ursell numer;

$x_1, \ldots, x_7$ = feature parameters;

$\mathbf{X}$ = feature parameter vector;

ϵ = relative error;

η = surface elevation (m);

$\eta_{\max}$ = maximum surface elevation in a wave cycle (m);

λ = cutoff value in the SINDy algorithm;

$\mathbf{\Theta}(\mathbf{X})$ = feature library;

ξ = coefficients of polynomial terms in the discovered equations;

ρ = water density (kg/m^3);

σ_{wave} = wave-induced bending stress on vegetation stems (Pa);

σ_{veg} = allowable bending stress of vegetation stems (Pa);

ω = angular frequency (Hz);

APPENDIX IV. SEQUENTIAL THRESHOLDED LEAST-SQUARES (STLS) ALGORITHM

The STLS algorithm is adapted from Brunton et al. (2016) and Rudy et al. (2017).

Algorithm 1 STLS(Θ, ; $\mathbf{y}$, λ, iters)

$\hat{\xi} = \arg\min_{\xi} \|\Theta\xi - \mathbf{y}\|_2^2$
bigcoeffs = $\{j : |\hat{\xi}_j| \geq \lambda\}$ ▹ Select large coefficients
$\hat{\xi}$[∼ bigcoeffs] = 0 ▹ Apply cutoff λ
$\hat{\xi}$[bigcoeffs] = STLS(Θ[:, bigcoeffs], $\mathbf{y}$, λ, iters-1) ▹ Recursively call STLS w/ fewer coefficients
return $\hat{\xi}$

REFERENCES

Anderson, M. and Smith, J. (2014). "Wave attenuation by flexible, idealized salt marsh vegetation." *Coastal Engineering*, 83, 82–92.

Babenko, K. I. (1987). "Some remarks on the theory of surface waves of finite amplitude." *Dokl. Akad. Nauk SSSR*, (5), 1033–1037.

Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., and der Vorst, H. V. (1994). *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*. SIAM, Philadelphia, PA.

Basco, D. (2020). *Design of Coastal Hazard Mitigation Alternatives for Rising Seas*. WORLD SCIENTIFIC, <https://www.worldscientific.com/doi/abs/10.1142/11462>.

Beji, S. (1995). "Note on a nonlinearity parameter of surface waves." *Coastal Engineering*, 25(1), 81–85.

Bonakdar, L., Oumeraci, H., and Etemad-Shahidi, A. (2015). "Wave load formulae for prediction of wave-induced forces on a slender pile within pile groups." *Coastal Engineering*, 102, 49–68.

Bongard, J. and Lipson, H. (2007). "Automated reverse engineering of nonlinear dynamical systems." *Proceedings of the National Academy of Sciences*, 104(24), 9943–9948.

Booij, N., Ris, R. C., and Holthuijsen, L. H. (1999). "A third-generation wave model for coastal regions: 1. model description and validation." *Journal of Geophysical Research: Oceans*, 104(C4), 7649–7666.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems." *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

Chen, Z., Liu, Y., and Sun, H. (2021). "Physics-informed learning of governing equations from scarce data.." *Nature Communication*, 12, 6136.

Clamond, D. and Dutykh, D. (2018). "Accurate fast computation of steady two-dimensional surface gravity waves in arbitrary depth." *Journal of Fluid Mechanics*, 844, 491–518.

Dalinghaus, C., Coco, G., and Higuera, P. (2023). *Using genetic programming for ensemble predictions of wave setup*. 1933–1939.

Dalrymple, R., Kirby, J., and Hwang, P. (1984). "Wave diffraction due to areas of energy dissipation." *Journal of Waterway, Port, Coastal and Ocean Engineering*, 110(1), 67–79.

Dean, R. (1974). "Evaluation and development of water wave theories for engineering application, volume i and volume ii." *Report No. Technical Report No. 14*, U. S. Army Coastal Engineering Research Center.

Fenton, J. (1988). "The numerical solution of steady water wave problems." *Computers & Geosciences*, 14(3), 357–368.

Fenton, J. D. (1985). "A fifth-order stokes theory for steady waves." *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 111(2), 216–234.

Formentin, S. M. and Zanuttigh, B. (2019). "A genetic programming based formula for wave overtopping by crown walls and bullnoses." *Coastal Engineering*, 152, 103529.

Hedges, T. S. (1995). "Regions of validity of analytical wave theories.." *Proceedings of the Institution of Civil Engineers - Water, Maritime and Energy*, 112(2), 111–114.

Heuner, M., Silinski, A., Schoelynck, J., Bouma, T., Puijalon, S., Troch, P., Fuchs, E., Schröder, B., Schröder, U., Meire, P., and Temmerman, S. (2015). "Ecosystem engineering by plants on wave-exposed intertidal flats is governed by relationships between effect and response traits." *PLoS ONE*, 10, e0138086.

Hoffmann, M., Fröhner, C., and Noé, F. (2019). "Reactive sindy: Discovering governing reactions from concentration data." *The Journal of Chemical Physics*, 150(2), 025101.

