OCEAN
ISSN 3008-1203

Open Access
Research Article

Proximal policy optimization–based noncausal control for wave energy conversion systems

Yao Zhang[1] ✉, Tianyi Zeng[2], Vincentius Wijaya[3], Yutong Song[1,3], Richard Bucknall[1], and Stephen Turnock[3]

[1] *Department of Mechanical Engineering, University College London, London WC1E 7JE, UK*
[2] *Rolls-Royce UTC, University of Nottingham, Nottingham NG8 1BB, UK*
[3] *School of Engineering, University of Southampton, Southampton SO16 7QF, UK*

Cite This: *Ocean*, **2026**, *2*, 9470015

Read Online

ABSTRACT: This paper proposes a framework for wave energy converter (WEC) control that utilizes proximal policy optimization (PPO), which is a state-of-the-art policy gradient reinforcement learning algorithm. Unlike conventional model-based controllers that rely on precise models, the proposed framework is model-free and noncausal, utilizing wave prediction to enhance energy harvesting. By integrating PPO with a noncausal controller, the proposed control scheme adaptively adjusts the control parameters in real time based on the response of the WEC and wave predictions. To the best of the authors' knowledge, this is the first attempt to utilize noncausal-based PPO in a WEC application. The proposed framework directly addresses the challenges of controlling emerging WEC systems, and it is particularly suited to soft-body devices, e.g., dielectric elastomer generators and dielectric fluid generators, which are difficult to model accurately. The proposed control scheme was evaluated with different wave prediction horizons and compared against a conventional reactive controller. The results demonstrate that the proposed control scheme obtains higher energy generation while maintaining stable operation. In addition, the proposed scheme exhibits robust performance under wave prediction error and uncertainty conditions.

KEYWORDS: wave energy converters, proximal policy optimization, noncausal control, reinforcement learning, wave prediction, policy gradient methods

1 Introduction

Ocean wave energy is a promising renewable energy resource that offers high energy density compared with other renewable sources, e.g., wind and solar energy [1,2]. With significant global potential, wave energy can contribute substantially to satisfying global energy demands [3]. In addition, various design concepts for wave energy converters (WEC) have been proposed, including point absorbers [4], overtopping devices [5], oscillating water columns [6], and attenuators [7]. However, WEC systems have not achieved widespread commercialization due to their high levelized cost of energy (LCoE) compared with more mature renewable technologies [8]. To address this issue, the Engineering and Physical Sciences Research Council in the United Kingdom (UK) published the Wave Energy Roadmap in 2020 [9], which outlines strategic objectives to reduce the LCoE to £90/MWh by 2035 and deliver an installed capacity of 22 GW by 2050. Achieving these targets requires improved control strategies that can reduce costs and increase the survivability of WEC systems.

There is a broad consensus that optimal control of WECs is a noncausal problem, and noncausal control means that the controller requires wave prediction information to achieve maximum energy generation. Early WEC control schemes include reactive control [10,11], latching [4,12], and declutching [13,14] systems, which are primarily based on matching the frequency of the device with the incoming wave frequency. These methods are effective for regular wave conditions; however, they frequently struggle when handling irregular ocean waves [12]. Thus, more advanced model-based approaches, e.g., model-predictive control (MPC) [15], have emerged. MPC can handle constraints and optimize energy generation performance over a finite prediction horizon. However, these methods are dependent on accurate mathematical models of the WEC systems, and obtaining such precise models is challenging due to the inherent hydrodynamic nonlinearities, viscous effects, and the complex interactions between the WEC and the surrounding fluid [16]. Nonetheless, MPC provides an optimal control solution for WEC systems in the absence of model mismatch. Furthermore, it is inherently noncausal and utilizes future wave information to optimize current control input, and this reliance on future wave elevation information highlights the importance of accurate wave prediction techniques.

Conventional wave prediction methods include autoregressive (AR) models, which estimate the future wave elevation based on past measurements [17]. Such methods have demonstrated satisfactory performance in low sea states; however, their accuracy tends to degrade under nonlinear sea conditions. To address this limitation, machine learning (ML)-based wave prediction methods, e.g., hybrid empirical mode decomposition (EMD)-based support vector regression models [18] and neural networks for excitation force forecasting [19,20], have been proposed and demonstrated improved accuracy. In addition, these prediction techniques can be

Received: September 18, 2025; **Revised:** December 4, 2025
Accepted: December 11, 2025

✉Address correspondence to Yao Zhang, Yao.Zhang@ucl.ac.uk

integrated into noncausal control frameworks to improve energy conversion efficiency.

In addition to the wave prediction task, ML technology presents opportunities for WEC control that address the limitations of conventional model-based methods. For example, reinforcement learning (RL) has demonstrated remarkable success in complex control tasks by learning optimal control policies through interaction with the environment rather than relying on explicit mathematical models [21]. Model-free RL is well suited for handling the modeling complexities inherent in WEC systems, and this is particularly true for emerging WEC systems, e.g., dielectric elastomer generators (DEGs) and dielectric fluid generators (DFGs). However, modeling these types of WECs accurately is challenging. Figure 1 shows the operational principles of soft-body DEGs and DFGs, highlighting their flexible nature and associated modeling complexities. These characteristics make them ideal candidates for model-free approaches [22,23].

Recent applications of RL to WEC control have shown promising results. For example, Anderlini et al. [24] investigated using RL to optimize damping coefficients in point absorber systems. Subsequent research extended this to reactive control of two-body point absorbers [25] and addressed real-time implementation challenges [26]. In addition, improved energy harvesting efficiency has been observed in oscillatory wave surge converters [27], and comprehensive wave-to-wire control frameworks that utilize RL were investigated by [28]. However, many existing RL-based WEC control methods focus on value-based algorithms, e.g., Q-learning and the double deep Q-network [29]. These methods are primarily designed for discrete action spaces and may not be effective for the continuous control actions in WEC systems. In contrast, policy gradient methods, e.g., the proximal policy optimization (PPO) algorithm, offer significant advantages for continuous control applications [30]. For instance, the PPO algorithm ensures stable, sample-efficient policy updates and exhibits robust performance in many continuous control tasks. In summary, learning-based control for offshore energy systems is a rapidly advancing research area that is expected to expand over the next decade. However, in its current state, learning-based control for a WEC system does not offer an "ultimate" solution to the control problem [31]. Nonetheless, such methods can be considered promising alternatives to conventional model-based strategies under appropriate operating conditions.

Policy gradient methods, e.g., the PPO algorithm, are well suited to continuous control; however, the integration of such methods in WEC noncausal control has not been investigated. Thus, this paper addresses these challenges by proposing a PPO-based noncausal control framework for WEC systems. To the best of our knowledge, this is the first application of PPO in the noncausal control of WEC systems. The primary contributions of this study are summarized as follows:

- A PPO-based noncausal control framework for WECs is proposed. The proposed framework integrates wave prediction with model-free, policy gradient methods in wave energy conversion.
- A real-time adaptive gain adjustment mechanism is implemented in the proposed framework using PPO, which allows us to handle the model uncertainties inherent in WEC systems.
- The effectiveness of the proposed control scheme is validated using realistic, irregular wave data gathered off the coast of Cornwall, UK.
- Generally, the proposed model-free framework is applicable to other types of WEC, especially for emerging soft-body devices, including DEGs and DFGs, which are difficult to model.
- The robustness of the proposed control scheme is evaluated, and the results demonstrate its ability to maintain stable performance even in the presence of model uncertainties and wave prediction errors.

The remainder of this paper is organized as follows. Section 2 presents the WEC energy maximization problem statement, and Section 3 introduces the proposed PPO-based noncausal control scheme, including the formulation of the PPO algorithm and its integration with wave prediction. Section 4 presents the experimental simulation setup and performance evaluation. Finally, the paper is concluded in Section 5, including a summary of future work.

