Model-Free Linear Noncausal Optimal Control of Wave Energy Converters via Reinforcement Learning

Siyuan Zhan, *Member, IEEE* and John V. Ringwood, *Fellow, IEEE*

***Abstract*— This article introduces a novel reinforcement learning (RL) method for wave energy converters (WECs), which directly generates linear noncausal optimal control (LNOC) policies on continuous action space. Unlike other existing WEC RL algorithms looking at the problem mainly from a learning perspective, the proposed RL approach adopts a control-theoretic approach by delving into the underlying WEC energy maximization (EM) optimal control problem (OCP). This leads to control-informed decisions on choosing the RL state, as well as developing the RL structure. The proposed model-free LNOC (MF-LNOC) offers substantial advantages, including significantly improved performance due to the use of noncausal information, a simplified RL with linear actor and quadratic critic structures, and remarkable fast convergence speeds, achieved using less than 150 s of data points, for a benchmarked point absorber, which can be further shortened using the replay technique. This reduction in training time allows for controller reconfiguration in pace with sea changes. Demonstrative numerical simulations are presented to verify the efficacy of the proposed methods. The proposed MF-LNOC also shows robustness against wave prediction inaccuracies and changing sea conditions. The MF-LNOC methodology can be highly attractive for WEC developers who want to design an efficient and reliable controller for WECs but also hope to avoid the challenge of establishing a control-oriented model that can preserve high fidelity over a wide range of sea conditions.**

***Index Terms*— Optimal control, reinforcement learning (RL), wave energy converter (WEC), wave prediction.**

I. Introduction

WAVE energy has a significant potential in supplying renewable energy to complement other renewable sources. However, the current levalised cost for wave energy is significantly higher than the other renewable (and conventional) sources, and it is well-known that a reliable and efficient wave energy converter (WEC) controller can reduce the unit cost of wave energy [1]. Early WEC control methods based on the impedance-matching principle are challenging to implement in the actual sea conditions which include a wide range of wave frequencies [2]. Optimal control provides a natural mechanism to maximize energy conversion from waves and has attracted significant research attention [3], [4], [5].

The WEC energy maximization (EM) optimal control problem (OCP) is essentially different from conventional OCPs in three respects. First, a WEC EM OCP aims to maximize the power conversion rate, represented by a product of the power take-off (PTO) force and the device relative velocity, leading to an indefinite stage cost, rather than the positive-definite stage cost for conventional tracking and regulating OCPs. Second, the impact of wave excitation, treated as a disturbance input in the WEC EM-OCP, is *beneficial* for the EM control objective. Therefore, the disturbance handling principle for a convention OCP cannot be adopted here because, in the WEC EM OCP, kinetic energy transferred from ocean waves needs to be captured/enhanced rather than attenuated. Third, the WEC EM-OCP is essentially a noncausal control problem; that is, the current optimal control action depends on the future wave information. Using wave the prediction, the energy conversion rate can be significantly increased, even doubled in some sea states [2], [6].

Based on this principle, many optimal control methods have been developed for WECs, with comprehensive reviews available in [1] and [3]. Targeting WECs with nonlinear dynamics, many online optimal control algorithms have been developed, including those based on dynamic programming [6], Prontrayin's minimality principle [4], spectral method [7], pseudospectral method [8], [9], and moment matching [10]. Despite being able to handle nonlinearities, those algorithms can sometimes be difficult to design and implement in real time. On the other hand, when a linear model can adequately describe WEC dynamics, the noncausal WEC EM OCP can be solved with a closed-form analytic control policy, referred to as the linear noncausal optimal control (LNOC), which consists of a causal feedback part and an anticausal feedforward part to incorporate wave prediction. Since the control coefficients of LNOC can be calculated offline, the implementation of which does not require online computation, developing representative linear WEC models for WEC LNOC becomes increasingly attractive. Some successful initial application examples of LNOC have been reported in [11], while an example of linear representative modeling is reported in [12].

Despite the advantages mentioned above, establishing a linear control-oriented model that can represent the WEC dynamics with adequate accuracy can be challenging for some WECs, since a linear WEC model described by (the physical)

Manuscript received 12 December 2023; revised 26 April 2024; accepted 8 May 2024. This work was supported in part by the SFI Frontiers for the Future Programme - Award under Grant 21/FFP-A/8997, in part by the Science Foundation Ireland for the Marine Renewable Ireland (MaREI) Centre under Grant 12/RC/2302_P2, and in part by the Royal Irish Academy Charlemont Sciences Grant. Recommended by Associate Editor A. Mesbah. *(Corresponding author: Siyuan Zhan.)*

Siyuan Zhan is with the School of Engineering, Trinity College Dublin, Dublin, D02 PN40 Ireland (e-mail: zhans@tcd.ie).

John V. Ringwood is with the Centre for Ocean Energy Research, Maynooth University, Maynooth, W23 F2H6 Ireland (e-mail: john.ringwood@mu.ie).

Digital Object Identifier 10.1109/TCST.2024.3401863

Cummins' equation [13], is usually obtained around the predefined equilibrium point with a small movement assumption. With a well-designed WEC controller, e.g., LNOC, large oscillations, to maximize energy output, result, particularly for those sea conditions containing more exploitable energy. This, paradoxically [14] invalidates the small movement assumption and, therefore, leads to a substantial deterioration in model fidelity. A WEC LNOC design based on such an inaccurate model can lead to significant performance degradation [15], [16]. Other aspects that could potentially lead to a change in WEC dynamics include: 1) changing sea state; 2) varying mooring dynamics due to slow drift motions of the WEC; 3) changing tidal elevation [17]; 4) marine growth on the WEC [18]; and 5) water leakage into the WEC and other noncritical subsystem partial failure [19].

To cope with model inaccuracy problems, WEC research engineers have developed adaptive mechanisms, which can automatically correct a WEC linear model, using measured data from sensors, to remain representative of the actual WEC dynamic behavior [15], [18]. In [20], an adaptive model-correcting mechanism is developed to update a representative WEC linear model based on which an LNOC is developed. However, a significant critical drawback of the cited adaptive methods is the lack of a passivity guarantee, a feature not only reflecting the principle that a WEC can not only generate more energy than it receives but also ensuring the existence, convergence, and stability of the resultant LNOC. Failure to enforce passivity in the model update process can lead to a loss of control stability, resulting in potential catastrophic failure for WEC operation.

An alternative way of dealing with model change is to generate control coefficients directly from real-time data. Reinforcement learning (RL), as a class of machine learning methods, has demonstrated its efficacy across a wide range of disciplines in finding optimal actions in uncertain environments [21], [22]. Despite significant success in other applications, applying RL to solve the WEC EM-OCP is less straightforward. First, the continuous control signal invalidates the direct application of those RL algorithms well-developed for discrete-action-space problems. To resolve this problem, pioneering WEC RL algorithms adopted a suboptimal approach, involving the use of a prefixed causal suboptimal feedback structure. The WEC EM-OCP of finding the continuous optimal control signal is therefore reformulated into finding the optimal feedback coefficients, which can be discretized in the action space. For example, in [19], an RL-based method is developed for WEC controllers to adapt the damping coefficient of a WEC resistive controller using input-output data. Subsequently, Anderlini et al. [23] develop an automatic data-driven reactive control tuning mechanism. As a complementary result, Q-learning methods, with tabular and function approximators, are benchmarked in [24].

Meanwhile, inspired by the recent development of RL algorithms for continuous action spaces, such as deterministic deep policy gradient (DDPG) [25], twin delayed DDPG (TD3) [26], and soft actor–critic (SAC) [27], WEC researchers attempt to develop real-time algorithms directly targeting the WEC EM-OCP without being restricted by the prefixed causal suboptimal feedback structures. In [28], featuring the development of a model-based MPC based on a linearized model, an RL algorithm is developed using the SAC structure, which shows its potential for improving energy performance in some sea states, compared with the MPC based on an inaccurate linearized model. Nevertheless, the SAC-based RL method adopted in [28], similar to other actor–critic (AC) methods, needs to use separate neural network (NN) function approximates to represent the actor and the critic, respectively. This feature significantly increases the required data points to train the AC structure. As a compromise, in [28], modeling efforts are required to initialize a reference model-based MPC to accelerate training. In [29], to improve the convergence performance of the SAC developed in [28], the authors adopted the Bayesian policy gradient with the AC framework, referred to as "BAC." The BAC developed in [29] achieves a performance close to the optimized feedback $u(t) = Fz(t) + Gv(t)$, with training time reduced from approximately 8.4, for SAC in [28], to 1.5 h.

Meanwhile, the OCP for linear systems with convex quadratic cost functions, often known as "linear quadratic regulators (LQRs)," has some favorable unique features, such as the existence of analytic linear feedback optimal control policies and analytic quadratic value functions. Those favorable features are preserved in the corresponding model-free LQR (MF-LQR) problems. In [30], a policy-iterative RL algorithm is developed, which shows that: 1) the Q-function is quadratically dependent on an augmented vector consisting of the state and control action; 2) the resultant control action is linearly dependent on the state; 3) only $(n_x + n_u)(n_x + n_u + 1)/2$ parameters need to be estimated in the training process, with n_x and n_u being the orders of the system state and control input, respectively; and 4) theoretical guarantees on convergence and stability can be established. Nevertheless, those approaches focusing on causal LQRs cannot be directly translated into the noncausal WEC control problem of maximizing energy output.

The presence of a persistent disturbance is another nontrivial problem in a learning-based control system since the disturbance-induced bias, when unaccounted in the design process, may prevent the convergence of control parameters to their optimal values [31], [32]. To resolve this problem, several approaches have been proposed. For example, Jiang and Jiang [33] integrate a learning-based approach with conventional disturbance attenuation tools, such as back-stepping and sliding mode control. In [34], the disturbance attenuation control problem is considered in an $\mathcal{H}_\infty$ structure. After reformulating the $\mathcal{H}_\infty$ problem into a zero-sum game structure, a data-driven RL method can be developed to solve the associated Hamilton–Jacobi–Isaacs (HJI) functions. In [32], by actively compensating for the disturbance-induced bias, robust optimal tracking is achieved for systems subject to an $\mathcal{L}_2$ disturbance. However, those approaches cannot solve the WEC EM-OCP of maximizing the benefit of disturbance utilization in optimizing the predefined performance index.

Therefore, despite some achievements made, the existing RL-based WEC control methods cannot optimally solve the panchromatic WEC EM control problem, which is inherently noncausal and use wave prediction information to improve performance. Furthermore, the ability to deal with slow variations in sea state, which define the spectral shape of the wave excitation, is also important.

Targeting those limitations, in this article, we present a novel model-free method to formulate LNOC (MF-LNOC) in a continuous action space, by developing a control-theoretic approach. Rather than solely relying on the learning techniques, as in SAC and BAC, to achieve convergence, MF-LNOC based on the control-theoretic approach directly investigates the underlying WEC EM-OCP, which gives the informed decision on selecting the state and the structure of RL. We show that, with the novel formulation, features such as fixed linear actor and quadratic critic structures, and guarantees of convergence and stability, are preserved from MF-LQR to MF-LNOC.

Solving the WEC control problem using the proposed MF-LNOC has the following advantages.