Isobe, M., Nishimura, H., and Horikawa, K. (1982). "Theoretical considerations on perturbation solutions for waves of permanent type." *Bull. Faculty of Engng, Yokohama National University*, 29–57.

Jadhav, R. S., Chen, Q., and Smith, J. M. (2013). "Spectral distribution of wave energy dissipation by salt marsh vegetation." *Coastal Engineering*, 77, 99–107.

Johnson, B., Kobayashi, N., and Gravens, M. (2012). "Cross-shore numerical model cshore for waves, currents, sediment transport, and beach profile evolution." *Report No. ERDC/CHL TR-12-22*, Vicksburg, MS: US Army Engineer Research and Development Center.

Journée, J. M. J. and Massie, W. W. (2001). *Offshore hydromechanics*. Delft University of Technology, 1st edition.

Journee, J. M. J. and Massie, W. W. (2001). *Offshore hydromechanics*. Delft Univ. of Technology, Delft, Netherlands, 1st ed edition, 1st ed edition.

Kamphuis, J. (1991). "Wave transformation." *Coastal Engineering*, 15, 173–184.

Lee, J.-S. and Suh, K.-D. (2019). "Development of stability formulas for rock armor and tetrapods using multigene genetic programming." *Journal of Waterway, Port, Coastal and Ocean Engineering*, 146(1), 04019027.

Lee, K., Trask, N., and Stinis, P. (2022). “Structure-preserving sparse identification of nonlinear dynamics for data-driven modeling.” *Proceedings of Mathematical and Scientific Machine Learning, PMLR*, <https://arxiv.org/abs/2109.05364>.

LeMehaute, B. (1976). *An introduction to hydrodynamics and water waves*. Springer Science + Business Media, LLC.

Li, S., Kaiser, E., Laima, S., Li, H., Brunton, S. L., and Kutz, J. N. (2019). “Discovering time-varying aerodynamics of a prototype bridge by sparse identification of nonlinear dynamical systems.” *Phys. Rev. E*, 100, 022220.

Loiseau, J.-C. and Brunton, S. L. (2018). “Constrained sparse galerkin regression.” *Journal of Fluid Mechanics*, 838, 42–67.

Longuet-Higgins, M. S. (1974). “On the mass, momentum, energy and circulation of a solitary wave.” *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 337(1608), 1–13.

Lu, C., Wang, J. D., and Mehaute, B. L. (1987). “Boundary integral equation method for limit surface gravity waves.” *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 113(4), 347–363.

Luhar, M. and Nepf, H. (2016). “Wave-induced dynamics of flexible blades.” *Journal of Fluids and Structures*, 61, 20–41.

Maklakov, D. V. (2002). “Almost-highest gravity waves on water of finite depth.” *European Journal of Applied Mathematics*, 13(1), 67–93.

Mendez, F. J. and Losada, I. J. (2004). “An empirical model to estimate the propagation of random breaking and nonbreaking waves over vegetation fields.” *Coastal Engineering*, 51(2), 103–118.

Morison, J. R., Johnson, J. W., and Schaaf, S. A. (1950). “The force exerted by surface waves on piles.” *Journal of Petroleum Technology*, 2, 149–154.

Petviashvili, V. I. (1976). “Equation of an extraordinary soliton.” *Fizika Plazmy*, 2, 469–472.

Pourzangbar, A. (19-21 Nov., 2012). “Determination of the most effective parameters on scour depth at seawalls using genetic programming (gp).” *The 10th International Conference on Coasts, Ports and Marine Structures (ICOPMASS 2012)*, Tehran, Iran.

Pourzangbar, A., Brocchini, M., Saber, A., Mahjoobi, J., Mirzaaghasi, M., and Barzegar, M. (2017a). “Prediction of scour depth at breakwaters due to non-breaking waves using machine learning approaches.” *Applied Ocean Research*, 63, 120–128.

Pourzangbar, A., Saber, A., Yeganeh-Bakhtiary, A., and Ahari, L. R. (2017b). “Predicting scour depth at seawalls using GP and ANNs.” *Journal of Hydroinformatics*, 19(3), 349–363.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2018). “Multistep neural networks for data-driven discovery of nonlinear dynamical systems, <https://arxiv.org/abs/1801.01236>.

Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2017). “Data-driven discovery of partial differential equations.” *Science Advances*, 3(4), e1602614.

Schmidt, M. and Lipson, H. (2009). “Distilling free-form natural laws from experimental data.” *Science*, 324(5923), 81–85.

Smits, G. F. and Kotanchek, M. (2005). “Pareto-front exploitation in symbolic regression.” *Genetic programming theory and practice II*, Springer, 283–299.

Udrescu, S.-M. and Tegmark, M. (2020). “Ai feynman: A physics-inspired method for symbolic regression.” *Science Advances*, 6(16), eaay2631.

U.S. Army Corps of Engineers (2011). “Coastal Engineering Manual - Part VI (September 28, 2011, Change 3).” *EM 1110-2-1100 (Part VI)*.