2 WEC problem statement

This section formulates the WEC energy optimization problem. The WEC control problem can be posed as an optimal control

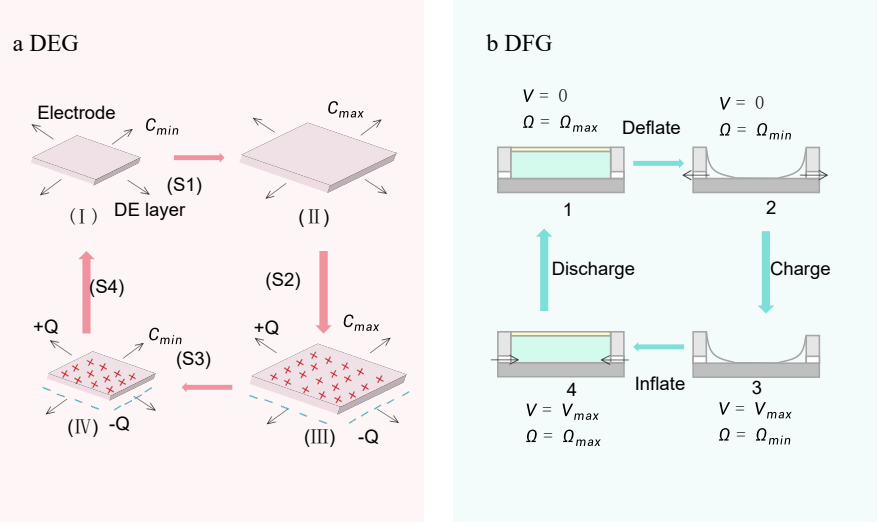

Figure 1 Operational principles of soft-body WEC devices. (a) DEG showing a four-stage cycle with large elastic deformations and time-varying capacitance. (b) DFG showing a charge–discharge cycle with significant volume changes. These complex dynamics increase the challenges associated with accurate modeling.

https://doi.org/10.26599/OCEAN.2026.9470015

problem, where the goal is to maximize the absorbed power. Here, for safety concerns, the heave motion must be constrained, and the power-takeoff (PTO) force must be constrained due to the limitations of the device. These can be expressed as follows:

$$\begin{aligned} |z_k| &\le z_{\max} \\ |u_k| &\le u_{\max} \end{aligned} \tag{1}$$

where z_k and u_k denote the discrete-time heave and control input, respectively. In addition, $z_{\max}$ and $u_{\max}$ denote the maximum heave motion and PTO force that can be handled by the device, respectively. Then, the optimal control problem can be formulated as follows:

$$\min_{u_0,\ldots,u_{N-1}} \frac{1}{N}\sum_{k=0}^{N-1} l_k(x_k,u_k), \tag{2}$$

subject to the system dynamics

$$x_{k+1} = A_{\mathrm{d}}x_k + B_{\mathrm{u,d}}u_k + B_{\mathrm{w,d}}w_k \qquad y_k = C_{\mathrm{d}}x_k, \tag{3}$$

where N denotes the prediction horizon step, $(A_{\mathrm{d}}, B_{\mathrm{u,d}}, B_{\mathrm{w,d}}, C_{\mathrm{d}})$ represents the discrete-time state-space realization, w_k denotes the wave elevation, y_k denotes the output, and x_k denotes the states. Furthermore, l_k is the stage cost, which is defined as

$$l_k(x_k,u_k) = \frac{1}{2}x_k^{\mathrm{T}}Qx_k + z_k u_k + \frac{1}{2}ru_k^2, \tag{4}$$

where Q and r are the weighting matrices for the states and control input, respectively. The components of the stage cost are described as follows.

1. $\frac{1}{2}x_k^{\mathrm{T}}Qx_k$ penalizes the state deviation, where the weight matrix Q influences the stability of the controlled system and serves as a tuning parameter to enforce the states constraint Eq. 1.
2. $-z_k u_k$ represents the instantaneous absorbed power by the PTO, which must be maximized. Note that the optimization problem is formulated as a minimization; thus, the sign convention is inverted in this expression.
3. $\frac{1}{2}ru_k^2$ penalizes excessive control force, where r functions as a tuning parameter for the control input constraint Eq. 1.

This optimal control problem Eq. 2 has the following analytical solution:

$$u_k = K_x^* x_k + K_w^* \boldsymbol{w}, \tag{5}$$

where K_x^* and K_ω^* denote the optimal feedback and feedforward gains, respectively, which can be precomputed offline. Furthermore, $\boldsymbol{w} = [w_k, \ldots, w_{k+n_{\mathrm{p}}-1}]$ is the vector of the predicted wave elevation from timestep k to $k+n_{\mathrm{p}}-1$, where n_{p} denotes the number of predicted steps. This formulation is optimal for a perfectly modeled WEC; however, real-world WECs frequently suffer from modeling inaccuracies due to hydrodynamic nonlinearities and complex dynamics, which is especially true for emerging WECs, e.g., DEGs, DFGs, and origami-based WECs. In these cases, the gain parameters in Eq. 5 must be adapted in real time to maintain sufficient performance. Thus, in the following section, an RL method is proposed as a data-driven approach to adjust these gains in an online manner.

3 Noncausal PPO for WEC control

This section introduces the proposed noncausal PPO scheme for WEC control. Section 3.1 discusses relevant background information about the PPO algorithm, and Section 3.2 presents the proposed PPO-based noncausal control scheme. Then, Section 3.3 discusses the implementation of the training and deployment of the proposed control scheme.

3.1 PPO background

The PPO algorithm is a model-free RL algorithm that does not require an explicitly formulated system model. This method only requires knowledge about the dimension of the state and action spaces. Compared with conventional supervised learning methods, the PPO algorithm learns a control policy directly through interactions with the environment without identifying the system model and designing a model-based controller. This simplifies the tuning process because only the PPO policy requires tuning. Compared with conventional RL methods, e.g., Q-learning, the PPO algorithm demonstrates improved training stability because policy updates are limited by a clipped objective function that constrains parameter changes.

The goal of RL is to learn an optimal policy $(\pi^*(a_k|s_k))$ that maximizes the expected cumulative discounted reward. This is commonly defined using the return (G_k) with timestep k:

$$G_k = \sum_{l=0}^{\infty} \gamma^l R_{k+l}, \tag{6}$$

where R_{k+l} is the reward received at timestep $k+l$, and $\gamma \in (0,1]$ is the discount factor. To measure the long-term quality of the states (s_k) under a given policy π, the following state-value function is introduced:

$$V^{\pi}(s_k) = \mathbb{E}_{\pi}\left[G_k|s_k\right]. \tag{7}$$

Note that optimizing policies directly to maximize the expected return can lead to unstable learning due to large parameter updates. The PPO algorithm addresses this issue by introducing a clipped surrogate objective that constrains updates, effectively achieving a balance between policy improvement and training stability [30]. The PPO objective is given as follows:

$$L^{\mathrm{CLIP}}(\theta) = \hat{\mathbb{E}}_k\left[\min\left(\rho_k(\theta)\,\widehat{\mathrm{Adv}}_k,\ \mathrm{clip}\left(\rho_k(\theta), 1-\Xi, 1+\Xi\right)\widehat{\mathrm{Adv}}_k\right)\right], \tag{8}$$

where

$$\rho_k(\theta) = \frac{\pi_\theta(a_k|s_k)}{\pi_{\theta_{\mathrm{old}}}(a_k|s_k)} \tag{9}$$

is the probability ratio between the new and old policies, a_k is the action at timestep k, $\Xi > 0$ is the clipping threshold, and $\widehat{\mathrm{Adv}}_k$ is the advantage estimate. In addition, $\hat{\mathbb{E}}_k[\cdot]$ denotes the empirical average over the sampled timestep. The clipping effectively keeps $\rho_k(\theta)$ within $[1-\Xi, 1+\Xi]$, which prevents destructive updates, thereby stabilizing training. The advantage estimate is from the generalized advantage estimator (GAE), and it is computed from temporal-difference (TD) errors using a backward recursion over the rollout:

$$\delta_k = R_k + \gamma V_{\mathrm{old}}(s_{k+1}) - V_{\mathrm{old}}(s_k), \tag{10}$$

$$\widehat{\mathrm{Adv}}_k = \delta_k + \gamma\lambda\widehat{\mathrm{Adv}}_{k+1}, \tag{11}$$

where R_k denotes the reward at timestep k, $V_{\text{old}}(\cdot)$ denotes the value estimate from the critic at rollout time, and $\lambda \in [0,1]$ controls the tradeoff between bias and variance. Here, if $\lambda = 0$, it reduces to one-step TD (high bias and low variance). In contrast, if $\lambda = 1$, it becomes Monte Carlo return (low bias and high variance). Practically, the recursion is initialized with $\widehat{\text{Adv}}_{\text{end}} = 0$ at the end of the rollout.

