Reactive Control of a Two-Body Point Absorber using Reinforcement Learning

E. Anderlini[a,b,c,d,*], D.I.M. Forehand[a,**], E. Bannon[b], Q. Xiao[c],
M. Abusara[d,**]

[a]*Institute of Energy Systems, University of Edinburgh, Faraday Building, Colin Maclaurin Road, Edinburgh, EH9 3DW, UK*
[b]*Wave Energy Scotland, 10 Inverness Campus, Inverness, IV2 5NA, UK*
[c]*Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, 100 Montrose Street, Glasgow, G4 0LZ, UK*
[d]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn Campus, Penryn, TR10 9FE, UK*

Abstract

In this article, reinforcement learning is used to obtain optimal reactive control of a two-body point absorber. In particular, the Q-learning algorithm is adopted for the maximization of the energy extraction in each sea state. The controller damping and stiffness coefficients are varied in steps, observing the associated reward, which corresponds to an increase in the absorbed power, or penalty, owing to large displacements. The generated power is averaged over a time horizon spanning several wave cycles due to the periodicity of ocean waves, discarding the transient effects at the start of each new episode. The model of a two-body point absorber is developed in order to validate the control strategy in both regular and irregular waves. In all analysed sea states, the controller learns the optimal damping and stiffness coefficients. Furthermore, the scheme is independent of internal models of the device response, which means that it can adapt to variations in the unit dynamics with time and does not present modelling errors.

Keywords: Reinforcement learning (RL), Q-learning, reactive control,

[*] Principal corresponding author
[**] Corresponding author

Email addresses: `E.Anderlini@ed.ac.uk` (E. Anderlini), `D.Forehand@ed.ac.uk` (D.I.M. Forehand), `elva.bannon@hient.co.uk` (E. Bannon), `Qing.Xiao@strath.ac.uk` (Q. Xiao), `M.Abusara@exeter.ac.uk` (M. Abusara)

point absorber, wave energy converter (WEC).

1. Introduction

Wave power is a renewable energy resource that can considerably contribute to the future energy generation thus reducing society's dependence on fossil fuels. Although a potential of up to 2.1 TW of power has been estimated globally [1], wave energy converter (WEC) devices are not economically viable yet, despite a large number of different designs having been suggested [2]. The design of an effective control strategy is fundamental in order to address this problem, since it can result in substantial gains in absorbed energy without additional hardware costs.

Over the years, different control strategies have been proposed for the maximization of power extraction of WECs. A review of the first studies can be found in [3], while [4] presents a review of recent techniques. From hydrodynamic considerations, complex-conjugate control would theoretically provide optimal energy absorption by achieving resonance between the WEC and the incident waves [3]. Nevertheless, delivering optimal control may be infeasible in reality due to the associated excessive motions and loads in extreme waves. Hence, alternative suboptimal control schemes have been implemented, which include physical constraints on the motions, forces and power rating of the device [4].

Latching, declutching, model-predictive and simple-but-effective control are instances of real-time WEC control schemes. Firstly suggested by [5], latching control achieves resonance conditions by adjusting the time period when the machine is locked in place through a dedicated mechanism [6, 7]. During the remaining part of the wave cycle, the device motions are linearly damped. Declutching control presents a similar concept, but in this case the power take-off (PTO) system is disconnected during part of the wave cycle through a by-pass valve (with hydraulic PTOs) as opposed to being fixed in place [8]. Model predictive control applies at each time step the force that is expected to result in maximum energy absorption over a future time horizon [9, 10, 11, 12]. Simple-but-effective control obtains an estimate for the optimal controller force by modelling the current excitation force as a narrow-banded function [13]. These control strategies can include constrains on the motions and loading of WECs. While it is hard to scale latching control to farms of WECs, model predictive control has been successfully implemented for multi-body devices and even small array problems [14, 15, 16, 17, 18]. However, model predictive control presents high computational requirements. Simple-but-effective control results in similar performance, but with a simpler implementation [4]. Nevertheless, these methods are strongly affected by the accuracy of the prediction of the future wave excitation force, usually over a short time horizon, as well as of the model of the machine dynamics [4].