Vuik, V., Heo, H. Y. S., Zhu, Z., Borsje, B. W., and Jonkman, S. N. (2018). “Stem breakage of salt marsh vegetation under wave forcing: A field and model study.” *Estuarine, Coastal and Shelf Science*, 200, 41–58.

Wang, W.-X., Yang, R., Lai, Y.-C., Kovanis, V., and Grebogi, C. (2011). “Predicting catastrophes in nonlinear dynamical systems by compressive sensing.” *Physical Review Letters*, 106(15).

Wang, Z., Estrada, J., Arruda, E., and Garikipati, K. (2021). “Inference of deformation mechanisms and constitutive response of soft material surrogates of biological tissue by full-field characterization and data-driven variational system identification.” *Journal of the Mechanics and Physics of Solids*, 153, 104474.

Zhu, L. and Chen, Q. (2015). “Numerical modeling of surface waves over submerged flexible vegetation.” *Journal of Engineering Mechanics*, 141(8), A4015001.

Zhu, L. and Chen, Q. (2017). “Attenuation of nonlinear waves by rigid vegetation: Comparison of different wave theories.” *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 143(5), 04017029.

Zhu, L. and Chen, Q. (2019). “Phase-averaged drag force of nonlinear waves over submerged and through emergent vegetation.” *Journal of Geophysical Research: Oceans*, 124(7), 4368–4388.

Zhu, L., Chen, Q., Ding, Y., Jafari, N., and Rosati, J. D. (2019). “Semianalytical model of depth-integrated vegetal drag force based on stokes second-order wave theory.” *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 145(2), 04018041.

Zhu, L., Chen, Q., Ding, Y., Jafari, N., Wang, H., and Johnson, B. D. (2023). “Towards a unified drag coefficient formula for quantifying wave energy reduction by salt marshes.” *Coastal Engineering*, 180, 104256.

List of Tables

TABLE 1. Four subsets of the dataset and the corresponding criterion.

subset	name	criterion
I	fully submerged vegetation in Stokes wave regime	$h_v < h + 0.5H$ and $\mathrm{Ur} \leq 40$
II	nearly submerged or emergent vegetation in Stokes wave regime	$h_v \geq h + 0.5H$ and $\mathrm{Ur} \leq 40$
III	fully submerged vegetation in cnoidal wave regime	$h_v < h + 0.7H$ and $\mathrm{Ur} > 40$
IV	nearly submerged or emergent vegetation in cnoidal wave regime	$h_v \geq h + 0.7H$ and $\mathrm{Ur} > 40$

The contour lines for Ur = 40 and Ur = 26 closely converge, as illustrated in Fig. 3. Isobe et al. (1982) and Hedges (1995) proposed criteria thresholds of Ur = 26 and 40, respectively, that distinguishes Stokes and cnoidal waves. Here, we label waves with Ur < 40 as Stokes waves; otherwise, cnoidal waves.

TABLE 2. Discovered equations from SINDy, and least squares regression method for linear waves.

Feature library	ξ for $M^*_{D\,\max}$			ξ for $F^*_{D\,\max}$		
	theoretical	SINDy	least squares	theoretical	SINDy	least squares
1	0	0	0.25492	0	0	0.15126
kh_v	0	0	0.94197	2	2.0008	0.62292
$\sinh 2kh_v$	0	0	-0.38072	1	0.99997	0.59769
$\cosh 2kh_v$	-1	-1.0001	-0.59953	0	0	0.35088
$(kh_v)^2$	2	2.0015	1.5905	0	0	0.86042
$kh_v \sinh 2kh_v$	2	2	1.0011	0	0	-0.0053816
$kh_v \cosh 2kh_v$	0	0	0.99442	0	0	0.017017
$\sinh^2 2kh_v$	-1	-1.0104	-0.15622	0	0	-0.11503
$\sinh 2kh_v \cosh 2kh_v$	0	0	0.057519	0	0	0.078802
$\cosh^2 2kh_v$	1	1.0104	0.098702	0	0	0.036228

TABLE 3. The count of terms in discovered equations ($\|\xi\|_0$), hyper-parameter (λ), coefficient of determination of exact and predicted $M_{D\,\max}$ and $F_{D\,\max}$ (R^2), and the maximum and mean of relative errors (ϵ).