Typically, PPO is implemented with an actor–critic architecture, as shown in Fig. 2. Here, the actor/policy ($\pi_\theta(a_k|s_k)$) outputs a distribution over actions, which makes them naturally suited for continuous action spaces, and the critic estimates state values for policy improvement. For continuous action spaces, the policy is modeled as a Gaussian distribution ($\mathcal{N}(\cdot)$) as follows:

$$\pi_\theta(a_k|s_k) = \mathcal{N}\big(\mu_\theta(s_k), \sigma_\theta^2(s_k)\big), \tag{12}$$

with the mean $\mu_\theta(s_k)$ and standard deviation $\sigma_\theta(s_k)$. During evaluation, the policy is typically set to be deterministic, ensuring repeatability by setting it to $\mu_\theta(s_k)$. The critic $V_\phi(s_k)$ approximates the state-value function and is trained by minimizing the squared error with respect to a bootstrapped target as follows:

$$L^{\text{VF}}(\phi) = \hat{\mathbb{E}}_k\big[(V_\phi(s_k) - V_k^{\text{targ}})^2\big], \tag{13}$$

where $V_k^{\text{targ}} = \widehat{\text{Adv}}_k + V_{\text{old}}(s_k)$ is the bootstrapped target computed. Here, $V_{\text{old}}(s_k)$ denotes the value estimate given by the critic at rollout time. Finally, the full PPO objective combines the clipped surrogate, the value loss, and an entropy regularization term $S(\pi_\theta)$ that promotes exploration:

$$L^{\text{PPO}}(\theta,\phi) = \hat{\mathbb{E}}_k\Big[L^{\text{CLIP}}(\theta) - c_1 L^{\text{VF}}(\phi) + c_2 S(\pi_\theta)\Big]. \tag{14}$$

Here, c_1 and c_2 are the contributions of the critic and entropy, respectively. Note that this combined loss is optimized using stochastic gradient descent across multiple epochs.

3.2 PPO-based noncausal control with wave prediction

This section introduces the proposed PPO-based noncausal control scheme. The PPO agent is designed to adaptively adjust the gains in Eq. 5 as follows:

$$u_k = K_{x,k} s_k + K_{w,k} \boldsymbol{w}, \tag{15}$$

where $s_k = [z_k, \dot{z}_k]^{\text{T}}$ represents the states, and $K_{x,k}$, $K_{w,k}$ are the time-varying gains. Note that some of the states (radiation and excitation) may not be available; thus, the states in Eq. 5 are modified to include only the available states (s_k). Therefore, the controller receives the real-time measurements of the WEC responses, e.g., heave displacement and velocity. Here, the K_x and K_w gains are adjusted by the action of the agent. Furthermore, the action space of the PPO agent is constrained as follows:

$$\mathcal{A} := \{\, a \in \mathbb{R}^{n_s+n_p} : \|a\|_\infty \le \psi \,\}, \qquad a_k \in \mathcal{A}, \qquad a_k = \begin{bmatrix} \Delta K_x \\ \Delta K_w^{(0)} \\ \vdots \\ \Delta K_w^{(n_p-1)} \end{bmatrix}. \tag{16}$$

Here, set $\mathcal{A}$ constrains the action space, $\psi > 0$ is a small constant bounding the increments, n_s denotes the number of states in s_k, and Δ represents the small increment for each gain. In this context, $\mathcal{A}$ is continuous; thus, all increments inside a_k may be updated simultaneously at each timestep. Then, the gain is updated as follows:

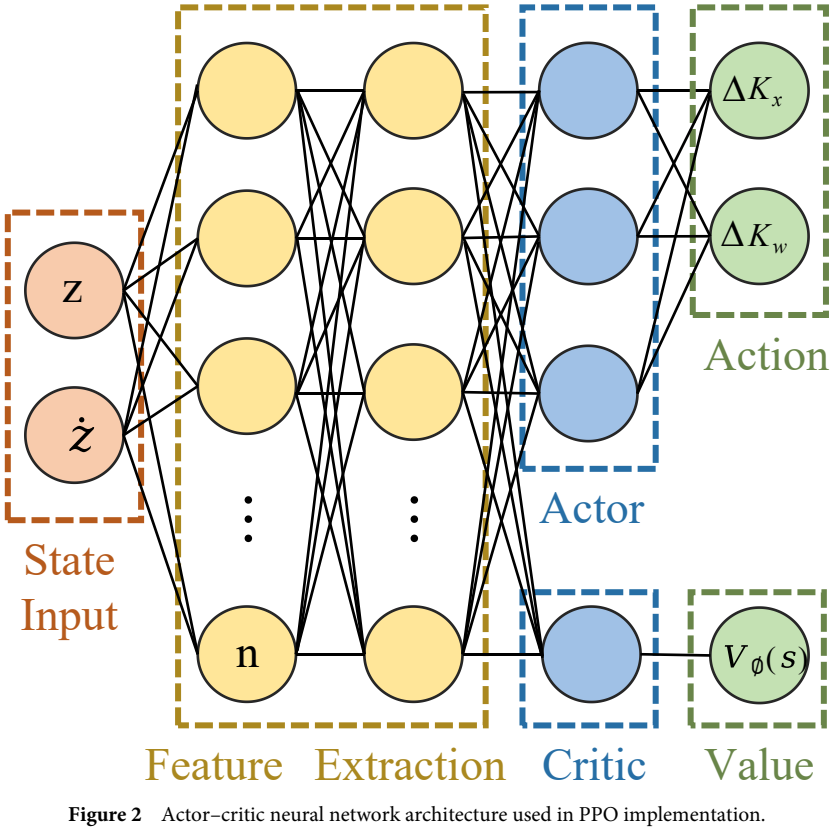

Figure 2 Actor–critic neural network architecture used in PPO implementation.

$$K_{x,k+1} = K_{x,k} + \Delta K_x, \tag{17}$$

$$K_{w,k+1} = K_{w,k} + \Delta K_w. \tag{18}$$

The reward function is designed to maximize energy extraction while penalizing the constraint violations:

$$R_k = \alpha P_k - \beta_1 u_k^2 - \beta_2 \mathbb{I}_{|u_k|>u_{\max}}, \tag{19}$$

where $P_k = -u_k \dot{z}_k$ is the instantaneous power (refer to Remark 2), α, β_1, and β_2 are weighting parameters, and $\mathbb{I}_{|u_k|>u_{\max}}$ is an indicator function that equals 1 if $|u_k| > u_{\max}$ (i.e., a constraint violation) and 0 otherwise. From the power, the energy can be calculated as the cumulative sum of the instantaneous power as follows:

$$E_k = \sum_{i=0}^{k} P_i T_s, \tag{20}$$

where E_k denotes the total energy up to step k, with T_s representing the sampling period.

Remark 1. *The proposed noncausal control scheme in Eq. 15 requires n_p wave measurement/prediction steps. Some of the available wave predictors include deterministic sea wave prediction* [32]*, ML methods, long short-term memory* [33]*, or AR methods* [34].

Remark 2. *The proposed noncausal control scheme in Eq. 15 utilizes the mechanical power of the WEC as the reward function in Eq. 19. Note that this is a common objective function in many previous WEC studies* [15,35]*. Although mechanical to electrical power requires conversion, many electrical power PTO share the same mechanical power formulation (where force is taken as the control input). Thus, the proposed formulation is broadly applicable.*

3.3 Implementation of noncausal PPO control

The PPO training process for WEC control is summarized in Algorithm 1. The overall architecture of the proposed PPO-based noncausal control scheme is shown in Fig. 3, and a summary of the corresponding algorithm is given in Algorithm 2.