1) *Significantly Improved Performance:* The proposed MF-LNOC solves the noncausal WEC EM-OCP, without being limited to any causal suboptimal structure. By directly incorporating future wave excitation in the RL, the MF-LNOC shows significant performance improvement over even the best-tuned causal linear feedback controller, which is regarded as the performance limit for existing WEC RL methods [19], [23], [28], [29].
2) *Simplified RL Structure:* By investigating the underlying control problem, MF-LNOC leads to control-informed selections of the RL structures, which, similar to the MF-LQR case, also have a prefixed linear actor and a quadratic critic. This avoids the need for multiple NNs for the actor and the critic as in [26], [28], and [29].
3) *Fast Convergence Speed:* In the proposed MF-LNOC, only a total of $N(N+1)/2$ parameters in the quadratic actor need to be determined from the linear least squares (LLSs) solution during the training process, where $N := n_x + n_u + n_p$, with n_x, n_u, and n_p being the orders of the system state and input, and the wave prediction horizon, respectively. Even without using the replay technique, the off-policy training process can achieve convergence within less than 150 s of exploration, thanks to the simplified structure, and the ensured convergence. This enables the MF-LNOC to update the RL parameters in pace with sea changes for the WEC to remain in its best performance.

In summary, the proposed MF-LNOC enables WEC developers to formulate a simple, computationally efficient, and control law, with guaranteed convergence and stability without the need for the first principles of WEC modeling, which is known to be challenging for some WEC designs. This also allows the tracking of model variations across sea states, in the sense of "representative linear models" [12], [35]. It will be shown that the convergence rate of the learning algorithm proposed in this article is of the order of 2–3 min, which is entirely adequate to deal with sea state changes which typically occur over the course of 20 min or more. This computational efficiency is crucial, compared with existing learning methods applied to wave energy control which take, at minimum, 1.5 h.

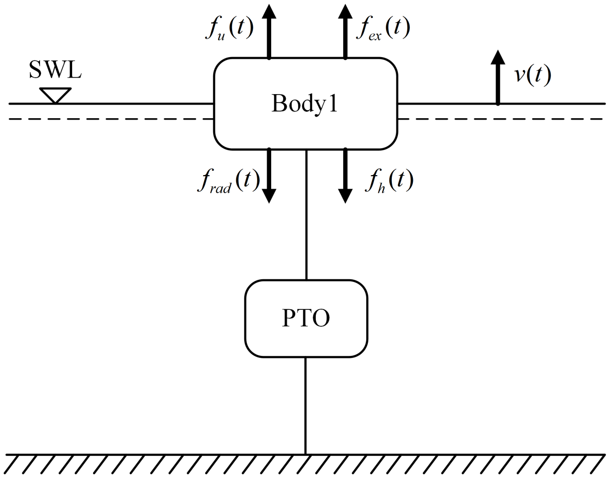


Fig. 1. Dynamic diagram of the float (SWL: still wave level and PTO: power take-off unit).

The remainder of this article is organized as follows. Section II presents the preliminaries, including WEC dynamics, WEC control problem formulation, and WEC RL backgrounds. Section III presents the main result of the MF-LNOC based on the control-theoretic approach. Some real-time implementation issues are discussed in Section IV. Numerical examples are given in Section V, and finally, this article is concluded in Section VI.

Notation: Let $\mathbb{R}^n$ and $\mathbb{R}^{a\times b}$ be the space of all real n-dimensional vectors, and all $a \times b$ dimensional matrices, respectively; $\mathbb{N}_{a:b}$ and $\mathbb{N}_{\geq a}$ denote a set of integers from a to b and greater than or equal to a, respectively. For column vectors z_1 and z_2, $[z_1, z_2]$ denotes a column vector $[z_1^{\mathrm{T}}\ z_2^{\mathrm{T}}]^{\mathrm{T}}$. $\boldsymbol{z}_{a:b} := [z_a, z_{a+1}, \ldots, z_b]$. An element with subscript $i|k$ represents horizon i, predicted/estimated/calculated at time step k. $\boldsymbol{w}_k^{n_p}$ denotes an n_p-step sequence $\boldsymbol{w}_{k:k+n_p-1}$, obtained at time step k. I_n denotes the $n \times n$ identity matrix. $0_{a\times b}$ denotes an $a \times b$ matrix composed entirely of zero entries. "s.t." is the abbreviation of "subject to." "w.r.t." is the abbreviation of "with respect to."

II. Preliminaries

A. WEC Dynamics

In this article, we consider a point absorber-type WEC, restricted in heave motion only, whose dynamic diagram is shown in Fig. 1.

From Newton's law, the motion of the float can be described by

$$M\dot{v}(t) = -f_h(t) - f_{\mathrm{rad}}(t) + f_{\mathrm{ex}}(t) + f_u(t) \qquad (1)$$

where M is the float mass; $f_u(t)$ is the PTO manipulable force; and $v(t)$ is the heave velocity. In (1), the hydrostatic restoring force $f_h(t)$ is modeled by

$$f_h(t) = k_h z(t) \qquad (2)$$

where $z(t)$ denotes the heave displacement, and hydrostatic stiffness $k_h = \rho g S_w$, where ρ, g, and S_w denote the water density, gravitational acceleration, and cross-sectional area of the buoy, respectively. The radiation force $f_{\text{rad}}(t)$ models the frequency-dependent damping effect due to the radiated waves produced by the buoy motion. Using the standard assumptions associated with linear potential theory [36], the radiation force can be modeled by a linear convolution of the radiation impulse response $h_r(t)$ and heave velocity $v(t)$ as follows:

$$f_{\text{rad}}(t) = \int_{-\infty}^{t} h_r(\tau)v(t-\tau)\, d\tau + \mu_\infty \dot{v}(t) \tag{3}$$

where μ_∞ is the added mass asymptote at infinite frequency. $h_r(t)$ and μ_∞ can be calculated via hydrodynamic codes, such as NEMOH [37]. The wave excitation force $f_{\text{ex}}(t)$, corresponding to the force experienced by the body due to wave action, is treated as a predictable additive disturbance. Please refer to [38] for a comprehensive review of methods to calculate, estimate, and/or predict the excitation force over a short-term interval. With (2) and (3), the dynamic equation (1) results in Cummins' [13] equation

$$\begin{aligned}(m+\mu_\infty)\dot{v} &= f_{\text{ex}}(t) - k_h z(t)\\ &\quad - \int_{-\infty}^{t} h_r(\tau)v(t-\tau)\, d\tau + f_u(t).\end{aligned} \tag{4}$$

To develop a control-oriented model, we use a state-space model with minimal realization to approximate the convolution term in (3) via

$$\begin{cases} \dot{x}_r(t) = A_r x_r(t) + B_r v(t) \\ y_r(t) = C_r x_r(t) + D_r v(t) \approx \displaystyle\int_{-\infty}^{t} h_r(\tau)v(t-\tau)\, d\tau \end{cases} \tag{5}$$

where $(A_r,\ B_r\ C_r,\ D_r)$ and $x_r \in \mathbb{R}^{n_r}$ are the state-space matrices and the associated state vector, respectively.

Defining the overall system state vector $x(t) := [z(t), v(t), x_r(t)]$, the control input $u(t) := f_u(t)$, and disturbance input $w(t) := f_{\text{ex}}(t)$, the WEC dynamics described by Cummins' equation (4) can be modeled by the following linear time-variant (LTI) system:

$$\dot{x}(t) = A_c x(t) + B_{\text{wc}} w(t) + B_{\text{uc}} u(t) \tag{6}$$

with coefficients

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ -\dfrac{k_h}{m} & -\dfrac{D_r}{m} & -\dfrac{C_r}{m} \\ 0 & B_r & A_r \end{bmatrix}, \quad B_{\text{wc}} = B_{\text{uc}} = \begin{bmatrix} 0 \\ \dfrac{1}{m} \\ 0 \end{bmatrix}$$
$$C_z = \begin{bmatrix} 1 & 0 & 0_{1\times n_r} \end{bmatrix}, \quad C_v = \begin{bmatrix} 0 & 1 & 0_{1\times n_r} \end{bmatrix}$$

and $m := M + m_\infty$.

B. WEC Optimal Control Problem Setup

The WEC EM OCP aims to maximize the energy converted in the PTO, $\int p(t)dt$, where $p(t)$ is the instantaneous power produced, with the consideration of quadratic PTO losses

$$p(x(t), u(t)) := -v(t)u(t) - ru^2(t). \tag{7}$$

In (7), the first term $-v(t)u(t)$ is converted energy, while the second term $-ru^2(t)$ represents the energy losses, with $r > 0$.

Assumption 1 (Passivity): The WEC model (6) is passive with respect to a virtual performance output $y_p(t)$, defined such that $u(t)y_p(t) = -p(x(t), u(t))$.

We assume that Assumption 1 holds throughout this article. In fact, passivity is a property characteristic of all WECs, since a WEC cannot generate more energy than it absorbs.

To facilitate the design of a linear noncausal optimal controller (LNOC), the continuous-time control-oriented model is discretized via a zero-order hold (ZOH) equivalent, with sampling period t_s, which leads to

$$x_{k+1} = Ax_k + B_u u_k + B_w w_k \tag{8}$$

where $(A,\ B_u,\ B_w)$ are the corresponding discrete-time state-space matrices and w_k is the instantaneous value of excitation force at time step k, $w_k = f_{\text{ex}}(kt_s)$. The energy captured, using the ZOH convention, i.e., $u(t) = u_k$ for $kt_s \leq t < (k+1)t_s$, for a single time interval starting at k, can be expressed by

$$\begin{aligned} e_k &= \int_{kt_s}^{(k+1)t_s} p(x(\tau), u(\tau))\, d\tau \\ &= -u_k C_z (x_{k+1} - x_k) - t_s r u_k^2 \end{aligned} \tag{9}$$

where $C_z := [1 \quad 0 \quad \cdots \quad 0] \in \mathbb{R}^{1\times n_x}$. The cumulative energy converted for a time period from 0 to k is $E_k := \sum_{i=0}^{k} e_i$.

C. RL Background

In an RL framework, an agent, which is in a state s_t at time t, interacts with the environment modeled by a Markovian decision process. By taking an admissible action $a_t \in \mathcal{A}$, where $\mathcal{A}$ denotes the action space of the state (in the learning sense), s_t transitions to the state of the next time instant s_{t+1}, with a reward observed as r_{t+1}. The "reinforcement" philosophy is reflected in evaluating the value functions of taking a_t, defined as the accumulated and weighted future rewards of taking action a_t. Subsequently, the best action a_t is selected by evaluating the values associated with taking action a_t.

Depending on the nature of action space $\mathcal{A}$, RL problems can be categorized into RL with a continuous action space and RL with a discrete action space, where the former often results in a more complex structure because the evaluation of policy has to be performance over infinitely many actions.

Motivated by this factor, most of the existing WEC RL algorithms use a prefixed suboptimal causal feedback control structure, e.g., $u(t) = Gv(t)$ in [19] and [24] or $u(t) = Fz(t) + Gv(t)$ in [28] and [39]. In this way, the problem of finding the continuous action $u(t)$ is converted into finding the optimal feedback parameters F and G. By discretizing the step changes in F and G as the action space, i.e., ∇F and ∇G, respectively, an RL problem with discrete action space can be formulated, where action-value RL methods, such as tabular Q-learning, Q-learning with function approximation, and deep Q-Networks (DQN) can be adopted. To prevent the fast-changing feed gains, in [19], [24], and [28], the parameters are learnt and updated on sea stage changes, and performance is evaluated based on the average power generated in the particular sea states. Nevertheless, the control policy updates can only be done after the sea state change is settled. Meanwhile, in [39], the parameter updates are performed on a wave-by-wave time scale to accelerate the training and policy update process. However, rapid fluctuations occur in the feedback gains, indicating a significantly less stable Q-Network, compared with [19], [24], and [28], which could limit the application of those algorithms on real WEC devices. This phenomenon is partially due to not having modeled the impact of wave excitation force in the sea states, either explicitly or implicitly, which has a substantial influence on the WEC dynamics.