Subsets	Variables	$\|\xi\|_0$	λ	Data types	R^2	max ϵ (%)	mean ϵ (%)
I	$M^*_{D\,\max}$	64 (40)‡	0.4	training / testing	0.999 / 0.999	2.27 / 2.27	0.22 / 0.22
	$F^*_{D\,\max}$	52	0.0875	training / testing	0.999 / 0.999	2.49 / 2.49	0.31 / 0.31
II	$M^*_{D\,\max}$	36 (33)‡	0.8	training / testing	0.999 / 0.999	1.76 / 1.76	0.04 / 0.04
	$F^*_{D\,\max}$	43	0.5	training / testing	0.999 / 0.999	1.06 / 1.06	0.01 / 0.01
III	$M^*_{D\,\max}$	45 (45)‡	0.0525	training / testing	0.999 / 0.999	6.21 / 6.21	1.20 / 1.19
	$F^*_{D\,\max}$	45	0.0525	training / testing	0.999 / 0.999	6.13 / 6.13	1.21 / 1.20
IV	$M^*_{D\,\max}$	35 (34)‡	0.7	training / testing	0.999 / 0.999	6.63 / 6.63	1.36 / 1.36
	$F^*_{D\,\max}$	34	0.6	training / testing	0.999 / 0.999	6.47 / 6.47	1.44 / 1.43

‡ The count of common terms in the discovered equations for $M^*_{D\,\max}$ and $F^*_{D\,\max}$ is listed in parentheses.

TABLE 4. Wave and cylinder conditions and dimensionless wave parameters in the example problems. $F_{D\,\max}$ and $M_{D\,\max}$ are computed from LWT, Fenton's algorithm, the discovered equations, and SSGW.

		Example 1: $H = 2.9$ m, $h = 5.0$ m, $T = 8$ s, $b_v = 70$ cm	Example 2: $H = 2.5$ m, $h = 4.5$ m, $T = 10$ s, $b_v = 30$ cm
	kh	0.59	0.44
	Ur	65	114
$F_{D\,\max}$ (N)	LWT	4651	1552
	Fenton's algorithm	11854	4256
	Discovered eqs.	10561	4220
	SSGW	**11230**	**4034**
$M_{D\,\max}$ (N·m)	LWT	12288	3602
	Fenton's algorithm	54739	17105
	Discovered eqs	48941	17270
	SSGW	**52313**	**16299**

List of Figures

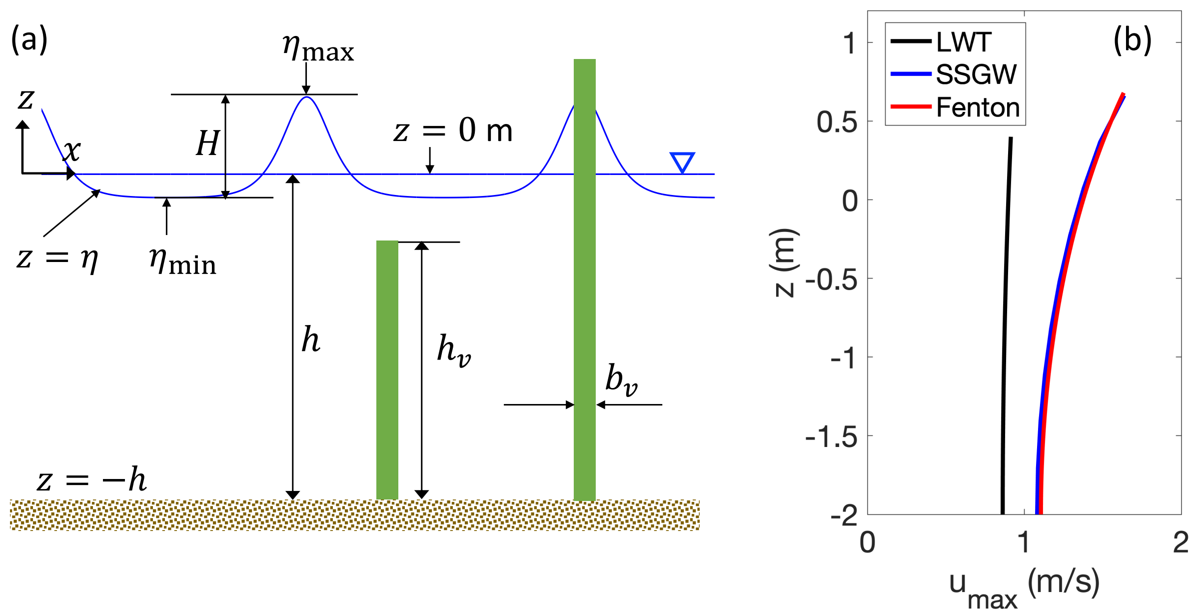

Fig. 1. (a) A sketch of submerged and emergent vegetation in free surface waves. (b) $u_{max}(z)$ as calculated from linear wave theory, Fenton's algorithm, and SSGW for a shallow water wave with $H = 0.8$ m, $h = 2.0$ m, and $T = 10.0$ s.