Algorithm 1 PPO training process for WEC Control

Initialize θ_0, ϕ_0, α, β_1, β_2, n_p

for iteration $k = 0, 1, 2, \ldots$ **do**

Collect trajectory with policy π_θ under n_p step of wave predictions with random initial condition x_0

Compute advantages $\widehat{\mathrm{Adv}}_k$ (GAE)

for epoch $e = 1, \ldots, n_{\text{epoch}}$ **do**

Using gradient descent, θ and ϕ are updated according to the total loss Eq. 14 with weights α, β_1, and β_2.

end for

end for

4 Simulation setup and results

In this section, the experimental simulation setup and evaluation results for the proposed PPO-based noncausal control scheme are presented. Here, the point absorber WEC was selected as the case study for the RL environment because it is a commonly used benchmark problem[24,36–44]. In the following, Section 4.1 introduces the model of the point absorber, and Section 4.2 discusses the training of the PPO agent and the corresponding hyperparameter tuning process. Finally, Section 4.3 evaluates the performance of the PPO agent under various wave prediction horizon conditions, as well as in the presence of disturbances and wave prediction errors.

4.1 Simulation model and parameters

Here, the model of the point absorber is introduced. Note that this model is only used to construct the environment for training. However, in practice, the proposed control scheme only needs to know the dimension of the states, control input, and wave prediction. The PTO of the point absorber can be hydraulic, and the realization of this design has been reported by [45]. Specifically, the hydraulic piston is connected to the hydraulic motor coupled with a synchronous generator. The power is then conditioned through back-to-back AC/DC/AC converters prior to reaching the grid. In the generator-side converters, the control input is the q-axis current that controls the electric torque of the generator. The generator torque is proportional to the force exerted by the hydraulic fluid in the cylinders on the piston. Since they are proportional, this piston force is taken as the control input. Furthermore, the variable z_v represents the vertical midpoint position of the point absorber, and z_w denotes the seawater level. The dynamics of a point absorber WEC can be described as follows according to Newton's second law:

$$m_s \ddot{z}_v = -f_s - f_r + f_e + f_u, \tag{21}$$

Algorithm 2 PPO-based noncausal control scheme for WEC

Initialize π_θ, n_p, x_0

for iteration $k = 0, 1, 2, \ldots$ **do**

Compute n_p step of wave prediction $\boldsymbol{w}$

Compute action based on the trained policy (π_θ)

Compute control input Eq. 15 and apply it to the WEC systems

end for

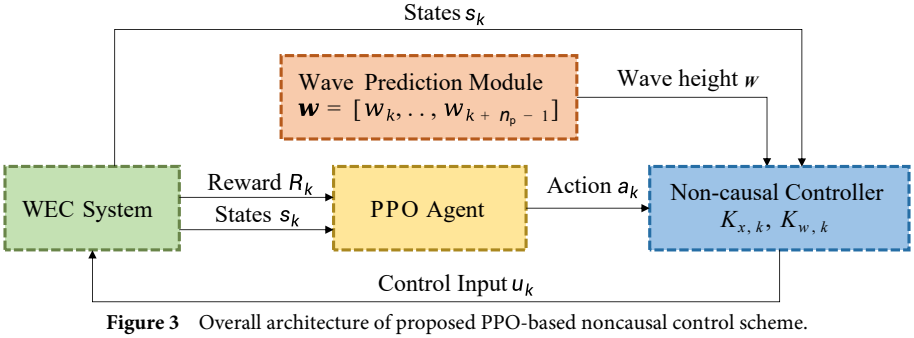

Figure 3 Overall architecture of proposed PPO-based noncausal control scheme.

where m_s, f_s, f_r, f_e, and f_u denote the point absorber mass, the hydrostatic restoring force, the radiation force, the wave excitation force, and the control force applied by the PTO system, respectively. Similar to Refs. [43,44,46], Eq. 21 can be transformed into a tenth-order continuous-time state-space representation:

$$\begin{cases} \dot{x} &= A_c x + B_{u,c} u + B_{w,c} w + \varepsilon \\ y &= C_c x \end{cases}, \tag{22}$$

with

$$A_c = \begin{bmatrix} 0 & 1 & 0_{1\times n_r} & 0_{1\times n_e} \\ -\dfrac{k_s}{m} & 0 & -\dfrac{C_r}{m} & \dfrac{C_e}{m} \\ 0_{n_r\times 1} & B_r & A_r & 0_{n_r\times n_e} \\ 0_{n_e\times 1} & 0_{n_e\times 1} & 0_{n_e\times n_r} & A_e \end{bmatrix}, B_{u,c} = \begin{bmatrix} 0 \\ \dfrac{1}{m} \\ 0_{n_r\times 1} \\ 0_{n_e\times 1} \end{bmatrix},$$
$$B_{w,c} = \begin{bmatrix} 0 \\ 0 \\ 0_{n_r\times 1} \\ B_e \end{bmatrix}, C_c = \begin{bmatrix} 0 & 1 & 0_{1\times(n_r+n_e)} \end{bmatrix}. \tag{23}$$

Here, $w = z_w$ denotes the wave elevation, which is assumed to be predictable. The total mass is defined as $m = m_s + m_\infty$, where m_∞ is the infinite frequency added mass of the point absorber. In addition, the hydrostatic stiffness is denoted k_s, and the radiation force is represented in state space form by $(A_r, B_r, C_r, 0)$, with associated states x_r and dimension n_r. The excitation force is described by $(A_e, B_e, C_e, 0)$, with states x_e and dimension n_e. Note that these state-space models can be obtained using hydrodynamic solvers, e.g., the WAMIT [47] and NEMOH [48] solvers. In addition, the ε term accounts for modeling uncertainties arising from imperfect wave prediction or the linear model assumption. The notation 0 without a subscript denotes a scalar zero, whereas a subscript indicates a zero matrix of the specified dimension. The output is $y = \dot{z}_v$, the state vector is $x = [z_v, \dot{z}_v, x_r, x_e]^T$, and the control input is $u = f_u$. For simplicity, the subscript v is omitted, with the first two states denoted z and $\dot{z}$.

The continuous-time formulation in Eq. 22 can be discretized using Euler's method or a zero-order hold:

$$\begin{cases} x_{k+1} &= A_d x_k + B_{u,d} u_k + B_{w,d} w_k + \varepsilon_k \\ y_k &= C_d x_k \end{cases}. \tag{24}$$

Here, $(A_d, B_{u,d}, B_{w,d}, C_d)$ represent the discrete-time version of the continuous-time matrices $(A_c, B_{u,c}, B_{w,c}, C_c)$. The specific numerical values for these matrices and the physical parameters of the WEC are described in the following section.

4.2 Training and hyperparameter tuning

This section discusses the training and hyperparameter tuning processes of the proposed PPO-based noncausal control scheme. For the wave energy converter model, the parameters are $m = m_s + m_\infty = 242 + 83.5 = 325.5$ kg and $k_s = 3866$ N/m, which correspond to a cylindrical body with a radius and draft of 0.35 m and 0.63 m, respectively. The dynamic representation of the radiation and excitation forces follows that reported in the literature [46]:

$$A_r = \begin{bmatrix} 0 & 0 & -17.9 \\ 1 & 0 & -17.7 \\ 0 & 1 & -4.41 \end{bmatrix}, B_r = \begin{bmatrix} 36.5 \\ 394 \\ 75.1 \end{bmatrix}, C_r = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}. \tag{25}$$

$$A_e = \begin{bmatrix} 0 & 0 & 0 & 0 & -400 \\ 1 & 0 & 0 & 0 & -459 \\ 0 & 1 & 0 & 0 & -226 \\ 0 & 0 & 1 & 0 & -64 \\ 0 & 0 & 0 & 1 & -9.96 \end{bmatrix}, B_e = \begin{bmatrix} 1549886 \\ -116380 \\ 24748 \\ -644 \\ 19.3 \end{bmatrix},$$
$$C_e = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{26}$$

In this study, the simulation employed wave height data collected off the coast of Cornwall, UK, and a segment of the data are shown in Fig. 4. Here, the first 200 s of the dataset were employed for training purposes. Given a simulation timestep of $T_s = 0.1$ s, this corresponds to 2000 steps per episode. Unlike supervised learning, RL collects data through interactions with the environment rather than relying on a fixed dataset. Thus, the amount of data used by PPO increases with the number of episodes. Here, the maximum number of training episodes was set to 500, with each episode comprising 2000 datapoints. Furthermore, the agent was evaluated periodically during the training process, and early stopping was triggered if the average reward did not improve. In addition, three prediction horizons were considered for training: $n_p = 1, 3, 5$. For each prediction horizon setting, training was terminated at different episodes, resulting in different total numbers of datapoints. The state and action spaces were constrained by the following bounds. The maximum control input was $u_{max} = 10$ kN, and the limits for heave displacement and velocity were $z_{max} = 2$ m and $\dot{z}_{max} = 2$ m/s, respectively. All simulations were executed on a CPU using a 24-core Intel Core i9-14900K processor.