Resistive and reactive control represent alternative types of schemes that rely on time-averaged sea states, so that stationary wave conditions are assumed [3]. Numerical simulations are performed so as to obtain the PTO damping (resistive control) or combination of damping and stiffness (reactive control) that result in maximum energy absorption in each sea state [19]. It is possible to include force saturation within the numerical model and displacement constraints in the cost function. On the one hand, these techniques may present a lower efficiency as compared with on-line control schemes [18]. On the other hand, resistive and reactive control are conceptually simple to understand, and they present much lower computational costs than real-time methods. Furthermore, they are easily scalable to multi-body or multiple-device problems, as for instance shown by [19].

The aforementioned schemes suffer from a significant problem: the optimal control action is determined based on internal models of the body dynamics. Therefore, modelling errors can severely affect the performance of these algorithms, with significant drops in efficiency. In addition, these control strategies do not account for changes in the device dynamics over time, e.g. due to slow marine growth or sudden non-critical subsystem failures. For these reasons, the authors have proposed the application of reinforcement learning (RL) to resistive control in a previous work [20]. With this machine learning algorithm, the controller learns the optimal PTO damping coefficient in every sea state directly from experience. Penalties for large displacements are included to prevent failures in extreme waves.

In this article, the developed control strategy based on RL is generalised to reactive control. Although WECs are expected to be deployed in arrays so as to exploit the advantage of economies of scale [19], we consider a single, axisymmetric device for simplicity. In particular, a more realistic WEC than that in [20] is analysed: a two-body point absorber, similar to the reference model 3 in [21, 22, 23]. Point absorbers, which extract energy by resisting the motions of a small floating body subject to wave loading through a PTO system, represent a well-understood and simple offshore WEC technology [2]. The performance of the algorithm is assessed in both regular and irregular

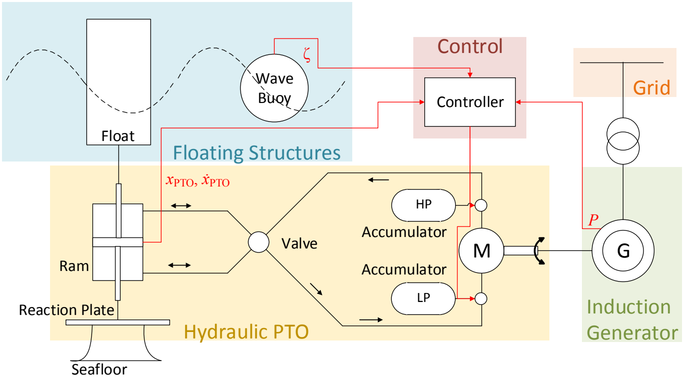


Figure 1: Diagram of the point absorber with its hydraulic PTO.

waves.

2. Optimal Reactive Control of a Point Absorber

2.1. System Description

The selected point absorber features a hydraulic PTO system as shown in Fig. 1, as envisioned by [21]. The mechanical energy associated with the relative motion between the float and the reaction plate is converted into electrical energy through a hydraulic stage. The advantages of a hydraulic PTO unit, whose design is inspired by [24, 25, 26], are its robustness, capacity for energy storage and speed control. Furthermore, no expensive, fully-rated power converters are necessary because through the PTO system it is possible to control the output current [26].

The point absorber comprises of two bodies: a float and a reaction plate connected to a vertical spar. The wave excitation causes the float and reaction plate to move. However, the oscillations of the reaction plate present a much lower magnitude than the float because of the higher inertia, viscous drag and depth of the plate. Hence, the motion difference is used to drive a two-way, single-degree-of-freedom ram that pumps high-pressure oil into the circuit. A rectifying valve prevents flow reversal. Furthermore, the flow is smoothed out through a gas accumulator system. In the reference model 3 [21], this comprises of four high-pressure (HP) cylinders and a low-pressure reservoir, designed to prevent cavitation [26]. The flow drives a hydraulic motor, which is connected to an induction generator. The produced electrical power is fed into the national grid after the voltage is stepped up through a transformer.

As can be seen from Fig. 1, the input variables to the controller are the generated power, P, the displacement and velocity at the PTO, x_{PTO} and $\dot{x}_{\mathrm{PTO}}$, respectively, and the wave elevation, ζ, from which the sea state is derived. The controller then adjusts the flow in the hydraulic circuit by opening or closing the valves connected to the accumulators. This corresponds to changing the damping and stiffness in the system.