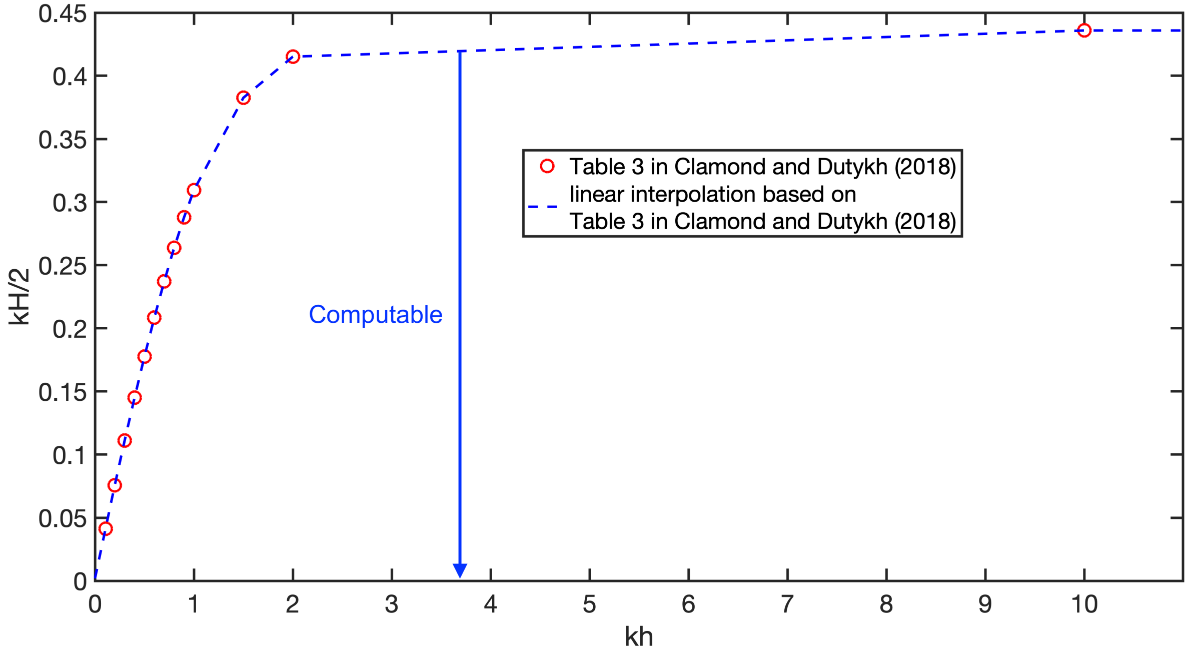

Fig. 2. Highest computable waves by SSGW, indicated by $kH/2$, vary as a function of kh.