The training environment and PPO model were implemented in Python using the Stable-Baselines3 (SB3) library [49]. In addition, the reward, advantage, and observation signals were normalized via running statistics, and the weighting parameters were set to $\alpha = 1$, $\beta_1 = 10^{-3}$, and $\beta_2 = 5$. The scalar coefficients α, β_1, and β_2 balance the relative importance of each term in the reward function, where a greater α value emphasizes maximizing the harvested power, a greater β_1 value enforces lower control input, and a greater β_2 value strongly discourages violating the control bound u_{max}. The hyperparameter optimization process was performed using the Optuna framework [50], and the search space and selected values for $n_p = 5$ are summarized in Table 1. Optuna, which can be installed from GitHub, is an open source hyperparameter

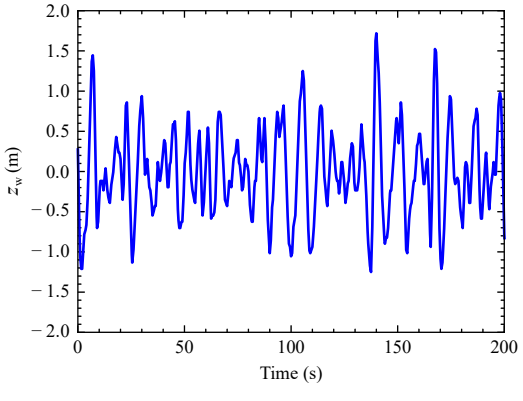

Figure 4 Segment of wave height data collected off the coast of Cornwall, UK. The wave data have been scaled appropriately according to the size of the device.

https://doi.org/10.26599/OCEAN.2026.9470015

optimization framework that automatically searches for optimal configurations in ML models. The details on how each hyperparameter is tuned and their typical values have been reported by Ref. [51]. In the current study, the hyperparameter tuning process followed the standard procedures for PPO. Here, gradient descent was performed using the Adam optimizer [52], with a linearly decaying learning rate:

$$\mathrm{lr}(p) = \begin{cases} \mathrm{lr}_{\mathrm{init}} + \dfrac{1-p}{p_{\mathrm{end}}}\left(\mathrm{lr}_{\mathrm{final}} - \mathrm{lr}_{\mathrm{init}}\right), & 0 \le 1-p \le p_{\mathrm{end}}, \\ \mathrm{lr}_{\mathrm{final}}, & 1-p > p_{\mathrm{end}}, \end{cases} \tag{27}$$

where $\mathrm{lr}_{\mathrm{init}}$ is the initial learning rate, $\mathrm{lr}_{\mathrm{final}}$ is the final learning rate, p_{end} is the end fraction of training, and $p \in [0,1]$ is the progress remaining, which decreases linearly from 1 at the start of training to 0 at the end. Furthermore, the hyperparameter optimization is based on the evaluation reward from the environment, where higher rewards indicate better configurations. In this study, 15 trials were performed with pruning, where the pruning stopped individual training early if their reward was less than the current best reward. Here, pruning was applied using Optuna's MedianPruner, which was applied after five initial warmup trials were performed without pruning. The candidate hyperparameters were sampled with the tree-structured Parzen estimator, which

Table 1 Optuna hyperparameter search space and selected values for PPO training of the WEC controller with $n_p = 5$.

Hyperparameter	Search range	Value
Initial learning rate ($\mathrm{lr}_{\mathrm{init}}$)	[1×10⁻⁴, 5×10⁻⁴], log scale	2.75×10⁻⁴
Entropy coefficient (c_2)	[1×10⁻³, 1×10⁻²], log scale	1.39×10⁻³
End fraction (p_{end})	[0.5, 1.0]	0.7899
Discount factor (γ)	[0.90, 0.90]	0.981
Value function coefficient (c_1)	[0.3, 0.7]	0.684
Max grad norm	[0.3, 0.5]	0.338
GAE parameter (λ)	[0.90, 0.98]	0.978
Clip parameter (Ξ)	[0.15, 0.25]	0.208
Minibatch size	{64, 128}	64
Number of steps to run for each environment per update	{1024, 2048}	2048
Number of epoch when optimizing the surrogate loss	{8, 10}	10
Policy and value network architecture	{64, 64;64}	64; 64
Activation function	{tanh, ReLU}	tanh

Note: Log scale indicates that the variable was sampled according to a log-uniform distribution, [·] denotes uniform sampling, and {·} corresponds to sampling from a categorical distribution. The range in the search space was adopted from [51], and a detailed explanation of each hyperparameter can be found in the SB3 documentation [49].

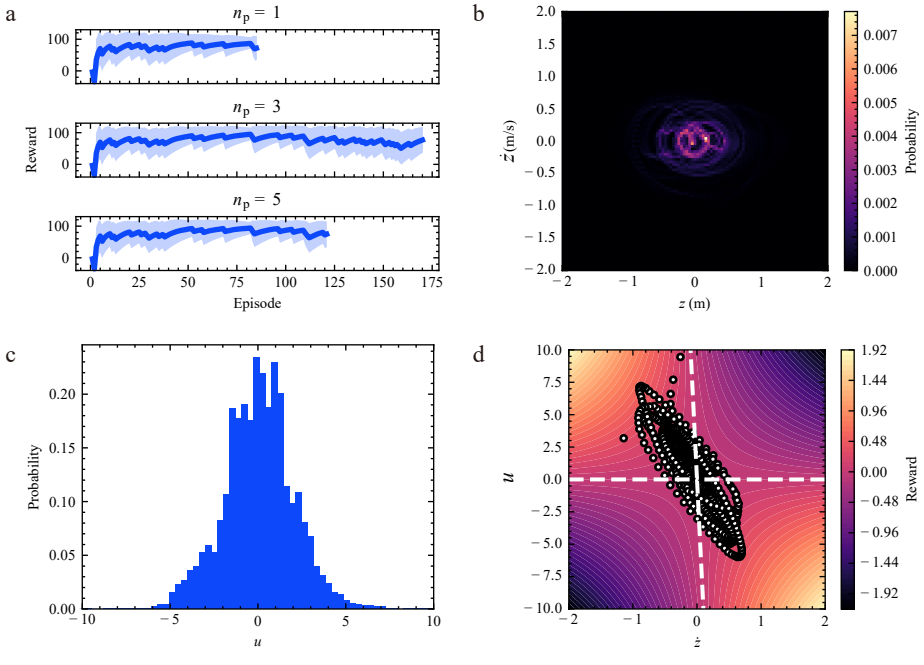

Figure 5 PPO training statistics. (a) Reward evolution over episode for each wave prediction steps (n_p). (b) Contour plot of the training data occurrence, normalized by the total number of samples, with respect to heave displacement (z) and heave velocity ($\dot{z}$) for $n_p = 5$. The contours illustrate the empirical probability of state occurrence across the training dataset. (c) Histogram plot of the training data occurrence, normalized by the total number of samples, with respect to the control input (u) for $n_p = 5$. (d) Contour plot of the reward landscape at timestep k, with heave velocity ($\dot{z}$) and control input (u) on the x-axis and y-axis, respectively. The contours were calculated using Eq. 19, and the scatter points indicate the individual data samples from the final training episode from $n_p = 5$. The white dotted line represents the boundary where the reward changes sign.

steers the search toward values that performed well in previous trials. Additional details can be found in the Optuna documentation [50].