Continuous-action-space RL algorithms have the potential to avoid the problem of being restricted by a suboptimal causal feedback structure. Unlike discrete-action-space counterparts, based on state–action value functions, most continuous-action-space RL algorithms adopt an AC structure, where optimal policies are obtained by searching over policy gradient estimates [40]. The application to WEC control problems starts from [28], where MPC based on an inaccurately linearized model is used to reduce the training time to 8.4 h. Later in [29], the result is improved by considering the Bayesian gradient, which further reduces the training time to 1.5 h. However, despite not using an explicitly suboptimal feedback structure, the best achievable performance of those two AC methods does not exceed the feedback policy $u(t) = Fz(t) + Gv(t)$, tuned with optimum parameters F and G via trial and error, for the particular wave segments. This is because, in both [28] and [29], the RL state contains only causal information, i.e., the state $:= [z(t), v(t), f_{\text{ex}}(t), \dot{f}_{\text{ex}}(t)]$ and cannot exploit wave prediction to form a noncausal optimal control law.

Given the disadvantages of the restricted solution space associated with discrete-action-space methods, we pursue a continuous-action-space framework, with a prefixed structure requiring only $(N+1)N/2$ variables, to be parametrized in the training process, by extending the results of MF-LQR to MF-LNOC for EM. Here, N is the dimension of the continuous (learning) state

$$N = \begin{cases} n_x + n_u, & \text{for MF-LQR} \\ n_x + n_u + n_p, & \text{for MF-LNOC} \end{cases} \tag{10}$$

where n_x, n_u, and n_p denoting the dimensions of the system state and input, and the wave prediction horizon, respectively. This control-theoretic study also offers insights into fundamental issues, which not only provides guarantees of convergence and stability but also establishes some level of interpretability, while avoiding the use of a purely black-box structure, with the associated opaqueness and uncertainty of convergence/stability.

III. Main Results

In this section, first, we present a feedback control reformulation of the WEC LNOC in Section III-A, assuming that the model is known. Using a policy iterative method, the LNOC formulated in the feedback form is shown to be equivalent to the conventional LNOC [2] in the feedforward form. Next, in Section III-B, a model-free policy iterative framework is developed upon the result in Section III-A.

A. WEC Control Reformulation

Recall our previous results on the model-based formulation [2]. The optimal control of the LNOC has a closed-form analytic form solution

$$u_k^* = K_x x_k + K_d \boldsymbol{w}_k \tag{11}$$

where K_x and K_d are the feedback and feedforward coefficients that can be determined in [2, Algorithm 2]. However, it is more straightforward to develop an RL formulation for control in feedback form. To resolve this problem, we reformulate the control problem setup using a feedback representation based on an augmented state, consisting of the systems state x and the prediction of excitation force w.

Similar to the implementation principle of WEC MPC, the proposed WEC EM control is designed and implemented following a *receding horizon* manner. Here, we use the subscript $i|k$ to denote the state or input at horizon i, predicted/estimated at time k. Assuming the availability of a prediction of the wave excitation force w_k for $k \in \mathbb{N}_{0:n_p-1}$, we define the following control problem to be solved, recursively, at each time step k

$$\begin{aligned} \mathcal{P}_k^{n_p}\colon \ & \inf_{\boldsymbol{u}} \sum_{k=0}^{\infty} L\big(x_{i|k}, x_{i+1|k}, u_{i|k}\big) \\ & \text{s.t. } x_{i+1|k} = A x_{i|k} + B_u u_{i|k} + B_w w_{k+i}, \quad i \in \mathbb{N}_{0:n_p-1} \\ & \phantom{\text{s.t. }} x_{i+1|k} = A x_{i|k} + B_u u_{i|k}, \quad i \in \mathbb{N}_{\geq n_p} \\ & \phantom{\text{s.t. }} x_{0|k} = x_k \end{aligned} \tag{12}$$

and apply the first element of the optimal solution as the control sequence, i.e., $u_k = u_{0|k}^*$. Here, the superscripts and subscripts of $\mathcal{P}$ represent the wave prediction length n_p, and the time instant k, respectively; the stage cost L is defined as the negative of e_k in (9), i.e.,

$$L\big(x_{i|k}, x_{i+1|k}, u_{i|k}\big) := u_{i,k} C_z \big(x_{i+1|k} - x_{i|k}\big) + t_s r u_{i,k}^2 \tag{13}$$

such that minimizing the cost function L corresponds to maximizing the accumulated converted energy; the resultant value function of $\mathcal{P}_k^{n_p}$, i.e., the optimal cost function, is denoted by $\mathcal{V}_k^{n_p}$.

Ideally, since WEC control does not have a natural termination time, we would like to investigate $\mathcal{P}_k^{n_p}$ as $n \to \infty$. However, this infinite-horizon ideal case is not achievable due to the dependence on accurate infinite-horizon wave prediction, while most existing wave prediction techniques can only provide wave prediction within a limited horizon (10–20 s) with acceptable accuracy. With only a limited look-ahead horizon for wave excitation w_k being available for WEC control, solving $\mathcal{P}_k^{n_p}$ becomes reasonable and, in the authors' previous result in [41], the solution of $\mathcal{P}_k^{n_p}$ provides the best approximation to the ideal intractable case.

With Assumption 1 and by partitioning the state and input trajectories into $\boldsymbol{x}_{0:\infty} = [\boldsymbol{x}_{0:n-1}, \boldsymbol{x}_{n:\infty}]$ and $\boldsymbol{u}_{0:\infty} = [\boldsymbol{u}_{0:n-1}, \boldsymbol{u}_{n:\infty}]$, we have that the value function $\mathcal{V}_k^{n_p}(x_k, \boldsymbol{w}_k^{n_p})$ is bounded. From an application perspective, the boundedness of $\mathcal{V}_k^{n_p}$ can also be verified by observing that $w_{k+i} = 0$ is assumed for $i \in \mathbb{N}_{\geq n_p}$ in $\mathcal{P}_k^{n_p}$, implying a WEC absorbing finite energy from waves can only generate finite usable energy (e.g., electricity).

Next, we investigate the solution of $\mathcal{P}$, predicted/calculated at time instant k. Define the augmented predicted state

$$\begin{aligned} X_{0|k} &:= \left[x_{0|k}, w_k, w_{k+1}, \ldots, w_{k+n_p-1}\right] \in \mathbb{R}^{n_x+n_p} \\ X_{1|k} &:= \left[x_{1|k}, w_{k+1}, \ldots, w_{k+n_p-1}, 0\right] \in \mathbb{R}^{n_x+n_p} \\ &\vdots \\ X_{n_p-1|k} &:= \left[x_{n_p-1|k}, w_{k+n_p-1}, 0, \ldots, 0\right] \in \mathbb{R}^{n_x+n_p} \\ X_{i|k} &:= \left[x_{i|k}, 0, \ldots, 0\right] \in \mathbb{R}^{n_x+n_p} \quad \text{for all } i \geq n_p. \end{aligned}$$

The associated dynamics in (12), using the defined augmented state, can be equivalently written as follows:

$$X_{i+1|k} = \boldsymbol{A} X_{i|k} + \boldsymbol{B} u_{i|k}, \quad i \in \mathbb{N}_{\geq 0} \tag{14}$$

with coefficients

$$\boldsymbol{A} := \begin{bmatrix} A & B_w D \\ 0 & \mathrm{T} \end{bmatrix}, \quad \boldsymbol{B} := \begin{bmatrix} B \\ 0 \end{bmatrix}.$$

Here, $D := [1 \quad 0 \quad \cdots \quad 0] \in \mathbb{R}^{1\times n_p}$, and $M \in \mathbb{R}^{n_p\times n_p}$ is a matrix such that, for a sequence $\boldsymbol{w}_k^{n_p} = [w_k, w_{k+1}, \ldots, w_{k+n_p-1}]$ and $M\boldsymbol{w}_k^{n_p} = [w_{k+1}, \ldots, w_{k+n_p-1}, 0]$, i.e., $\mathrm{T} := \left[\begin{smallmatrix} 0 & I_{n_p-1} \\ 0 & 0 \end{smallmatrix}\right]$.

Using the augmented state $X_i|k$, defined in (14), the stage cost $L(x_{i|k}, x_{i+1|k}, u_{i|k})$ can be equivalently written as follows:

$$L\left(X_{i|k}, u_{i|k}\right) := u_{i|k} C_X X_{i|k} + (1/2) R u_{i|k}^2 \tag{15}$$

where $C_X := \left[\begin{smallmatrix} C_z(A-I) \\ C_z B_w D \end{smallmatrix}\right]$ and $R := 2t_s r + 2C_z B_u$. Defining augmented states $X_{i|k}$, and using (14) and (15), the control problem $\mathcal{P}_k^{n_p}$ can be equivalently expressed as follows:

$$\begin{aligned} \mathcal{P}_k^{n_p}\colon \ &\inf_{\boldsymbol{u}} \sum_{k=0}^{\infty} L\left(X_{i|k}, u_{i|k}\right) \\ &\text{s.t. } X_{i+1|k} = \boldsymbol{A} X_{i|k} + \boldsymbol{B} u_{i|k} \\ &\phantom{\text{s.t. }} X_{0|k} = X_k. \end{aligned} \tag{16}$$

Lemma 1 (Preservation of Passivity): Suppose Assumption 1 holds. The following augmented dynamics

$$\boldsymbol{X}_{k+1} = \boldsymbol{A} X_k + \boldsymbol{B} u_k$$

are strictly passive w.r.t. a virtual performance output $Y_k \in \mathbb{R}$ defined such that $u_k Y_k = L(X_k, u_k)$.

Proof: See the Appendix. ■

With guaranteed passivity from Lemma 1, we can guarantee the existence and uniqueness of solution for $\mathcal{P}_k^{n_p}$ in the following theorem.

Theorem 1 [42]: The optimal control action at time k, computed from OCP $\mathcal{P}_k^{n_p}$, with a receding horizon implementation, takes the form of

$$u_k = u_{0|k}^* = F X_{0|k} = F X_k \tag{17}$$

where the feedback coefficient

$$F := -\left(R + \boldsymbol{B}^T H \boldsymbol{B}\right)^{-1}\left(C_X + \boldsymbol{B}^T H \boldsymbol{A}\right) \tag{18}$$

with $H \in \mathbb{R}^{(n_x+n_p)\times(n_x+n_p)}$ is the unique and stabilizing solution of the discrete-time algebraic Riccati equation (DARE)

$$H = \boldsymbol{A}^T H \boldsymbol{A} - \left(C_X + \boldsymbol{B}^T H \boldsymbol{A}\right)^{\mathrm{T}}\left(R + \boldsymbol{B}^T H \boldsymbol{B}\right)^{-1} \times \left(C_X + \boldsymbol{B}^T H \boldsymbol{A}\right) \tag{19}$$

Algorithm 1 Offline Model-Based PI to Solve DARE (19) and to Calculate F

```
Initialise $j = 0$, $F^0 := [0 \ \ -f_d \ \ 0 \ \ \ldots \ \ 0]$;
while $H^i$ and $F^i$ are not convergent do
    Policy evaluation: Solve for $H^{i+1}$ using
        $H^{j+1} = (\boldsymbol{A} + \boldsymbol{B}F^j)^T H^j \, (\boldsymbol{A} + \boldsymbol{B}F^j) + {F^j}^T R F^j + 2{F^j}^T C_X$;
    Policy improvement: Solve for $F^{i+1}$ using
        $F^{i+1} = -(R + \boldsymbol{B}^T H^{i+1} \boldsymbol{B})^{-1}(C_X + \boldsymbol{B}^T H^{i+1} \boldsymbol{A})$;
end while
```

with the corresponding value function $\mathcal{V}_k^{n_p}(X_k) = (1/2) X_k^T H X_k$.