2.2. Optimum Reactive Control

In reactive control, the controller force is calculated as the sum of a damping and a stiffness term [19]:

$$F_{\mathrm{PTO}}(t) = -B_{\mathrm{PTO}}\dot{x}_{\mathrm{PTO}}(t) - C_{\mathrm{PTO}}x_{\mathrm{PTO}}(t), \tag{1}$$

where x_{PTO} is the displacement at the PTO. It is assumed that the PTO damping and stiffness coefficients, B_{PTO} and C_{PTO}, respectively, can be modified by changing the pressure within the hydraulic circuit. By varying B_{PTO} and C_{PTO} directly, the developed algorithm can be easily applied also to other PTO systems such as electromechanical or direct-drive.

In reality, the PTO force is saturated with a limit F_{Max} due to the generator rating, as shown in Fig. 6 in Sec. 4.1. The generated power P can be calculated as:

$$P\mathrm{gen}(t) = -F_{\mathrm{PTO}}(t)\dot{x}_{\mathrm{PTO}}(t), \tag{2}$$

and the power fed into the grid is given by:

$$P(t) = \begin{cases} \eta P_{\mathrm{gen}} & \text{if } P_{\mathrm{gen}} \geq 0 \\ P_{\mathrm{gen}}/\eta & \text{if } P_{\mathrm{gen}} < 0 \end{cases}. \tag{3}$$

For simplicity, in Eq. 3 a single measure is employed for the overall efficiency of the PTO unit: η=80% [21]. In Eq. 2, P_{gen} is the generated power. From Eq. 3 it is clear that with reactive control not only is power extracted from the waves, but during part of the wave cycle it is also fed into the environment in order to increase the motions of the device through resonance and thus increase energy absorption [3]. From this behaviour comes the name of the algorithm "reactive control".

The optimal PTO damping and stiffness coefficients that result in maximum energy extraction depend on the wave period in regular waves [27] or the energy wave period, T_{e}, in irregular waves. If the force saturation is included, the optimum B_{PTO} and C_{PTO} values become also functions of the significant wave height, H_{s}. Similarly, the maximum displacement at the PTO, which is of interest to prevent failures in extreme waves, is also a function of the sea state, given by H_{s} and T_{e}.

The state-of-the-art optimum reactive control algorithm employs a tabular approach, where the optimal PTO damping and stiffness coefficients are stored in a table for the main sea states that are encountered at the operational site, given by the combinations of a number of discrete values of B_{PTO} and C_{PTO}. During the operation of the WEC, the controller tries to achieve the prescribed PTO stiffness and damping coefficients in the current sea state through the hydraulic PTO system. The optimal coefficients are usually precalculated using an optimization algorithm, such as the Nelder-Mead simplex algorithm as in [19], with a time-domain hydrodynamic model. For this reason, this technique can be affected by modelling errors and it cannot account for changes in the device response with time, e.g. due to ageing or marine biofouling.

3. Reinforcement Learning Control

In reinforcement learning, the controller learns an optimal behaviour, or *policy*, from direct interaction with the *environment*. In this work, the online, off-policy Q-learning algorithm [28] is selected as in [20]. With this strategy, at each time step n the *agent*, which is in a specific *state* s_n, selects an action a_n. As a result of the interaction with the surrounding environment, the controller lands in a new state, s_{n+1}, while observing a *reward*, r_{n+1}, which depends on the outcome of the chosen action. The action selection, modelled as a Markov decision process, depends on the *value function*, which is a measure of the expected future reward. By considering present as well as future rewards, RL is able to learn the optimal policy with time for the maximization of the total reward [28].