After hyperparameter tuning, the performance of the best-trained agent was recorded, and the results are presented in Fig. 5a, which shows the evolution of the total reward per episode. Here, the mean values and standard deviation are shown as solid lines and shaded regions, respectively. As can be seen, training terminated at episodes 85, 170 and 121 for the n_p= 1, 3 and 5 cases, respectively. These results demonstrate that increasing the prediction horizon can shorten the training process because training can be terminated earlier. In addition, the increase in the reward indicates that the agent successfully learned an effective control policy.

In addition to the reward evolution, the distributions of heave elevation, heave velocity, and control input during the training process are shown in Figs. 5band 5c. As shown, the heave displacement and velocity are concentrated around the origin, which is consistent with the expected oscillatory behavior of the WEC. In addition, the control input remains within the imposed constraint of $|u| \leq 10$ kN, with violations occurring in only 0.01% of the training data. This result confirms that the soft constraints embedded in the reward function are satisfied.

Figure 5d shows the reward contour for the final training episode with $n_p = 5$. Here, the reward surface, as computed using 19, shows that the trajectory generally remains in regions of positive reward, which confirms that the agent learned to exploit favorable operating conditions. Note that occasional passages through the negative region near $(\dot{z}, u) = (0, 0)$ occur, which is expected because strictly zero velocity and control input cannot be maintained in practice.

4.3 Simulation results

In this study, the trained PPO agent was evaluated in a real-wave environment using a 200 s segment of data distinct from the training data. Here, in addition to the PPO-based noncausal controllers, a baseline controller without wave prediction was included to facilitate an effective comparison. The corresponding results are shown in Figs. 6a–6e. Figures 6a and 6b show the buoy heave elevation and velocity. As can be seen, both remain within the imposed limits of $z_{max} = 2$ m and $\dot{z}_{max} = 2$ m/s across all of the compared controllers. Furthermore, the overall heave and heave velocity response ranges are comparable among all controllers. However, toward the end of the simulation, the n_p = 5 case exhibits

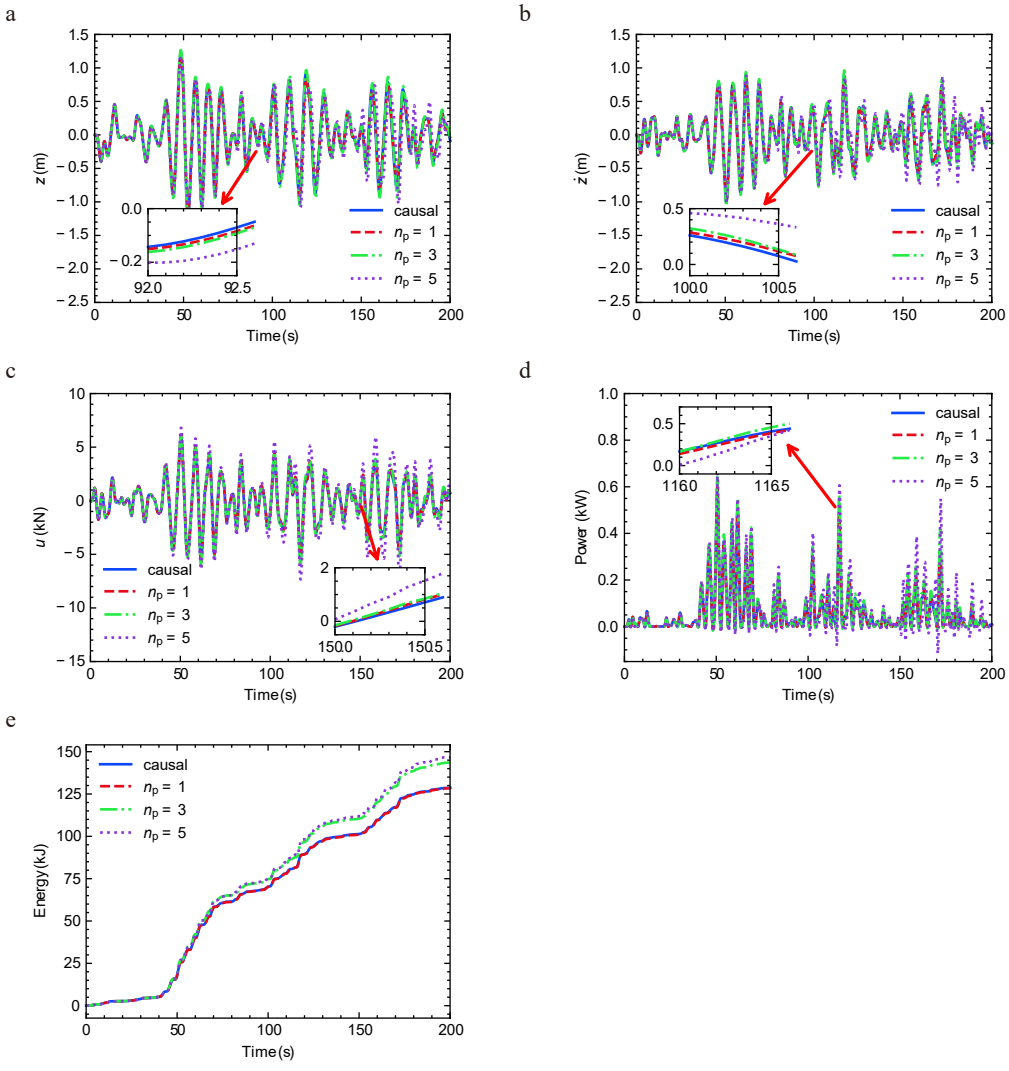

Figure 6 Simulation of trained noncausal PPO agent ($n_p = 1, 3, 5$) against a baseline causal controller (passive damper): (a) z, (b) $\dot{z}$, (c) u, (d) power, and (e) energy.

https://doi.org/10.26599/OCEAN.2026.9470015

stronger oscillations, which suggests that this configuration may involve more aggressive motion. Figure 6c shows the control input, which remains under the constraint of 10 kN. However, as shown, the higher wave prediction steps correspond to higher utilization of the control input, which is particularly evident from the peaks and troughs, where the n_p = 5 controller generates stronger forces compared with the other cases.

Figures 6d and 6e show the key performance metrics of the WEC, i.e., the generated power and cumulative energy. The results demonstrate that the causal baseline method and the n_p = 1 case achieved similar energy levels, and the n_p = 3 and n_p = 5 cases yielded higher energy than the baseline causal controller. Note that the improvement in the energy generation output from n_p = 3 to n_p = 5 is small; however, the training cost was reduced because fewer episodes were required for training. A quantitative comparison of the energy performance and control effort (CE) is shown in Table 2. The CE quantifies the total energy of the control actions applied over the simulation, computed as the sum of squared control inputs, which is defined as follows:

$$\mathrm{CE} = \sum_{k=0}^{N_{\mathrm{sim}}} |u_k|^2 T_s, \tag{28}$$

where N_{sim} denotes the total number of simulation steps (2000 steps). The results presented in Table 2 demonstrate that the cumulative energy for $n_p = 1$ is slightly lower than the baseline causal controller, which indicates that a longer prediction horizon is required to improve performance. Increasing the horizon from $n_p = 3$ to $n_p = 5$ yielded a small energy increase of 2.1%; however, the training converged faster with the $n_p = 5$ case. The highest energy output was obtained with $n_p = 5$, yielding a 13.9% improvement over the baseline causal controller. However, this improvement was obtained at the cost of higher CE, with $n_p = 5$ showing the largest values. Despite this increase, the control inputs remained within the constraint, as confirmed in the previous results.