Remark 1: Despite being formulated in a similar way to MPC, but rather than being implemented online, the WEC control problem $\mathcal{P}_k^{n_p}$ yields a *closed-form* analytic optimal control law

$$u_k = F X_k \tag{20}$$

for each time instant k. The control coefficient F is fixed and is precalculated offline in the design stage, that is, no online optimization is required.

Remark 2: By partitioning $F = \left[F_x \ \ F_w\right]$, where $F_x \in \mathbb{R}^{1\times n_x}$ and $F_w \in \mathbb{R}^{1\times n_p}$, the optimal control law (20) can be rewritten as follows:

$$u_k = F_x x_k + F_w \boldsymbol{w}_k^{n_p}$$

which consists of a state feedback part, similar to conventional optimal control, and a feed-forward part, which incorporates wave forecast information to further improve EM control performance.

Instead of directly solving DARE (19) to obtain the optimal control policy (20), the controller can also be equivalently calculated using recursive iterations. We rewrite DARE (19) as follows:

$$(\boldsymbol{A} + \boldsymbol{B}F)^T H (\boldsymbol{A} + \boldsymbol{B}F) - H + F^T R F + 2F^T C_X = 0. \tag{21}$$

Based on (21), we develop the following learning-based algorithm to calculate H and F, implemented in a policy iteration (PI) manner. Hereinafter, we use the superscript j to denote an element updated at iteration step j.

Remark 3: H^i and F^i, calculated in Algorithm 1, converge to H and F calculated via (19) and (20), respectively. The algorithm is an application of Hewer's method, the convergence properties of which have been studied in [43].

Thus, we have formulated the LNOC using PI, based on complete knowledge of the system dynamics. With Algorithm 1, the associated augmented Ricatti equation can be solved with stable feedback from the augmented state, which leads to an equivalent result to the noncausal optimal controller calculated via the existing approach [2]. The resultant value function, i.e., the optimal cost function, depends quadratically on the augmented state.

B. Model-Free Reformulation of the Policy Iterative Method

To obviate the requirement for a WEC model, in Section III-B, a model-free method will be developed, to solve for H and F, where neither the policy evaluation nor policy improvement steps require model information.

To begin with, define $J(X_k)$ as the control cost function associated with $u_{i|k} = F^j x_{i|k}$ and subject to the same dynamics as follows (16):

$$J(X_k) = \sum_{i=0}^{\infty} L\big(X_{i|k}, u_{i|k}\big). \tag{22}$$

With fixed F^j, the associated cost function has the structure of $J(X_k) = (1/2)X_k^T H^j X_k$. Following the principle of RL, we iteratively update H^j and F^j, such that they can converge to H and F, respectively.

1) Policy Improvement Step Design: In the policy improvement step $j+1$, we fix H^{j+1}. The control Bellman equation gives

$$\begin{aligned} J(X_k) &= L\big(X_{0|k}, u_{0|k}\big) + J\big(X_{1|k}\big) \\ &= L(X_k, u_k) + J\big(X_{1|k}\big). \end{aligned} \tag{23}$$

Here, $X_{1|k}$ is determined by state transitions within the predictive control horizon using $X_{1|k} = \boldsymbol{A}X_k + \boldsymbol{B}u_0$; $L(X_k, u_k)$ is the the immediate control cost $L(X_k, u_k)$ at state X_k of taking action u_k; and $J(X_{1|k})$ is the accumulated cost forever after, defined as $J(X_{1|k}) := \sum_{i=1}^{\infty} L(X_{i|k}, u_{i|k})$.

Note that, since convergence and stability can be established in the control-theoretic formulation, we can directly target the accumulated energy, without the need for a discount factor γ in accumulated costs (22) and Bellman equation (23), avoiding the risk of a "short-sighted" bias in such WEC control formulations.

Next, we define an action-value-function $Q(X_k, u_k)$ as the right-hand-side of the Bellman equation (23)

$$Q(X_k, u_k) := L(X_k, u_k) + J\big(X_{1|k}\big). \tag{24}$$

With the defined action-value-function (24), we can formulate the policy improvement procedure. Using the augmented dynamics in (14), (24) can be equivalently rewritten as follows:

$$\begin{aligned} &Q(X_k, u_k) \\ &\quad= u_k C_X X_k + \frac{1}{2}Ru_k^2 + \frac{1}{2}X_{1|k}^{\mathrm{T}} H^{j+1} X_{1|k} \\ &\quad= u_k C_X X_k + \frac{1}{2}Ru_k^2 \\ &\qquad+ \frac{1}{2}(\boldsymbol{A}X_k + \boldsymbol{B}u_k)^{\mathrm{T}} H^{j+1}(\boldsymbol{A}X_k + \boldsymbol{B}u_k) \\ &\quad= \frac{1}{2}\begin{bmatrix} X_k \\ u_k \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}} H^{j+1}\boldsymbol{A} & \boldsymbol{A}^{\mathrm{T}} H^{j+1}\boldsymbol{B} + C_X^{\mathrm{T}} \\ \boldsymbol{B}^{\mathrm{T}} H^{j+1}\boldsymbol{A} + C_X & R + \boldsymbol{B}^{\mathrm{T}} H^{j+1}\boldsymbol{B} \end{bmatrix} \begin{bmatrix} X_k \\ u_k \end{bmatrix}. \end{aligned} \tag{25}$$

Here, we know from (25) that the action-value-function $Q(X_k, u_k)$ has a quadratic structure depending on X_k and u_k. Therefore, consider the action-value function $Q(X_k, u_k)$ written in the general quadratic form

$$Q(X_k, u_k) = \frac{1}{2}\begin{bmatrix} X_k \\ u_k \end{bmatrix}^{\mathrm{T}} M^{j+1} \begin{bmatrix} X_k \\ u_k \end{bmatrix} \tag{26}$$

where M^{j+1} can be partitioned into

$$M^{j+1} = \begin{bmatrix} M_{XX}^{j+1} & M_{uX}^{j+1\,\mathrm{T}} \\ M_{uX}^{j+1} & M_{uu}^{j+1} \end{bmatrix}$$

with $M_{XX}^{j+1} \in \mathbb{R}^{(n_x+n_p)\times(n_x+n_p)}$, $M_{ux}^{j+1} \in \mathbb{R}^{1\times(n_x+n_p)}$, and $M_{uu}^{j+1} \in \mathbb{R}$. Applying the first order condition, i.e., $\partial Q(X_k, u_k)/\partial u_k = 0$, we have $u_k = FX_k$, where

$$\begin{aligned} F^{j+1} &= -\big(R + \boldsymbol{B}^T H^{j+1}\boldsymbol{B}\big)^{-1}\big(C_X + \boldsymbol{B}^T H^{j+1}\boldsymbol{A}\big) \\ &= -M_{uu}^{j+1^{-1}} M_{uX}^{j+1}. \end{aligned} \tag{27}$$

Remark 4: Rather than solving for the value function $(1/2)X_k^T H X_k$ as in the model-based PI formulated in Section III, the algorithm developed in this section solves for an action-value-function $Q(X_k, u_k)$ (24). In this manner, the optimal control policy update is achieved by (27), the calculation of which requires no information on the WEC dynamics.

2) Policy Evaluation Step Design: Next, we develop the policy evaluation step by assuming F^j is fixed. Observe that the Q-function satisfies

$$J(X_k) = Q\big(X_k, F^j X_k\big), \quad J\big(X_{1|k}\big) = Q\big(X_{1|k}, F^j X_{1|k}\big)$$

and that the cost function J satisfies

$$J(X_k) = L\big(X_k, F^j X_k\big) + J\big(X_{1|k}\big).$$

Therefore, we have the Q-function Bellman equation

$$Q\big(X_k, F^j X_k\big) = L\big(X_k, F^j X_k\big) + Q\big(X_{1|k}, F^j X_{1|k}\big). \tag{28}$$

Since $x_{1|k} = Ax_{0|k} + B_u u_{0|k}^* + B_w w_k = Ax_k + B_u u_k + B_w w_k = x_{k+1}$, we can define

$$\bar{X}_{k+1} := \big[x_{k+1}, w_{k+1}, \ldots, w_{k+n_p-1}, 0\big] = X_{1|k}.$$

Here, $\bar{X}_{k+1}$ can be obtained from X_{k+1}, measured at time step $k+1$, by replacing the last element w_{k+n_p} with 0. By substituting the Q-function in (24) with (26), the Bellman equation for the Q-function becomes

$$\begin{aligned} 2L\big(X_k, F^j X_k\big) = &\begin{bmatrix} X_k \\ F^j X_k \end{bmatrix}^{\mathrm{T}} M^{j+1} \begin{bmatrix} X_k \\ F^j X_k \end{bmatrix} \\ &- \begin{bmatrix} \bar{X}_{k+1} \\ F^j \bar{X}_{k+1} \end{bmatrix}^{\mathrm{T}} M^{j+1} \begin{bmatrix} \bar{X}_{k+1} \\ F^j \bar{X}_{k+1} \end{bmatrix}. \end{aligned} \tag{29}$$

In (29), the augmented state X_k and $\tilde{X}_{k+1}$, and the running cost L, can be measured. Next, we develop a computing method to update $\tilde{M}$, from (29).

For a vector $z \in \mathbb{R}^l$, $l := n_x + n_p + 1$, define a vector operator $\zeta(.) : \mathbb{R}^l \mapsto \mathbb{R}^{n_\theta}$, $n_\theta := l(l+1)/2$, such that

$$\zeta(z) = \big[z_1^2, z_1 z_2, \ldots, z_1 z_l, z_2^2, z_2 z_3, \ldots, z_{l-1} z_l, z_l^2\big]. \tag{30}$$

For a symmetric matrix $W \in \mathbb{R}^{l\times l}$, define a matrix operator $\theta(.) : \mathbb{R}^{l\times l} \mapsto \mathbb{R}^{n_\theta}$, such that

$$\theta(W) := \left[W_{11}, 2W_{12}, \ldots, W_{1l}, W_{22}, 2W_{23}, \ldots, 2W_{(l-1)l}, W_{ll}^2\right] \tag{31}$$

where W_{ab} denotes the a and b entry of matrix W. Here, we call $\theta(W)$ the vectorisation of W and, from $\theta(W)$, the matrix W can be straightforwardly reconstructed. With the defined operators (30) and (31), (29) is rearranged into

$$Y_k = Z_k^{\mathrm{T}}\Theta \tag{32}$$

where

$$\begin{aligned} Y_k &= 2L\left(X_k, F^jX_k\right)\\ Z_k &= \zeta\left(\left[X_k, F^jX_k\right]\right) - \zeta\left(\left[\bar{X}_{k+1}, F^j\bar{X}_{k+1}\right]\right)\\ \Theta &= \theta\left(M^{j+1}\right). \end{aligned} \tag{33}$$

Remark 5: Using (32), Θ^{j+1}, and the corresponding M^{j+1}, can be estimated based on LLSs method [44], using the collected data of the augmented state trajectories of X_k and the running cost L, rather than requiring full knowledge of the WEC dynamics.