Model-free RL techniques employ the action-value table $\boldsymbol{Q}$, which presents an entry for every combination of discrete states and actions. For instance, $\boldsymbol{Q}(s_n, a_n)$ represents the action-value for the current state and action. The one-step update of the Q-learning algorithm is given by [28]:

$$\boldsymbol{Q}_{n+1}(s_n, a_n) = \boldsymbol{Q}_n(s_n, a_n) + \alpha_n \left[r_{n+1} + \gamma \max_{a' \in A} \boldsymbol{Q}_n(s_{n+1}, a') - \boldsymbol{Q}_n(s_n, a_n) \right], \tag{4}$$

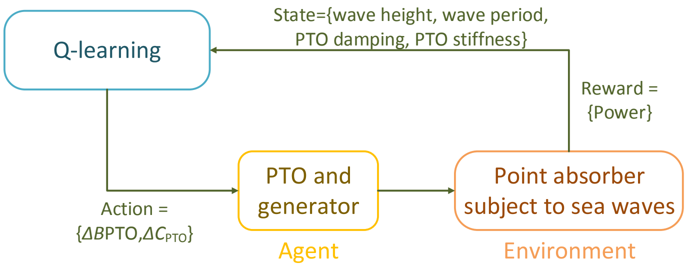


Figure 2: Block diagram of the RL control of the point absorber.

where α_n is defined as the *learning rate*, which determines the proportion of previous learning that is retained in the update of the action-value table, and γ_n is the *discount factor*, which can be used to stress either current or future rewards.

3.1. Application to the Reactive Control of Wave Energy Converters

Fig. 2 shows how Q-learning is used to learn the optimal combination of PTO damping and stiffness coefficients in each sea state without relying on any internal models of the device dynamics. At each step of the algorithm, the controller selects a step change in the coefficients (action), which is implemented by the PTO unit (agent). After interaction with the waves (environment), the controller receives a reward, which is a function of the generated power, and moves to a new state, as given by the significant wave height, the energy wave period, and the PTO damping and stiffness coefficients.

The generated power must be averaged over multiple wave cycles so as to ensure transient effects from changes in B_{PTO} and C_{PTO} do not affect the learning process. In particular, a longer time is required in irregular waves due to their random nature. Hence, the averaging is performed over a time horizon, H, during which the state s_n and action an are constant. As a result, the time steps of the Q-learning algorithm now have length H. As a new action is selected, there is an immediate change of state to s_{n+1} and a new averaging process.

3.1.1. State Space

As aforementioned, the selected state variables are the significant wave height, the energy wave period, and the PTO damping and stiffness coefficients. Hence, the adopted RL state space is given by:

$$S = \left\{ s | s_{i,j,k,l} = (H_{\mathrm{s},i}, T_{\mathrm{e},j}, B_{\mathrm{PTO},k}, B_{\mathrm{PTO},l}), \begin{array}{c} i = 1:I, \\ j = 1:J, \\ k = 1:K, \\ l = 1:L \end{array} \right\}. \tag{5}$$

The choice of I, J, K, and L is based on a compromise between avoiding slow convergence associated with large values and ensuring sufficient learning accuracy, which may be affected by small values. In particular, due to the extra state variable as compared with resistive control [20] the learning time can become an issue with reactive control if large values are selected. I and J are usually determined by the wave resource at the deployment site. Typical ranges of the significant wave height and energy wave period are $H_{\mathrm{s}} = [0, 9]$ m and $T_{\mathrm{e}} = [5, 14]$ s, in steps of 1 m and 1 s, respectively [29]. With a hydraulic PTO system, K and L are set by the number of accumulators.

3.1.2. Action Space

For reactive control, the action is a combination of increase, decrease, or not change the PTO damping and stiffness coefficients. This gives 9 possible actions as opposed to only 3 in the case of resistive control [20]. It has been preferred, however, to vary only one variable at a time in order to limit the action-state space thus decreasing the size of the Q-table. This has a direct consequence on the overall learning time. The action space A is now given by:

$$A = \{a| \, [(-\Delta B_{\mathrm{PTO}}, 0), (0, -\Delta C_{\mathrm{PTO}}), (0, 0), (+\Delta B_{\mathrm{PTO}}, 0), (0, +\Delta C_{\mathrm{PTO}})]\}, \tag{6}$$

where ΔB_{PTO} and ΔC_{PTO} are predefined step changes in the PTO damping and stiffness coefficients respectively.

The states corresponding to the minimum or maximum PTO damping and stiffness coefficients, i.e. $B_{\mathrm{PTO},1}$, $B_{\mathrm{PTO},K}$, $C_{\mathrm{PTO},1}$ and $C_{\mathrm{PTO},L}$, present a smaller action state to prevent the controller from exceeding the state space boundary. For instance, for $C_{\mathrm{PTO},L}$, the action ΔC_{PTO} is invalid.