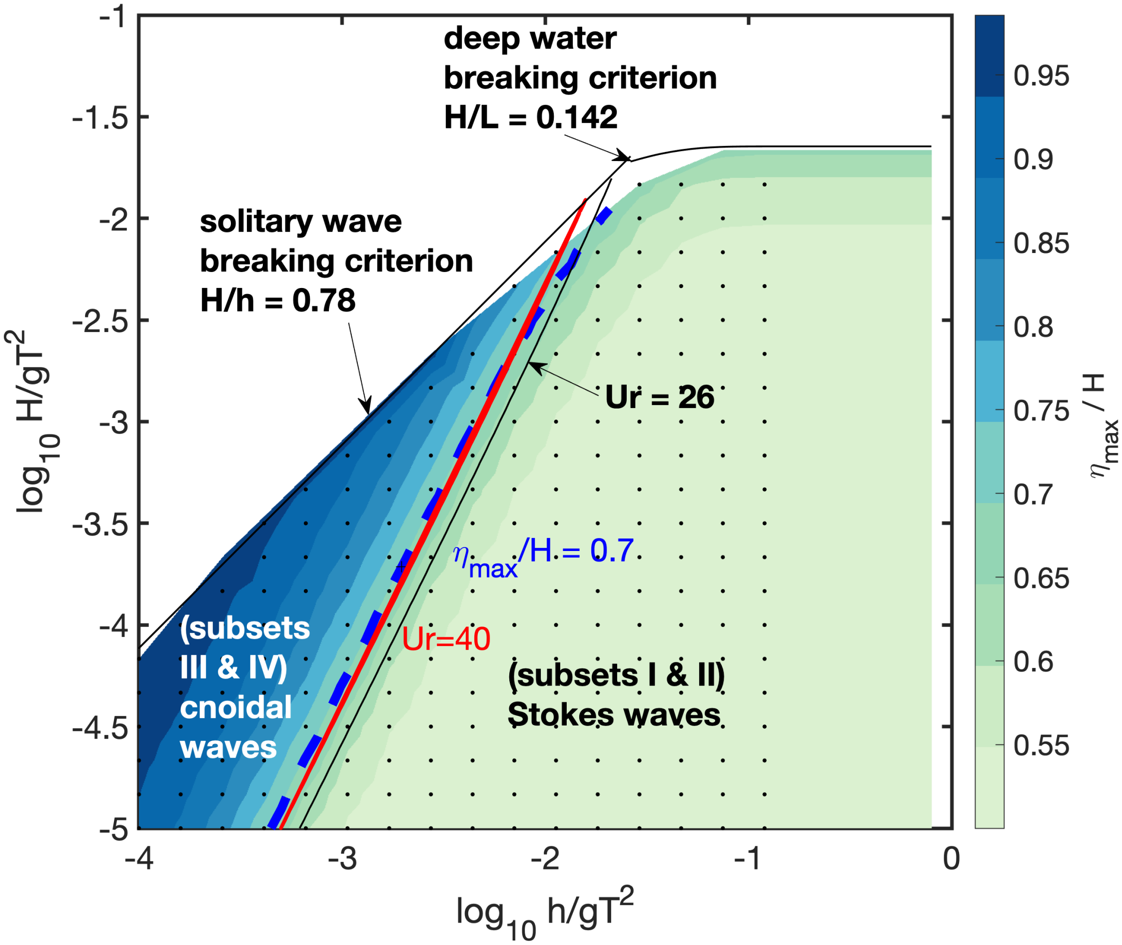

Fig. 3. Crest elevation to wave height ratios, as computed from SSGW. The blue dashed line represents the contour of $\eta_{\max}/H = 0.7$. The red line represents the contour of Ur = 40. Black dots represent wave scenarios in the synthetic dataset.

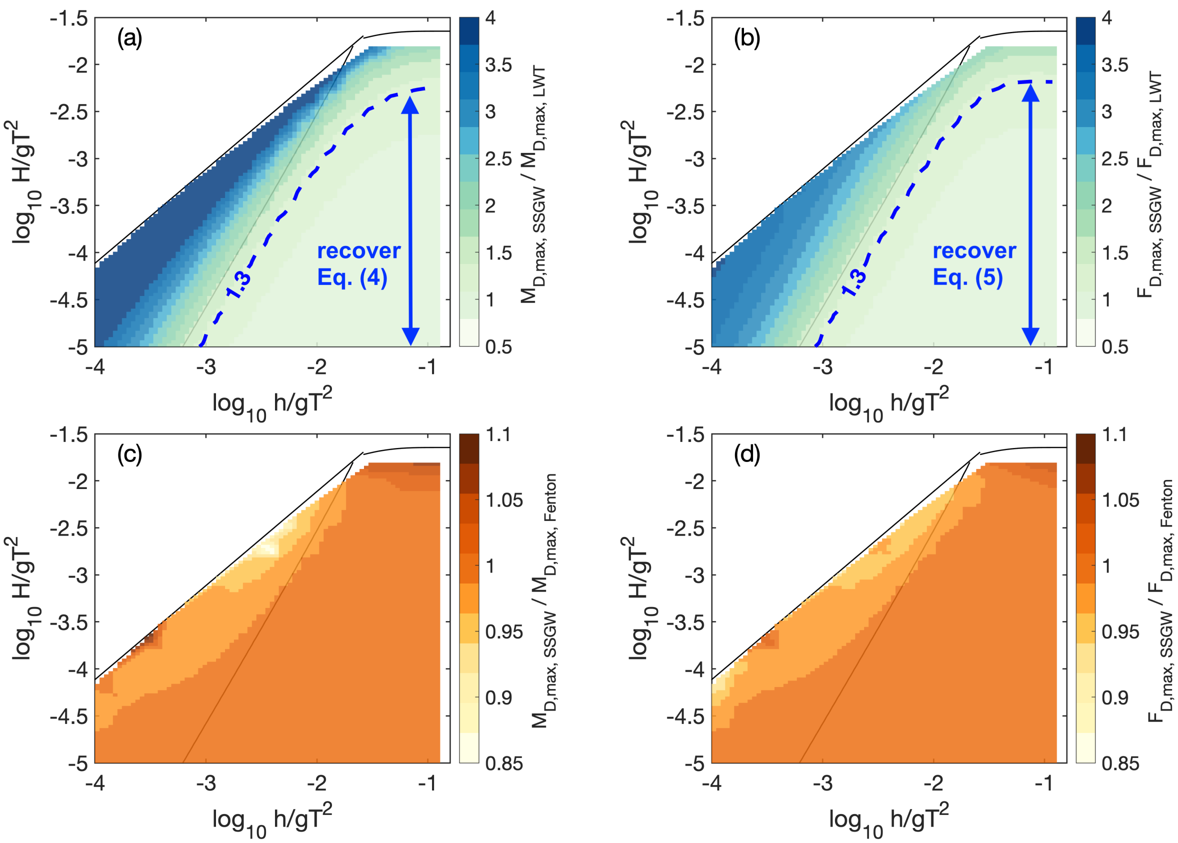

Fig. 4. $M_{D\,\max,\text{SSGW}}/M_{D\,\max,\text{LWT}}$ and $F_{D\,\max,\text{SSGW}}/F_{D\,\max,\text{Fenton}}$ on an emergent vegetation.