The variation of the gains K_x and K_w is shown in Figs. 7a and 7b. As can be seen, the gain K_x remains close to its initial value, exhibiting limited variation. In contrast, K_w does not converge to a steady-state solution as in typical adaptive controllers but instead varies over time. This behavior reflects the RL framework, where the agent selects an action to change K_w dynamically at each timestep based on the policy, with the selected value considered optimal for the current state.

To assess the robustness of the proposed control scheme, the trained PPO agent with $n_p = 5$ was evaluated in an uncertain environment without retraining. Here, uncertainty was introduced through test environment model mismatch and noisy wave predictions. Model uncertainty is represented by $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$, where I is an identity matrix. The wave prediction errors are modeled as follows:

$$w_{\mathrm{pred},k}^{\mathrm{noisy}} = w_{\mathrm{pred},k} + \eta_k, \tag{29}$$

where $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$ denotes the prediction noise with varying signal-to-noise ratios. The result of the simulated energy output is shown in Fig. 8. As shown, with wave prediction noise and model uncertainty ranges of 0–20 dB and 0–0.1, respectively, the trained $n_p = 5$ agent still generated energy and remained stable.

5 Conclusion

This paper has proposed a PPO-based noncausal control framework for WECs. By incorporating short-term wave prediction into the RL process, the proposed framework adaptively tunes time-varying gains to enhance energy absorption while respecting motion and actuation limits. Simulation results obtained using real-world sea wave height data demonstrated that longer prediction horizons ($n_p = 3, 5$) improved the energy capture compared with both a causal baseline controller and the $n_p = 1$ case, with the $n_p = 5$ case realizing up to 13.9% higher energy extraction compared with the baseline. In addition, training terminated faster for the $n_p = 5$ case, thereby reducing training costs. The tradeoff was an increase in CE; however, all controllers operated within the

Table 2 CE and energy at the end of the simulation for causal method (conventional PD controller) and noncausal PPO with n_p = 1, 3, 5.

Method	Control effort (kN²·s)	Energy at end of simulation, E_{2000} (kJ)
Causal	808.78	128.72
PPO, noncausal with $n_p = 1$	877.64	128.42
PPO, noncausal with $n_p = 3$	943.93	143.59
PPO, noncausal with $n_p = 5$	1396.40	146.65

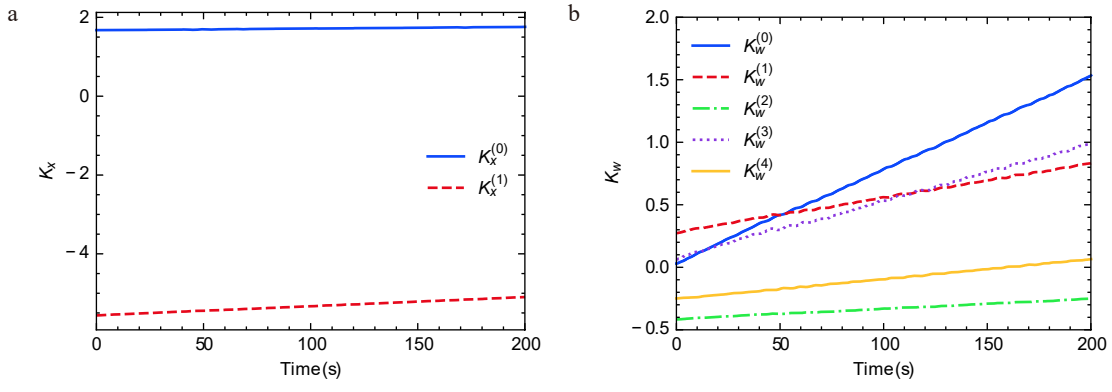

Figure 7 Simulation of trained noncausal PPO agent ($n_p = 5$): (a) K_x and (b) K_w.

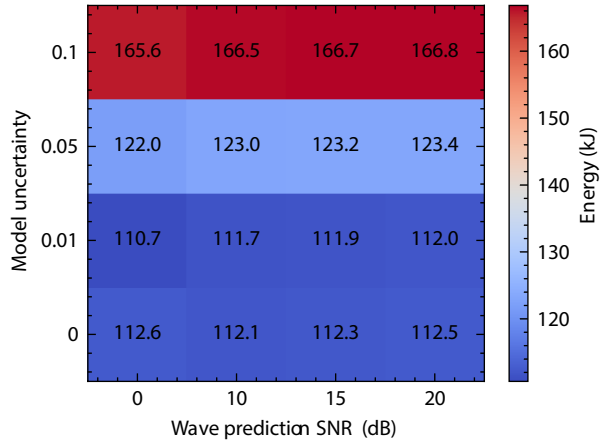

Figure 8 Simulation of trained noncausal PPO agent ($n_p = 5$) against disturbances and wave prediction signal-to-noise ratio.

actuator constraints. In addition, robustness evaluations further demonstrated that the proposed PPO-based noncausal control scheme maintained stable performance under test environment model mismatch and noisy wave predictions. Future work will focus on extending the proposed framework to experimental validation in hardware-in-the-loop environments, which will be performed in a dSPACE environment that can assess the real-time computation and constraints of the proposed controller. Furthermore, electrical power conversion beyond mechanical power will also be considered in future work.

Acknowledgments

None.

Author contribution statement

Yao Zhang: Conceptualization, Methodology, Validation, Writing - Original Draft, Visualization, Supervision, Funding acquisition. Tianyi Zeng: Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. Vincentius Versandy Wijaya: Methodology, Software, Investigation, Writing - Original Draft, Visualization. Yutong Song: Methodology, Software, Investigation, Writing - Original Draft, Visualization. Richard Bucknall: Conceptualization, Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. Stephen Turnock: Conceptualization, Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. All the authors have approved the final manuscript.

Data availability

Data that support the findings have been given in the article.

Declaration of competing interest

All the contributing authors report no conflicts of interest in this work.

Funding

This work is granted by Wave Energy Scotland Direct Generation Competition, the UK Royal Society IEC-NSFC 301 (Grant No. 223485), and Innovate UK GREENPORTSIDE project.

Use of AI statement

None.

References

[1] Falnes, J. (2007). A review of wave-energy extraction. *Mar. Struct.* 20, 185–201.
[2] Ahamed, R., McKee, K., Howard, I. (2020). Advancements of wave energy converters based on power take off (PTO) systems: a review. *Ocean Eng.* 204, 107248.
[3] Drew, B., Plummer, A. R., Sahinkaya, M. N. (2009). A review of wave energy converter technology. *Proc. Inst. Mech. Eng., Part A: J. Power Energy* 223, 887–902.
[4] Budar, K., Falnes, J. (1975). A resonant point absorber of ocean-wave power. *Nature* 256, 478–479.
[5] Kofoed, J. P., Frigaard, P., Friis-Madsen, E., Sørensen, H. C. (2006). Prototype testing of the wave energy converter wave dragon. *Renewable Energy* 31, 181–189.
[6] Falcão, A. F. O., Henriques, J. C. C. (2016). Oscillating-water-column wave energy converters and air turbines: a review. *Renewable Energy* 85, 1391–1424.
[7] Yemm, R., Pizer, D., Retzler, C., Henderson, R. (2012). Pelamis: experience from concept to connection. *Phil. Trans. R. Soc. A.* 370, 365–380.
[8] Smart, G., Noonan, M. (2018). Tidal Stream and Wave Energy Cost Reduction and Industrial Benefit. Glasgow, UK: ORE Catapult.
[9] EPSRC, 2020. Wave Energy Road Map. Technical Report. EPSRC.
[10] Falnes, J. (1980). Radiation impedance matrix and optimum power absorption for interacting oscillators in surface waves. *Appl. Ocean Res.* 2, 75–80.
[11] Hals, J., Falnes, J., Moan, T. (2011). Constrained optimal control of a heaving buoy wave-energy converter. *J. Offshore Mech. Arct. Eng.* 133, 011401.
[12] Babarit, A., Clément, A. H. (2006). Optimal latching control of a wave energy device in regular and irregular waves. *Appl. Ocean Res.* 28, 77–91.
[13] Falcão, A. F. D. O. (2010). Wave energy utilization: a review of the technologies. *Renewable Sustain. Energy Rev.* 14, 899–918.
[14] Clément, A. H., Babarit, A. (2012). Discrete control of resonant wave energy devices. *Phil. Trans. R. Soc. A* 370, 288–314.
[15] Faedo, N., Olaya, S., Ringwood, J. V. (2017). Optimal control, MPC and MPC-like algorithms for wave energy systems: an overview. *IFAC J. Syst. Control* 1, 37–56.
[16] Penalba, M., Giorgi, G., Ringwood, J. V. (2017). Mathematical modelling of wave energy converters: a review of nonlinear approaches. *Renewable Sustain. Energy Rev.* 78, 1188–1207.
[17] Fusco, F., Ringwood, J. V. (2010). Short-term wave forecasting for real-time control of wave energy converters. *IEEE Trans. Sustain. Energy* 1, 99–106.
[18] Duan, W. Y., Han, Y., Huang, L. M., Zhao, B. B., Wang, M. H. (2016). A hybrid EMD-SVR model for the short-term prediction of significant wave height. *Ocean Eng.* 124, 54–73.
[19] Peña-Sanchez, Y., Mérigaud, A., Ringwood, J. V. (2020). Short-term forecasting of sea surface elevation for wave energy applications: the autoregressive model revisited. *IEEE J. Oceanic Eng.* 45, 462–471.
[20] Mahmoodi, K., Nepomuceno, E., Razminia, A. (2022). Wave excitation force forecasting using neural networks. *Energy* 247, 123322.
[21] Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press.
[22] Moretti, G., Herran, M. S., Forehand, D., Alves, M., Jeffrey, H., Vertechy, R., Fontana, M. (2020). Advances in the development of dielectric elastomer generators for wave energy conversion. *Renewable Sustain. Energy Rev.* 117, 109430.
[23] Moretti, G., Rosati Papini, G. P., Daniele, L., Forehand, D., Ingram,