Remark 6: With the proposed control-theoretic formulation, the MF-LNOC developed in this article extends the results of MF-LQR [30] to WEC MF-LNOC, enabling the use of disturbance preview to maximize energy, and establishing theoretical guarantees on convergence and stability.

Despite having continuous state and action spaces, the proposed MF-LNOC aims to learn the action-value-function $Q(X_k, u_k)$, which has a fixed analytic quadratic structure, with a total of only $N(N+1)/2$ variables to be parameterized in the policy evaluation step, significantly less than training multiple NNs as in conventional RL, see [28], [29].

IV. Implementation

Based on the results of the control-theoretic analysis in Section III, we present the implementation of the proposed approach. In the MF-LNOC framework, the state and action of learning at time step k are defined as follows:

$$\text{state} := \left[x_k, w_k, \ldots, w_{k+n_p-1}\right], \quad \text{action} := u_k$$

where x_k is the control state; $w_k, \ldots, w_{k+n_p-1}$ are the wave excitation force values between time step k–$k+n_p-1$; and u_k is the manipulated PTO force. In the training process, since the training process takes place after completing one training epoch, we can use the measured/estimated values from w_k to w_{k+n_p-1}, which is considered noncausal at time step k.

However, when not all the states are available, a compromise has to be made by using all the available state information. Therefore, assuming the heave displacement z_k and heave velocity v_k can be measured/estimated, the state and action of learning is chosen as follows:

$$\text{state} := \left[z_k, v_k, w_k, \ldots, w_{k+n_p-1}\right], \quad \text{action} := u_k.$$

The RL reward function is defined as follows:

$$\text{reward} := \text{energy converted in one sampling interval.}$$

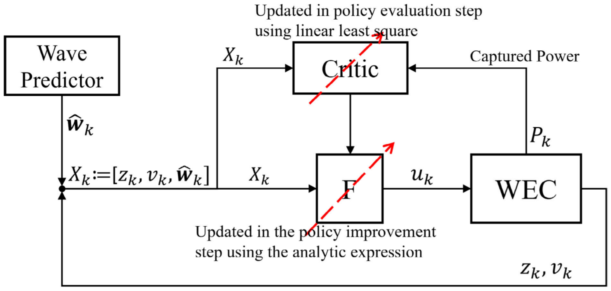

Updated in policy evaluation step
using linear least square
Wave
Predictor
X_k
Critic
Captured Power
$\hat{w}_k$
P_k
X_k:=[$z_k, v_k, \hat{w}_k$]
X_k
F
u_k
WEC
Updated in the policy improvement
step using the analytic expression
z_k, v_k

Fig. 2. Implementation framework of the MF-LNOC based on the control-theoretic approach.

Algorithm 2 Design Procedure of the Proposed MF-LNOC

```
Initialise $M^0$ and $F^0$;
while $M^j$ and $F^j$ do not converge do
    Policy Evaluation:
        while for each step $k$ of the collected data do
            Compute $X_k = [x_k, \boldsymbol{w}_k^{n_p}], \bar{X}_{k+1} = [x_{k+1}, \boldsymbol{w}_{k+1}^{n_p-1}, 0]$;
            Compute $Y_k$ and $Z_k$ via (33); If it's the first training epoch, replace $F^jX_k$ and $F^j\bar{X}_{k+1}$ with random exploring $u_k$ and $u_{k+1}$, respectively;
        end while
        Apply linear least square on (32) to obtain $\Theta$;
        $M^{j+1} = \theta^{-1}(\Theta)$;
    Policy Improvement:
        Update the LNOC coefficient $F^{j+1} = -{M_{uu}^{j+1}}^{-1}M_{uX}^{j+1}$;
end while
```

The accumulated reward, considered in MF-LNOC, represents the total energy that can be generated from the wave excitation effects between k and $k+n_p-1$. The proposed approach adopts a receding-horizon control implementation like MPC, in line with the recursive updates of wave excitation predictions.

The design procedure of the MF-LNOC, based on the control-theoretic approach, is summarized in Algorithm 2. The policy improvement step aims to update the critic, which has a fixed *quadratic* structure, depending on X_k

$$J(X_k) = Q\left(X_k, F^jX_k\right)$$

and the policy evaluation steps aim to update the actor (control) with a fixed *linear* structure, depending on X_k

$$u_k(X_k) = F^jX_k.$$

Here, $Q(.,.)$ and F are defined in (26) and (27), respectively. The implementation is illustrated in Fig. 2.

Compared with SAC [28] and BAC [29], which use NN function approximations, with this control-theoretic approach, only $N(N+1)/2$ parameters need to be estimated from the training process. In addition, this control-theoretic approach also minimizes the effort of choosing hyperparameters. In fact, if we use a batched least square (BLS) algorithm, the implementation does not involve any hyperparameter tuning, which further simplifies the controller design process.

Remark 7 (Persistent Excitation (PE) Condition): Similar to other data-driven control methods, implementation of the

MF-LNOC requires a PE condition to ensure sufficient exploration of the space, such that the kernel matrix M, and the control coefficient F, converge to their desired values. Therefore, in the exploration mode, we take random input to the system [45].

Remark 8 (Convergence and Stability Guarantee): The PI-based learning method, adopted in Algorithm 2, is a trivial generalization of [46], and using the PE condition, the convergence guarantee is proven in [46]. The stability is therefore preserved from model-based LNOC, established in Theorem 1, to MF-LNOC.

V. Numerical Simulation

In this section, we present several sets of numerical examples to verify the efficacy of the proposed method. The first set of simulations will be presented based on a reduced 2nd-order system, where the perfect prediction of wave excitation force is available. In this fully observable, accurate forecasting scenario, we verified that MF-LNOC converges to the true optimal values calculated using a model-based approach. Next, we present three sets of simulations based on a full-order system, with partial state observability of only the heave displacement and heave velocity. The focus will be on benchmarking performance, the data points required to train an MF-LNOC, adaptation to cope with the changing sea conditions, and sensitivity to wave excitation forecasting errors.

A. Simulation Case I—With Full State Information

The first simulation presented here is based on a reduced-order model, the result of which is easier to reproduce and verify, for demonstration of the design procedure, the implementation, and convergence to the optimal model-based controller.

The adopted parameters are

$$A = \begin{bmatrix} 0.7726 & 0.1834 \\ -2.1783 & 0.7614 \end{bmatrix}, \quad B_u = B_w = \begin{bmatrix} 0.0588 \\ 0.5635 \end{bmatrix} \times 10^{-3}$$
$$R = 0.0011. \tag{34}$$

The wave excitation force is generated using a JONSWAP spectrum [45], with a significant wave high of 3 m, a peak period of 5 s, and a peakedness parameter of 3.3. The excitation force dynamics, i.e., the LTI subsystem representing the input/output relationship between wave elevation and wave excitation force, are taken from [47, eqs. (54)–(56)]. The wave excitation force profile w_k is shown in Fig. 3.

When the model (34) is available, the control coefficients can be calculated either by solving the DARE (19) or by model-based PI via Algorithm 1. Both methods lead to the following identical result:

$$F = \begin{bmatrix} 81.2804 & -65.2976 & 0.0148 & 0.0537 \end{bmatrix}. \tag{35}$$

Next, we show that the MF-LNOC can recover the value of F from data collected in the past, without any knowledge of the system dynamics.

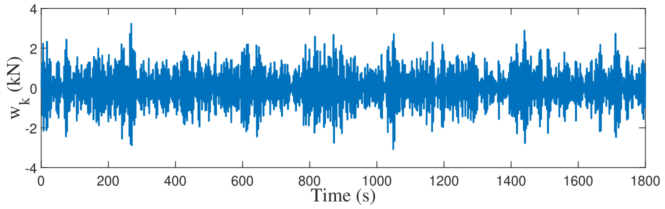


Fig. 3. 1000-s of wave excitation force profile generated using a JONSWAP wave spectrum, with a significant wave height 3 m, a peak period 5 s, and a peakedness parameter 3.3, and using wave excitation dynamics from [47].

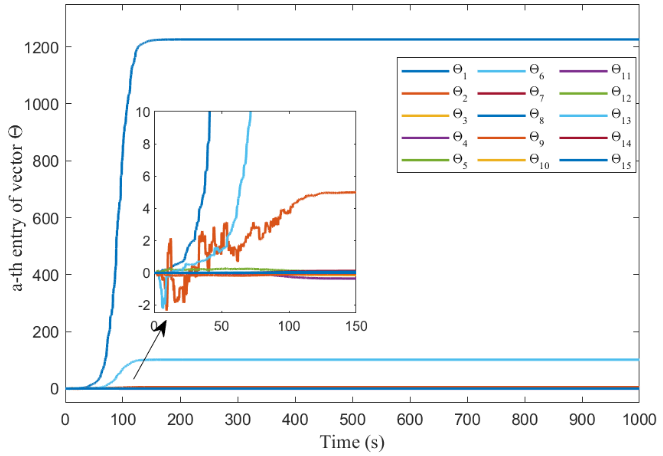


Fig. 4. ath element of vector Θ_k, estimated in the policy evaluation steps using Algorithm 2, for the full state observation case. Convergence of all 15 MF-LNOC parameters is achieved after 100 s.

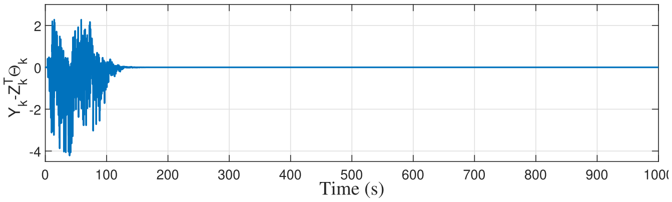


Fig. 5. Residue estimation error of the linear mapping, i.e., $Y_k - Z_k^k\Theta$ with estimated Θ in Fig. 4, for the full-state observation case.

The implementation of the MF-LNOC follows Algorithm 2, where the data are assumed to have been generated with w_k in Fig. 3 and a random signal generator input with maximal magnitude of 50-N. Since the WEC is strictly dissipative, any bounded input used in generating the data satisfies the initial stable requirement for the PI.

To better demonstrate the process of conversion, we demonstrate based on a recursive least square (RLS) estimator [44]. Initialize the RLS estimator with a vector $\Theta_0 = 0_{n_\theta \times 1}$ and a positive definite matrix $P_0 = I_{n_\theta}$. For each iteration, update P_{k+1} and Θ_{k+1} via the following two steps:

$$P_{k+1} = \frac{1}{\lambda} P_k - \frac{1}{\lambda} P_k Z_k \big(\lambda I_{n_\theta} + Z_k^T P_k Z_k\big)^{-1} Z_k^T P_k$$
$$\Theta_{k+1} = \Theta_k + P_{k+1} Z_k \big(Y_k - Z_k^{\mathrm{T}} \Theta_k\big) \tag{36}$$

where the convergence of Θ_k to a least squares estimate of Θ is proven in [48]. The corresponding M is calculated by $M = \theta^{-1}(\Theta_k)$.