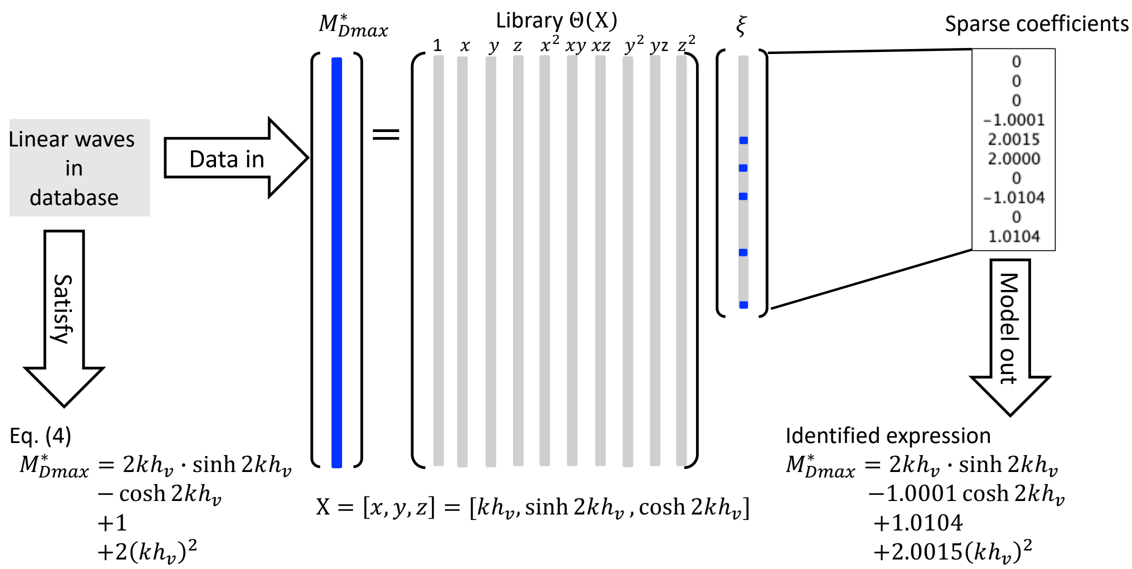

Fig. 5. Schematic of the SINDy algorithm for discovering the theoretical equations for $M^*_{D\,\max,\mathrm{LWT}}$.

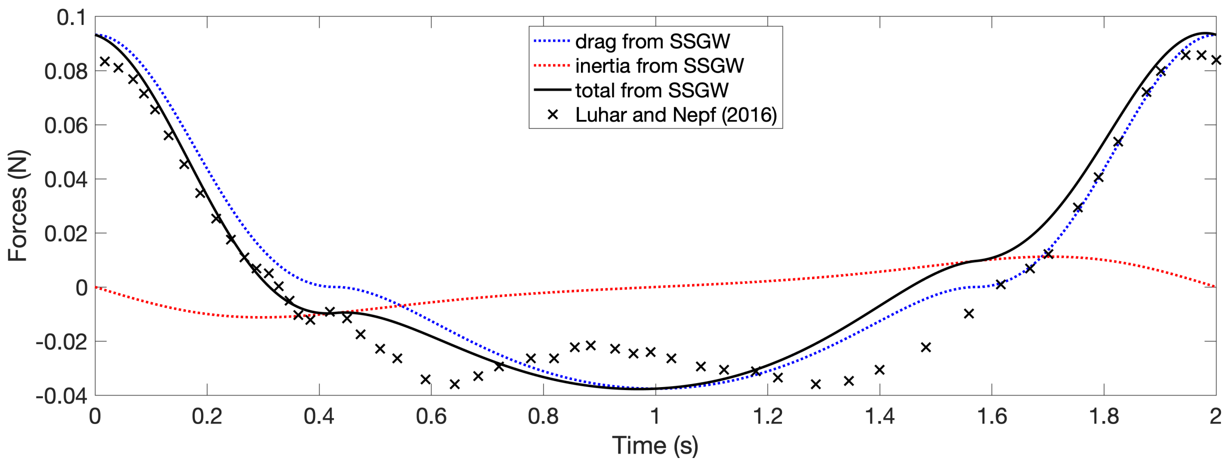

Fig. 6. Computed and measured horizontal forces exerted on the mimic vegetation.