D., Vertechy, R., Fontana, M. (2019). Modelling and testing of a wave energy converter based on dielectric elastomer generators. *Proc. R. Soc. A*. 475, 20180566.

[24] Anderlini, E., Forehand, D. I. M., Stansell, P., Xiao, Q., Abusara, M. (2016). Control of a point absorber using reinforcement learning. *IEEE Trans. Sustain. Energy* 7, 1681–1690.

[25] Anderlini, E., Forehand, D. I. M., Bannon, E., Xiao, Q., Abusara, M. (2018). Reactive control of a two-body point absorber using reinforcement learning. *Ocean Eng.* 148, 650–658.

[26] Anderlini, E., Husain, S., Parker, G. G., Abusara, M., Thomas, G. (2020). Towards real-time reinforcement learning control of a wave energy converter. *J. Mar. Sci. Eng.* 8, 845.

[27] Ye, M., Zhang, C., Ren, Y. R., Liu, Z. Y., Haidn, O. J., Hu, X. Y. (2025). Adaptive optimization of wave energy conversion in oscillatory wave surge converters via SPH simulation and deep reinforcement learning. *Renewable Energy* 246, 122887.

[28] Zou, S. Y., Zhou, X., Weaver, W., Abdelkhalik, O. (2022). Deep reinforcement learning control of wave energy converters. *IFAC-PapersOnLine* 55, 305–310.

[29] Wang, H. Z., Wijaya, V., Zeng, T. Y., Zhang, Y. (2024). Deep reinforcement learning-based non-causal control for wave energy conversion. *Ocean Eng.* 311, 118860.

[30] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint*, arXiv: 1707.06347.

[31] Pasta, E., Faedo, N., Mattiazzo, G., Ringwood, J. V. (2023). Towards data-driven and data-based control of wave energy systems: classification, overview, and critical assessment. *Renewable Sustain. Energy Rev.* 188, 113877.

[32] Belmont, M. R., Christmas, J., Dannenberg, J., Hilmer, T., Duncan, J., Duncan, J. M., Ferrier, B. (2014). An examination of the feasibility of linear deterministic sea wave prediction in multidirectional seas using wave profiling radar: theory, simulation, and sea trials. J. Atmos. *Oceanic Technol.* 31, 1601–1614.

[33] Meng, Z. F., Chen, Z., Khoo, B. C., Zhang, A. M. (2022). Long-time prediction of sea wave trains by LSTM machine learning method. *Ocean Eng.* 262, 112213.

[34] Shi, S., Patton, R. J., Liu, Y. H. (2018). Short-term wave forecasting using Gaussian process for optimal control of wave energy converters. *IFAC-PapersOnLine* 51, 44–49.

[35] Ringwood, J. V., Bacelli, G., Fusco, F. (2014). Energy-maximizing control of wave-energy converters: the development of control system technology to optimize their operation. *IEEE Control Syst. Mag.* 34, 30–55.

[36] Genest, R., Ringwood, J. V. (2016). A critical comparison of model-predictive and pseudospectral control for wave energy devices. *J. Ocean Eng. Mar. Energy* 2, 485–499.

[37] Hals, J., Falnes, J., Moan, T. (2011). A comparison of selected strategies for adaptive control of wave energy converters. *J. Offshore Mech. Arct. Eng.* 133, 031101.

[38] Cretel, J., Lewis, A. W., Lightbody, G., Thomas, G. P. (2010). An application of model predictive control to a wave energy point absorber. *IFAC Proc.* 43, 267–272.

[39] Kracht, P., Perez-Becker, S., Richard, J. B., Fischer, B. (2015). Performance improvement of a point absorber wave energy converter by application of an observer-based control: results from wave tank testing. *IEEE Trans. Ind. Appl.* 51, 3426–3434.

[40] Richter, M., Magana, M. E., Sawodny, O., Brekken, T. K. A. (2013). Nonlinear model predictive control of a point absorber wave energy converter. *IEEE Trans. Sustain. Energy* 4, 118–126.

[41] Jia, Y. B., Meng, K., Dong, L., Liu, T. M., Sun, C. Y., Dong, Z. Y. (2021). Economic model predictive control of a point absorber wave energy converter. *IEEE Trans. Sustain. Energy* 12, 578–586.

[42] Gu, Y. F., Ding, B. Y., Sergiienko, N. Y., Cazzolato, B. S. (2021). Power maximising control of a heaving point absorber wave energy converter. *IET Renewable Power Gen.* 15, 3296–3308.

[43] Zhang, Y., Li, G. (2020). Non-causal linear optimal control of wave energy converters with enhanced robustness by sliding mode control. *IEEE Trans. Sustain. Energy* 11, 2201–2209.

[44] Zhan, S. Y., Li, G. (2019). Linear optimal noncausal control of wave energy converters. *IEEE Trans. Control Syst. Technol.* 27, 1526–1536.

[45] Weiss, G., Li, G., Mueller, M., Townley, S., Belmont, M. R. (2012). Optimal control of wave energy converters using deterministic sea wave prediction. In Fuelling the Future: Advances in Science and Technologies for Energy Generation, Transmission and Storage. Mendez-Vilas, A., Ed. London: Brown-Walker Press, p 396–400.

[46] Yu, Z., Falnes, J. (1995). State-space modelling of a vertical cylinder in heave. *Appl. Ocean Res.* 17, 265–275.

[47] Lee, C. H. (1995). WAMIT Theory Manual. Cambridge: Massachusetts Institute of Technology, Department of Ocean Engineering.

[48] Kurnia, R., Ducrozet, G. (2023). NEMOH: open-source boundary element solver for computation of first-and second-order hydrodynamic loads in the frequency domain. *Comput. Phys. Commun.* 292, 108885.

[49] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N. (2021). Stable-baselines3: reliable reinforcement learning implementations. *J. Mach. Learn. Res.* 22, 1–8.

[50] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, USA, p 2623–2631.

[51] Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. (2020). What matters in on-policy reinforcement learning? A large-scale empirical study. *arXiv preprint*, arXiv: 2006.05990.

[52] Kingma, D. P., Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint*, arXiv: 1412.6980.