A forgetting factor $\lambda = 0.98$ is chosen. Figs. 4 and 5 show the trajectory of the estimated Θ, and the estimation error characterized by $Y_k - Z_k^{\mathrm{T}}\Theta$, respectively. Note that Θ converges after a 120-s estimator "warming-up" period. Asymptotically,

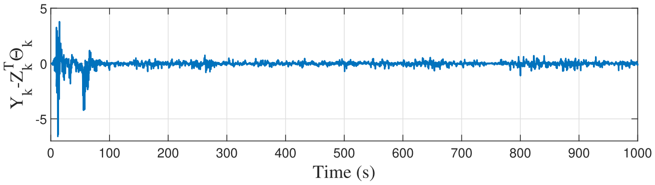


Fig. 6. Residue estimation error $Y_k - Z_k^k \Theta$ with Θ estimated using an RLS with $\lambda = 0.982$, for the partial state observation case. Estimation error is dramatically reduced, but does not diminish for the partial state observation case.

the estimation error converges to 0. The estimated action-value-function kernel matrix is

$$M = \begin{bmatrix} 1226 & 2.502 & -0.0713 & -0.1790 & 0.0424 \\ 2.502 & 101.3 & 0.05764 & 0.03054 & -0.03407 \\ -0.0713 & 0.05764 & 3.71\times10^{-5} & 2.80\times10^{-5} & 7.70\times10^{-6} \\ -0.1790 & 0.03054 & 2.80\times10^{-5} & 3.72\times10^{-5} & 2.80\times10^{-5} \\ 0.0424 & -0.03407 & 7.70\times10^{-6} & 2.80\times10^{-5} & -5.217\times10^{-4} \end{bmatrix}.$$

The LNOC coefficient is calculated, in the policy improvement step, as follows:

$$\begin{aligned} F &= -M_{uu}^{-1} M_{uX} \\ &= \begin{bmatrix} 81.2804 & -65.2976 & 0.0148 & 0.0537 \end{bmatrix} \end{aligned} \tag{37}$$

which shows the efficacy of the proposed MF-LNOC in accurately recovering the control coefficient F in (35) calculated from model-based approaches. The result of Simulation Case I verify that, with the control-informed decision, indeed, gives a correct guess of the RL structure.

B. Simulation Case II—With Partial State Information

The second simulation set is based on a full-order WEC model, the parameters of which are adopted from [2] and [47]. The state-space approximation of the radiation dynamics (3) is

$$A_r = \begin{bmatrix} 0 & 0 & -17.9 \\ 1 & 0 & -17.7 \\ 0 & 1 & -4.41 \end{bmatrix}, \quad B_r = \begin{bmatrix} 38.6 \\ 379 \\ 89 \end{bmatrix}$$

$$C_r = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T, \quad D_r = 0.$$

The simulation parameters are $r = 5 \times 10^{-3}$ for (7), sampling period $t_s = 0.2$ s and wave excitation prediction step $n_p = 5$. Heave elevation z_k, and heave velocity v_k, are assumed to be directly measurable.

However, now due to the unavailability of knowledge of the WEC dynamics, a model-based WEC state estimator cannot be designed. Therefore, the MF-LNOC is formulated based on reduced-order augmented state trajectories $X_k := [z_k, v_k, \boldsymbol{w}_k^{n_p}]$ and $\bar{X}_{x+1} := [z_{k+1}, v_{k+1}, \boldsymbol{w}_{k+1}^{n_p-1}, 0]$. Fig. 6 shows the training error when using the RLS estimator (36). Compared with Figs. 4 and 5, we can see that, after a 100-s "warming-up" period, the estimate of the actor-value-function remains close to 0, rather than converging to 0. This is because, when the system has higher order dynamics, with partial availability of the state, the proposed MF-LNOC can only find the best estimate of the actor-value-function from the reduced-order states, for the full-order system.

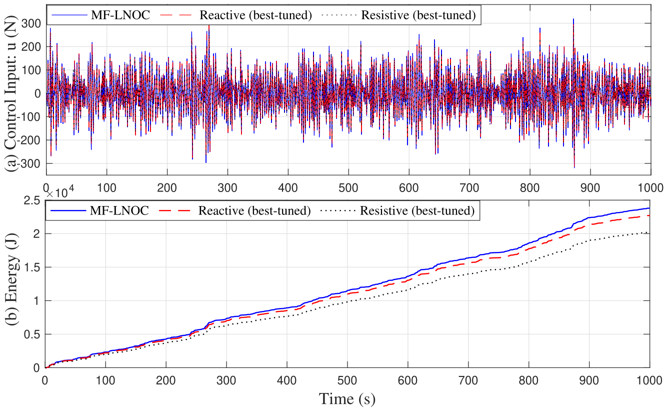


Fig. 7. Performance benchmark among the trained MF-LNOC (blue solid line), a best-tuned resistive control via trial and error (black dotted line), and a best-tuned reactive control via trial and error, based on a 1000-s wave segment, generated from a JONSWAP spectrum, with a significant wave height of 4 m, a peak period of 6 s, and a peakedness factor of 3.3. Energy output: 2.381×10^4 J for MF-LNOC, 2.271×10^4 J for "reactive" and 2.023×10^4 J for "resistive." (a) Control input response. (b) Energy output response.

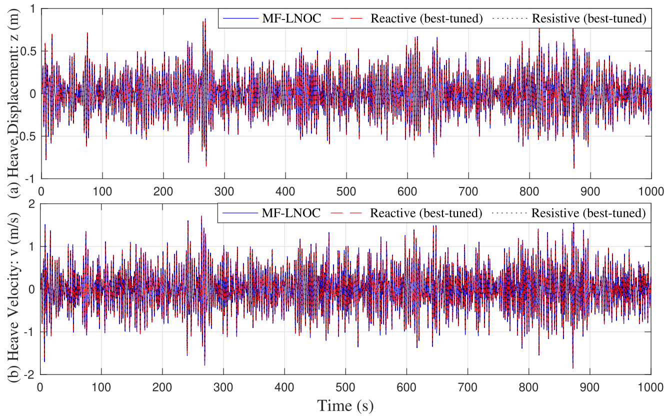


Fig. 8. Corresponding state responses of Fig. 7. (a) Heave displacement z. (b) Heave velocity v response.

After the training process, we have the MF-LNOC control coefficient: $F = [173.2\ -136.9\ -0.0147\ -0.0552\ 0.0929\ -0.0603\ \ -0.0077]$.

Next, we present a set of time simulations, based on a 1000-s wave segment, generated from a JONSWAP spectrum, with a significant wave height of 4 m, a peak period of 6 s, and a peakedness factor of 3.3. The MF-LNOC is benchmarked with two feedback control frameworks.

1) A resistive controller, $u_k = G v_k$, referred to as "resistive."
2) A reactive controller with the structure $u_k = F z_k + G v_k$, referred to as "reactive."

Here, both resistive and reactive controllers are optimally tuned via trial and error based on the same waveform segments, and fixed with parameters that lead to the most energy being produced.

Fig. 7 shows the energy produced for the three controllers. Figs. 7 and 8 show the control input and energy produced, and heave displacement and velocity responses, respectively.

We can observe that despite having similar control input magnitudes and resulting in similar levels of oscillations, the

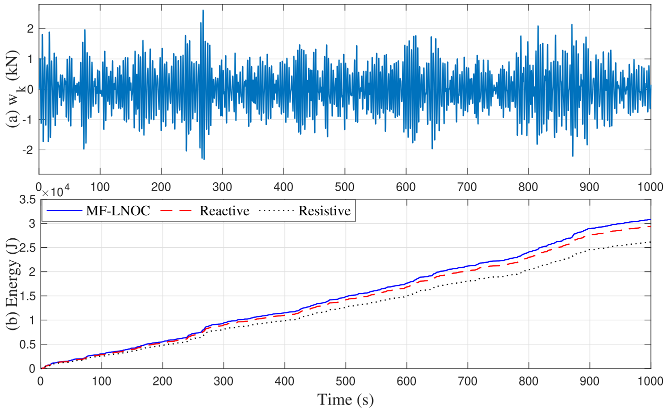


Fig. 9. Minor sea state change. (a) 1000-s wave excitation force w_k segment, based on a different sea state with a significant wave height of 3.8 m, a peak period of 5.5 s, and the same peakedness factor of 3.3. Energy output performance benchmark study based on w_k profile above, using the same controllers as shown in Fig. 7 (bottom). (b) Energy output: 3.080 × 10^4 J for MF-LNOC, 2.937 × 10^4 J for "reactive," and 2.613 × 10^4 J for "resistive."

energy generated using MF-LNOC reaches 2.381 × 10^4 J, compared to 2.271 × 10^4 J with the best-tuned "reactive" $u_k = 125z_k - 161v_k$ and 2.023 × 10^4 J with the best-tuned "resistive" $u_k = -164v_k$. This result is significant because even with two feedback controllers specifically tuned with optimal parameters for the testing wave segments, MF-LNOC, without the need for sea segment-based tuning can still achieve a 4.81% performance advantage.

Next, we introduce a minor change in the sea state to test the inherent robustness of MF-LNOC. Fig. 9 shows a benchmark study, based on the JONSWAP-generated wave segment, with the significant wave height changed from 4 to 3.3 m, and the peaked period from 6 to 5 m. The peakedness factor remains unchanged. Since there is no major change in WEC dynamics, MF-LNOC does not need to be retrained/reparametrised. The energies generated are 3.080 × 10^4 J for the proposed MF-LNOC, 2.937 × 10^4 J for the reactive controller and 2.613 × 10^4 J, respectively, representing a 4.87% and a 18.87% performance advantages of using the proposed MF-LNOC, compared with "reactive" and "resistive" controller, respectively.

To further test the adaption to major sea state changes that lead to changes in dynamics, we assume that the significant wave height changes from 4 to 2.5 m, the peak period changes from 6 to 4 s, and the radiation dynamics change to

$$A_r = \begin{bmatrix} 0 & 0 & -19 \\ 1 & 0 & -17 \\ 0 & 1 & -4.6 \end{bmatrix}, \quad B_r = \begin{bmatrix} 38.6 \\ 370 \\ 85 \end{bmatrix}$$
$$C_r = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^{\mathrm{T}}, \quad D_r = 0. \tag{38}$$

In this case, we reparameterized MF-LNOC by switching on the training mechanism in Algorithm 2.

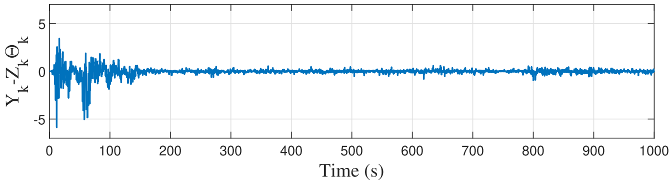


Fig. 10. Major sea state change: residual estimation error $Y_k - Z_k^k\Theta$ with Θ estimated using the estimator as shown in Fig. 6, tested in a case of both dynamics and sea changes.

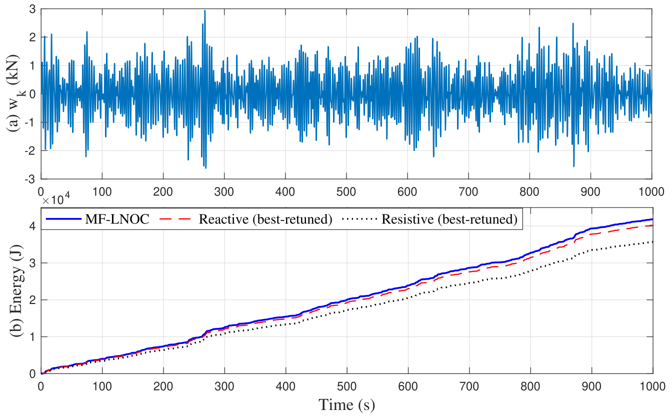


Fig. 11. Major sea state change. (a) 1000-s wave excitation force w_k segment used for simulation, which was generated based on a sea state with a significant wave height of 2.5 m, a peak period of 4 s, and the same peakedness factor of 3.3. (b) Energy output: 4.189 × 10^4 J for MF-LNOC, 4.021 × 10^4 J for the best-retuned "reactive," and 3.568 × 10^4 J for the best-retuned "resistive," respectively.