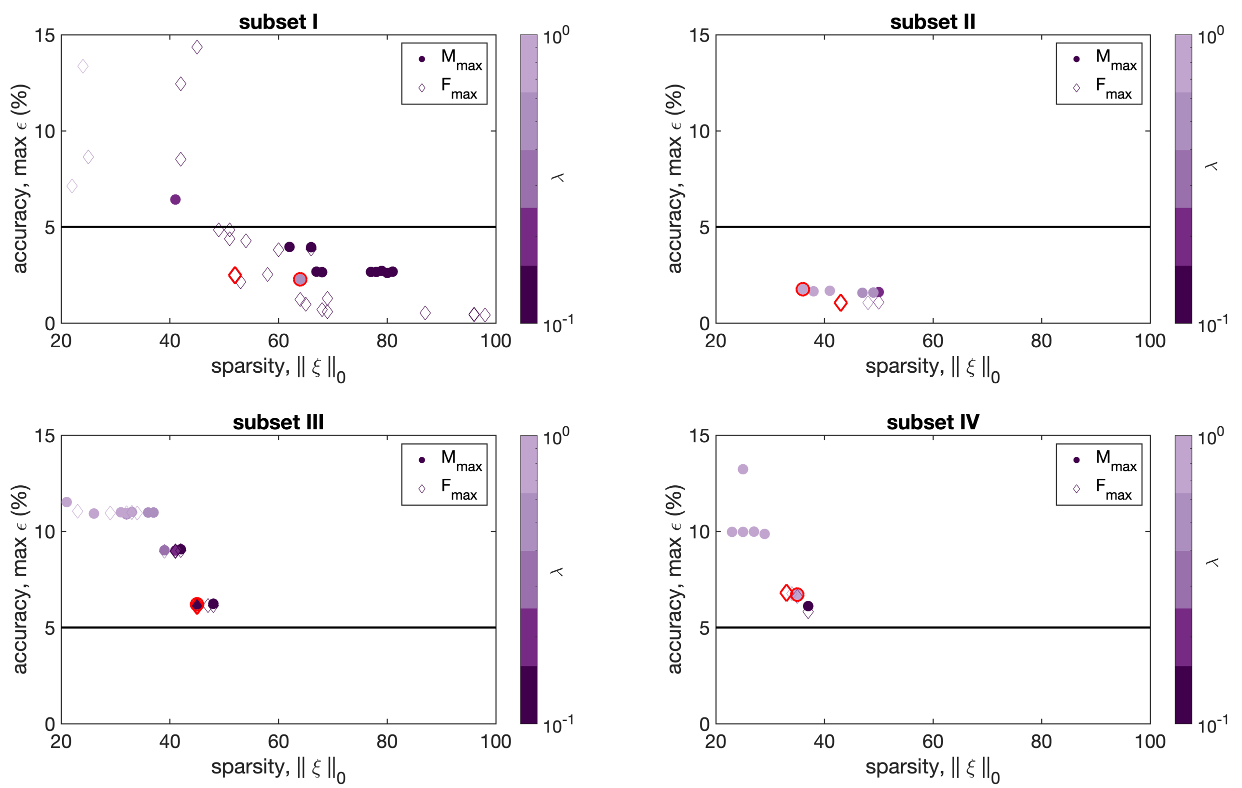

Fig. 7. Pareto front for $M^*_{D\,\max}$ and $F^*_{D\,\max}$. Red symbols represent the best trade-offs between accuracy and sparsity.

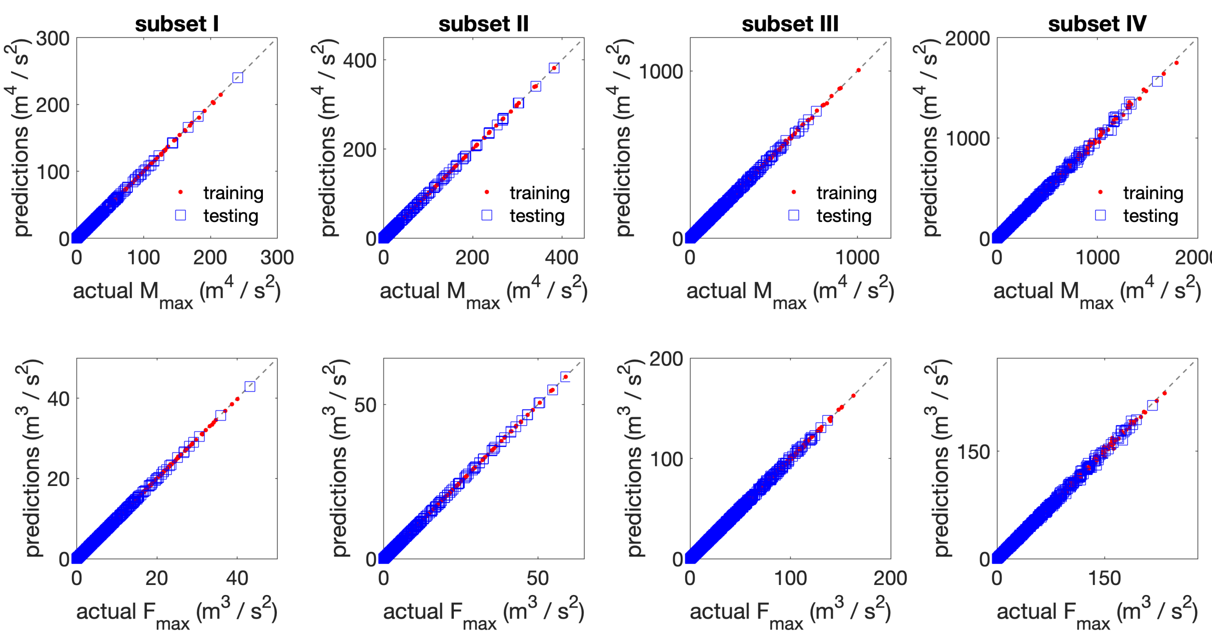

Fig. 8. Comparisons between predictions and ground truth for $M_{D\,\max}$ (upper panel) and $F_{D\,\max}$ (lower panel).