Fig. 10 shows the residue error in the retraining process. After 150 s, the error is dramatically reduced and maintained at a minimum level. The corresponding MF-LNOC coefficients have been updated to $F = [183.7 \quad -145.5 \quad -0.0301 \quad -0.0266 \; 0.0447 \quad -0.0213 \quad -0.0156]$.

Fig. 11 presents a comparative time simulation of MF-LNOC, compared with "reactive" and "resistive," both retuned to their optimal values. The energy generated is 4.189 × 10^4 J for the retrained MF-LNOC, 4.021 × 10^4 J for the best-retuned "reactive" ($u_k = 120z_k - 155v_k$), and 3.568 × 10^4 J for the best-retuned "resistive" ($u_k = -160v_k$), respectively. This represents a 4.18% and 17.40% performance advantage of using the MF-LNOC over the best achievable performances of the "reactive" and "resistive," respectively.

Note that the above optimal performance of the best-retuned feedback policies is only achievable for existing RL algorithms [23], [25], [28], [29], after completing the retraining process for the existing WEC RL algorithm, among which the fastest converging algorithms will need approximately 1.5 h, as reported in [29], although the time may be slightly reduced for retraining. This further highlights the advantage of the proposed MF-LNOC, which allows control reparametrization to be completed in just 150 s, well in pace with a typical sea change of 30 min [49].

Since MF-LNOC requires wave prediction, a natural question is how the performance will be affected when realistic imperfect wave prediction is used, which inevitably introduces inaccuracies. Therefore, we present the time simulations using the same settings, and wave segments, as shown in Fig. 11. Nevertheless, rather than using the true value of f_{ex} as in the previous simulations, we generate f_{ex} predictions using a widely used autoregressive (AR) model [50], where the

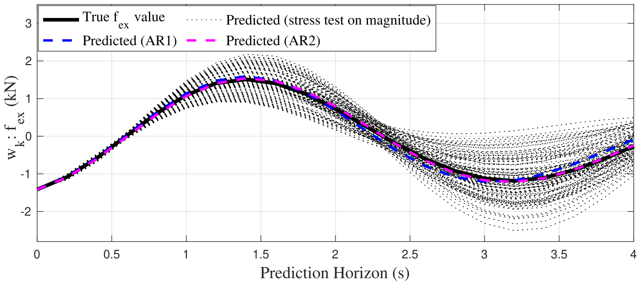


Fig. 12. Snapshot of the inaccurate f_{ex} predictions used in the sensitivity study of magnitude error. [Black solid line: true f_{ex} value; black dotted line: 50 scenarios of f_{ex}, with artificially induced error on magnitudes, for the Monte Carlo stress test; blue dashed line: f_{ex} predicted using "AR1" (AR parametrized based on a different sea state); and purple dashed line: f_{ex} predicted using "AR2" (AR parametrized based on the same sea state)].

p-step-ahead wave excitation force, predicted at time instant k, is calculated using

$$\hat{w}_{k+p} = \sum_{i=1}^{H} \phi_i \hat{w}_{k+p-i} \tag{39}$$

where $\hat{w}$ denotes the predicted value of w; H denoted the AR model order, set as $H = 15$; $\hat{w}_{k+p-i} = w_{k+p-i}$ for $p < i$ since the value of w is known up to k.

Two parameterization of the AR coefficients ϕ are benchmarked.

1) AR parameterised using f_{ex} in Fig. 9(a) of a different sea state, referred to as "AR1."
2) AR parameterised using f_{ex} in Fig. 11(a), of the same sea state, referred to as "AR2."

A snapshot of the f_{ex} predictions, calculated using predictors "AR1" and "AR2" are shown in Fig. 12, in the blue and purple dashed lines, respectively. The energy produced is 4.184×10^4 and 4.186×10^4 J, when "AR1" and "AR2" are used, respectively, showing only minor decreases in energy produced, even when the AR model is trained from a different sea state. This is because, as in the model-based case, the performance of MF-LNOC heavily depends on the wave prediction between 0 and 1 s, which is fairly accurate for both cases.

To further test the sensitivity of the MF-LNOC, we present, in Figs. 12 and 13, two sets of Monte Carlo stress tests, with artificially induced errors in magnitude and phase, respectively. In Fig. 12, a snapshot of 50 scenarios of f_{ex} predictions used in the Monte Carlo magnitude stress test is shown in black dotted lines. The energy generated is 4.182×10^4 J on average for the 50 scenarios, with a maximum of 4.196×10^4 J and a minimum of 4.164×10^4 J. The result shows only 0.06% performance degradation, even for the worst case, implying that the proposed MF-LNOC is robust to f_{ex} prediction errors in magnitudes.

Fig. 13(a) shows a snapshot of the inaccurate f_{ex} prediction used in the second Monte Carlo stress test, where the predicted f_{ex} deviates from the true value in time delays from 0 to 1.5 s, with a step increase of 0.05 s. The corresponding responses for energy generated are shown in Fig. 13(b). A substantial decrease occurs, with generated energy reduced to 3.962×10^4 J, with a 0.5-s phase lag, and further reduced to 3.411×10^4 J, with a phase lag of 1.5 s. This means that MF-LNOC is sensitive to phase errors in f_{ex} predictions. However, as shown in the blue and purple dashed lines in Fig. 13(a), predictions are fairly accurate within 1 s of prediction, even using a simple AR with simple model structure parameterized from a different sea state.

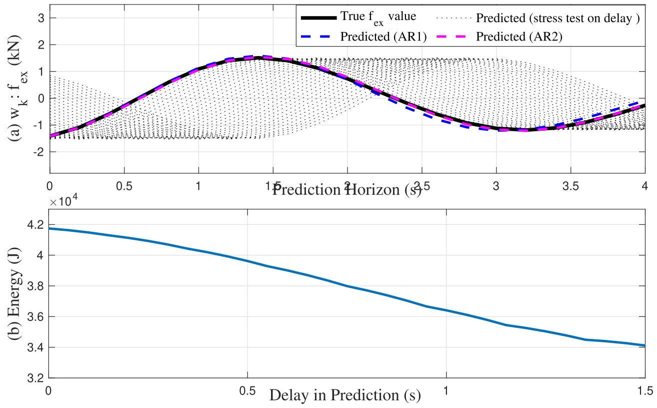


Fig. 13. (a) Snapshot of the inaccurate f_{ex} predictions used in the sensitivity study of magnitude error. (Black solid line: true f_{ex} value; black dotted line: f_{ex} prediction, with artificially induced time delays from 0 to 1.5 s; and blue/purple dashed line: f_{ex} predicted using "AR1"/"AR2.") (b) Energy output response.

VI. Conclusion

This article develops a control-theoretic approach for the EM control problem for WECs, via the formulation of a model-free linear noncausal optimal controller (MF-LNOC), based on reinforcement Q-learning. The MF-LNOC developed in this article offers several advantages over existing WEC RL-based control algorithms. First, the MF-LNOC algorithm can leverage wave-by-wave prediction information to improve energy conversion efficiency. Second, unlike existing learning-based algorithms that rely on prefixed parametric feedback control structures, the MF-LNOC algorithm avoids such limitations, enabling the full utilization of the performance potential of WECs. Third, the MF-LNOC eliminates the need for discretizing the decision space into a small number of action spaces, allowing for a more precise and comprehensive exploration of the solution space, and resulting in improved control precision. Finally, the MF-LNOC offers practical benefits regarding data requirements for training. It requires significantly fewer data points to converge than other RL-based approaches, reducing the time and resources needed to train the controller. The results demonstrate that the MF-LNOC achieves performance close to that of the (exact) model-based controller. This is significant, as it is well known that both the linearizing assumptions (especially small movement) for boundary-element-based hydrodynamic codes, and sensitivity to modeling errors, are challenged in the design of WEC control systems [12], [16]. Furthermore, the proposed algorithm can be employed in two ways: 1) formulation of a controller or 2) adaptation of the controller to accommodate model changes, which may occur as a result of operating point changes (device nonlinearity or sea state changes). The MF-LNOC proposed in this article expands the possibilities for WEC developers to formulate simple, reliable, and efficient controllers, flexible and robust to sea state variations and changes in the implicit "representative linearized model," without relying on complex first principles modeling.

Appendix
Proof of Lemma 1

Assumption 1 implies that, for the undisturbed system $\dot{x}(t) = Ax(t) + B_u u(t)$, there exists a positive definite matrix $P_c \in \mathbb{R}^{n_x \times n_x} > 0$ such that

$$\frac{d}{dt}\left(\frac{1}{2}x^{\mathrm{T}}(t)P_c x(t)\right) < -p(x(t), u(t)) \tag{40}$$

holds for all $x(t)$ and $u(t)$, where $p(x, u)$ is defined in (7). Integrating both sides of (40) from $t = kt_s$ to $(k+1)t_s$, and with $u(t) = u_k$, we have

$$\frac{1}{2}(Ax_k + Bu_k)^T P_c(Ax_k + B_u u_k) - \frac{1}{2}x_k^T P_c x_k < -\int_{kt_s}^{(k+1)t_s} p(x(\tau), u(\tau))d\tau. \tag{41}$$

Here, we have shown passivity for the undisturbed discrete-time system. Next, we show the passivity for the augmented system, considering a disturbance. Equation (41) can be equivalently written in the form of a linear matrix inequality

$$\frac{1}{2}\begin{bmatrix} A^T P_c A - P_c & A^T P_c B_u - C_x \\ B_u^T P_c A - C_x & B_u^T P_c B_u - R \end{bmatrix} \leq -\delta_0 I_{n_x+1} \tag{42}$$

for some $\delta_0 > 0$, where R is defined in (15); $C_x := C_z(A - I)$.

This further implies (via the Schur complement) that there exist positive constants $0 < \delta_1 < \delta_0$ and $\gamma_0 > 0$ such that

$$\frac{1}{2}\begin{bmatrix} A^T P_c A - P_c & A^T P_c B_u - C_x & A^T P_c B_w \\ B_u^T P_c A - C_x & B_u^T P_c B_u - R & B_u^T P_c B_w \\ B_w^T P_c A & A^T P_c B_u - C_x^{\mathrm{T}} & -\gamma_0 \end{bmatrix} \leq -\delta_1 I_{n_x+1+n_w}. \tag{43}$$

By pre- and post-multiplying (43) with $[x_k, u_k, w_k]^{\mathrm{T}}$ and $[x_k, u_k, w_k]$, respectively, we have that, for disturbed system $x_{k+1} = Ax_k + B_u u_k + B_w w_k$

$$\frac{1}{2}\left(x_{k+1}^T P_c x_{k+1} - x_k^T P_c x_k - \gamma_0 w_k^2\right) \leq -L(X_k, u_k) - \delta_1\left(x_k^T x_k + u_k^2\right).$$

Define a block diagonal matrix

$$P_{\mathrm{aug}} = \begin{bmatrix} P_c & & & \\ & \gamma_0 & & \\ & & \ddots & \\ & & & \gamma_{n_p-1} \end{bmatrix}$$

where $\gamma_{i+1} = \gamma_i + 2\delta_1$ for $i \in \mathbb{N}_{0:i-2}$. For the augmented dynamics $X_{k+1} = \boldsymbol{A}X_k + \boldsymbol{B}u_k$, we have

$$\begin{aligned}\frac{1}{2}&\left(X_{k+1}^{\mathrm{T}} P_{\mathrm{aug}} X_{k+1} - X_k^{\mathrm{T}} P_{\mathrm{aug}} X_k\right) \\ &= \frac{1}{2}\left(x_{k+1}^T P_c x_{k+1} - x_k^T P_c x_k - \gamma_0 w_k^2\right) - \delta_1 \sum_{i=1}^{n_p-1} w_{k+i}^2 \\ &\leq -L(X_k, u_k) - \delta_1\left(X_k^T X_k + u_k^2\right)\end{aligned}$$

which completes the proof.

References

[1] J. V. Ringwood, G. Bacelli, and F. Fusco, "Energy-maximizing control of wave-energy converters: The development of control system technology to optimize their operation," *IEEE Control Syst. Mag.*, vol. 34, no. 5, pp. 30–55, Oct. 2014.

[2] S. Zhan and G. Li, "Linear optimal noncausal control of wave energy converters," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 4, pp. 1526–1536, Jul. 2019.

[3] N. Faedo, S. Olaya, and J. V. Ringwood, "Optimal control, MPC and MPC-like algorithms for wave energy systems: An overview," *IFAC J. Syst. Control*, vol. 1, pp. 37–56, Sep. 2017.

[4] S. Zou, O. Abdelkhalik, R. Robinett, G. Bacelli, and D. Wilson, "Optimal control of wave energy converters," *Renew. Energy*, vol. 103, pp. 217–225, Apr. 2017.

[5] G. Bacelli and J. V. Ringwood, "Numerical optimal control of wave energy converters," *IEEE Trans. Sustain. Energy*, vol. 6, no. 2, pp. 294–302, Apr. 2015.

[6] G. Li, G. Weiss, M. Mueller, S. Townley, and M. R. Belmont, "Wave energy converter control by wave prediction and dynamic programming," *Renew. Energy*, vol. 48, pp. 392–403, Dec. 2012.

[7] A. Mérigaud and J. V. Ringwood, "Towards realistic non-linear receding-horizon spectral control of wave energy converters," *Control Eng. Pract.*, vol. 81, pp. 145–161, Dec. 2018.

[8] R. Genest and J. V. Ringwood, "Receding horizon pseudospectral control for energy maximization with application to wave energy devices," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 1, pp. 29–38, Jan. 2017.

[9] G. Li, "Nonlinear model predictive control of a wave energy converter based on differential flatness parameterisation," *Int. J. Control*, vol. 90, no. 1, pp. 68–77, Jan. 2017.

[10] N. Faedo, G. Scarciotti, A. Astolfi, and J. V. Ringwood, "Nonlinear energy-maximizing optimal control of wave energy systems: A moment-based approach," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 6, pp. 2533–2547, Nov. 2021.

[11] Z. Liao, N. Gai, P. Stansby, and G. Li, "Linear non-causal optimal control of an attenuator type wave energy converter M4," *IEEE Trans. Sustain. Energy*, vol. 11, no. 3, pp. 1278–1286, Jul. 2020.

[12] J. Davidson, S. Giorgi, and J. V. Ringwood, "Linear parametric hydrodynamic models for ocean wave energy converters identified from numerical wave tank experiments," *Ocean Eng.*, vol. 103, pp. 31–39, Jul. 2015.

[13] W. Cummins, "The impulse response function and ship motions," in *Schiffstechnik*. Heft: Band, 1962, pp. 101–109.

[14] C. Windt, N. Faedo, M. Penalba, F. Dias, and J. V. Ringwood, "Reactive control of wave energy devices—The modelling paradox," *Appl. Ocean Res.*, vol. 109, Apr. 2021, Art. no. 102574.

[15] S. Zhan, J. Na, G. Li, and B. Wang, "Adaptive model predictive control of wave energy converters," *IEEE Trans. Sustain. Energy*, vol. 11, no. 1, pp. 229–238, Jan. 2020.

[16] J. V. Ringwood, A. Mérigaud, N. Faedo, and F. Fusco, "An analytical and numerical sensitivity and robustness analysis of wave energy control systems," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 4, pp. 1337–1348, Jul. 2020.

[17] J. Davidson and J. Ringwood, "Mathematical modelling of mooring systems for wave energy Converters—A review," *Energies*, vol. 10, no. 5, p. 666, May 2017.

[18] J. Davidson, R. Genest, and J. V. Ringwood, "Adaptive control of a wave energy converter," *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, pp. 1588–1595, Oct. 2018.

[19] E. Anderlini, D. I. M. Forehand, P. Stansell, Q. Xiao, and M. Abusara, "Control of a point absorber using reinforcement learning," *IEEE Trans. Sustain. Energy*, vol. 7, no. 4, pp. 1681–1690, Oct. 2016.

[20] S. Zhan, B. Wang, J. Na, and G. Li, "Adaptive optimal control of wave energy converters," *IFAC-PapersOnLine*, vol. 51, no. 29, pp. 38–43, 2018.

[21] D. Bertsekas, *Reinforcement Learning and Optimal Control*. Nashua, NH, USA: Athena Scientific, 2019.

[22] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 3rd Quart., 2009.

[23] E. Anderlini, D. I. M. Forehand, E. Bannon, Q. Xiao, and M. Abusara, "Reactive control of a two-body point absorber using reinforcement learning," *Ocean Eng.*, vol. 148, pp. 650–658, Jan. 2018.

[24] E. Anderlini, D. I. M. Forehand, E. Bannon, and M. Abusara, "Control of a realistic wave energy converter model using least-squares policy iteration," *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1618–1628, Oct. 2017.

[25] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[26] J. Trigueiro, M. A. Botto, S. Vieira, and J. Henriques, "Control of a wave energy converter using reinforcement learning," in *Proc. APCA Int. Conf. Autom. Control Soft Comput.*, Caparica, Portugal. Cham, Switzerland: Springer, 2022, pp. 567–576.

[27] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[28] E. Anderlini, S. Husain, G. G. Parker, M. Abusara, and G. Thomas, "Towards real-time reinforcement learning control of a wave energy converter," *J. Mar. Sci. Eng.*, vol. 8, no. 11, p. 845, Oct. 2020.

[29] L. G. Zadeh, A. S. Haider, and T. K. A. Brekken, "Bayesian actor-critic wave energy converter control with modeling errors," *IEEE Trans. Sustain. Energy*, vol. 14, no. 1, pp. 3–11, Jan. 2023.

[30] S. Bradtke, "Reinforcement learning applied to linear quadratic regulation," in *Proc. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 5, 1992, pp. 295–302.

[31] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 1, pp. 142–160, Jan. 2021.

[32] S. A. A. Rizvi, A. J. Pertzborn, and Z. Lin, "Reinforcement learning based optimal tracking control under unmeasurable disturbances with application to HVAC systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7523–7533, Dec. 2022.

[33] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming with an application to power systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1150–1156, Jul. 2013.

[34] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Syst. Mag.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.

[35] M. Farajvand, D. García-Violini, and J. V. Ringwood, "Representative linearised models for a wave energy converter using various levels of force excitation," *Ocean Eng.*, vol. 270, Feb. 2023, Art. no. 113635.

[36] J. Falnes and A. Kurniawan, *Ocean Waves and Oscillating Systems: Linear Interactions Including Wave-Energy Extraction*, vol. 8. Cambridge Univ. Press, 2020.

[37] M. Penalba, T. Kelly, and J. V. Ringwood, "Using nemoh for modelling wave energy converters: A comparative study with wamit," in *Proc. Eur. Wave Tidal Energy Conf. (EWTEC)*, Cork, Ireland, 2017, pp 631.1–631.10.

[38] Y. Peña-Sanchez, M. Garcia-Abril, F. Paparella, and J. V. Ringwood, "Estimation and forecasting of excitation force for arrays of wave energy devices," *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, pp. 1672–1680, Oct. 2018.

[39] S. Zou, X. Zhou, I. Khan, W. W. Weaver, and S. Rahman, "Optimization of the electricity generation of a wave energy converter using deep reinforcement learning," *Ocean Eng.*, vol. 244, Jan. 2022, Art. no. 110363.

[40] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.

[41] S. Zhan, Y. Chen, and J. Ringwood, "Terminal weight and terminal set design for economic model predictive control for wave energy converters," *Int. J. Robust Nonlinear Control*, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/rnc.6841

[42] O. I. Olanrewaju and J. M. Maciejowski, "Implications of dissipativity on stability of economic model predictive control—The indefinite linear quadratic case," *Syst. Control Lett.*, vol. 100, pp. 43–50, Feb. 2017.

[43] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Trans. Autom. Control*, vol. AC-16, no. 4, pp. 382–384, Aug. 1971.

[44] S. A. U. Islam and D. S. Bernstein, "Recursive least squares for real-time implementation [lecture notes]," *IEEE Control Syst. Mag.*, vol. 39, no. 3, pp. 82–85, Jun. 2019.

[45] K. Hasselmann and D. Olbers, "Measurements of wind-wave growth and swell decay during the joint North Sea wave project (JONSWAP)," *Ergänzung zur Deut. Hydrogr. Z., Reihe A*, vol. 12, no. 8, pp. 1–95, 1973.

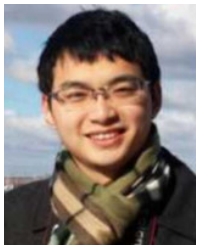
[46] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, Mar. 2007.

[47] Z. Yu and J. Falnes, "State-space modelling of a vertical cylinder in heave," *Appl. Ocean Res.*, vol. 17, no. 5, pp. 265–275, Oct. 1995.

[48] S. Dasgupta and Y.-F. Huang, "Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 3, pp. 383–392, May 1987.

[49] J. V. Ringwood, S. Zhan, and N. Faedo, "Empowering wave energy with control technology: Possibilities and pitfalls," *Annu. Rev. Control*, vol. 55, pp. 18–44, Apr. 2023.

[50] Y. Peña-Sanchez, A. Mérigaud, and J. V. Ringwood, "Short-term forecasting of sea surface elevation for wave energy applications: The autoregressive model revisited," *IEEE J. Ocean. Eng.*, vol. 45, no. 2, pp. 462–471, Apr. 2020.



Siyuan Zhan (Member, IEEE) received the B.Sc. degree from Shanghai Jiaotong University, Shanghai, China, in 2013, the M.Sc. degree from the University of Pennsylvania, Philadelphia, PA, USA, in 2014, and the Ph.D. degree from the Queen Mary University of London, London, U.K., in 2018.

He is currently an Assistant Professor with the School of Engineering, Trinity College Dublin, Dublin, Ireland, and an Honorary Research Fellow with the Centre for Ocean Energy Research, Maynooth University, Maynooth, Ireland, with a research focus on mechanical engineering, model predictive control, learning-based control, and marine renewable energy systems.

John V. Ringwood (Fellow, IEEE) received the Diploma degree in electrical engineering from Dublin Institute of Technology, Dublin, Ireland, in 1981, and the Ph.D. degree in control systems from Strathclyde University, Glasgow, U.K., in 1985.

He is currently a Professor of electronic engineering and the Director of the Centre for Ocean Energy Research, Maynooth University, Maynooth, Ireland. His research interests include time series modeling, ocean energy, and biomedical engineering.

Dr. Ringwood is a Chartered Engineer and a fellow of Engineers Ireland. He is a member of the editorial boards of IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, *IET Renewable Power Generation*, and the *Journal of Ocean Engineering and Marine Energy*. He received the Chevalier des Palmes Academiques from French Government for contributions to wave energy and the Outstanding Paper Prize Awards for *IEEE Control Systems Magazine* in 2016 and IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY in 2023.