3.1.3. Reward

In this work, the same reward function, which represents the goal the controller needs to maximise, as in [20] is used. As shown in Fig. 2, the reward is dependent on the absorbed power. Nevertheless, the significant wave height can have stronger influence on the mean generated power, P_{avg}, than variations in B_{PTO} and C_{PTO}. As a result, $P_{\mathrm{avg}}/H_{\mathrm{s}}^2$ is used instead because the absorbed power is proportional to the square of the significant wave height [29]. Additionally, in order to help the learning process by filtering out the noise associated with random seas, the reward function is in fact based on the mean value of a number M of P_{avg} values (which are themselves time-averaged) for each RL state. This is necessary because of the discretization of the state variables and the stochastic nature of irregular waves. Hence, the M most recent $P_{\mathrm{avg}}/H_{\mathrm{s}}^2$ values are stored for each RL state in a matrix, $\boldsymbol{R}$, which presents at most $n_s \cdot M$ entries so as to prevent memory issues, where n_s is the total number of states, $n_s = I \cdot J \cdot K \cdot L$. Thus, the mean value corresponding to each state can be calculated and then expressed with the vector $\boldsymbol{m} = \langle \boldsymbol{R}(s, m) \rangle_{m=1:(M \vee \mathrm{end})}$ of size n_s, with $\langle\rangle$ indicating the averaging process. In this vector, the states are arranged with a vectorised version of Eq. (5) so that the discrete values of B_{PTO} correspond to the innermost loop, C_{PTO} the inner middle loop, T_{e} the outer middle loop and H_{s} the outermost loop.

In order to speed up the learning process, it is advantageous to present a cost function that is equal to one at the optimum and zero everywhere else. This is achieved by first normalizing the values of $\boldsymbol{m}$ by the maximum in each sea state, i.e. for the same H_{s} and T_{e}. This means that the maximum value is searched between the indices $o = \mathrm{floor}((s_n - 1)/(K \cdot L)) \cdot K \cdot L + 1$ and $p = \mathrm{floor}((s_n - 1)/(K \cdot L)) \cdot K \cdot L + K \cdot L$ of the vector $\boldsymbol{m}$. Then, the normalized values should be raised to a very high power, $u = 25$, so that the optimum will present a value of one and the other terms will tend to zero. This process is necessary because the location of the optimum is unknown, and results in the algorithm giving greater importance to the optimum over suboptimal PTO coefficients, even if they result in mean generated power with only a slightly smaller magnitude. Fig. 3 enables the user to fully understand this point, which is of primary importance in the derivation of a suitable reward function for the control of WECs. As an example in Fig. 3, the reward function is assumed to be given by a Weibull distribution [29] with scale and shape parameters 0.6 and 1.5 respectively, whose values are normalized. From Fig. 3, it is clear that a greater value of u corresponds to a more pronounced peakiness. However, a plateau is reached for large values.

Additionally, with reactive control, negative mean power values are possible, which may present a magnitude greater than the maximum power by which the corresponding value in $\boldsymbol{m}$ is normalized. In this case, it is best not

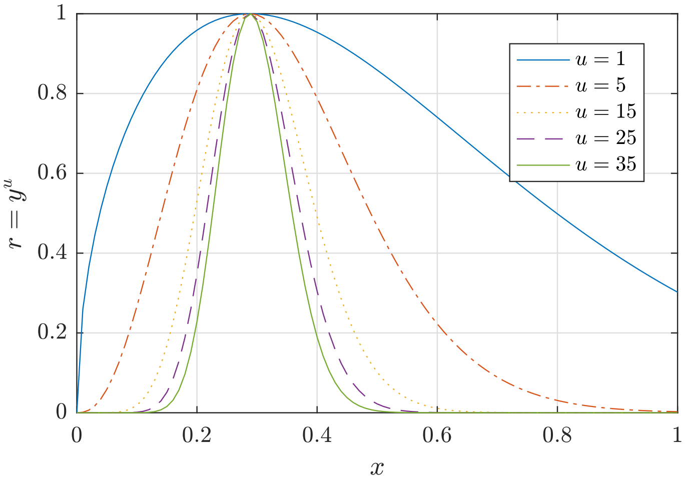
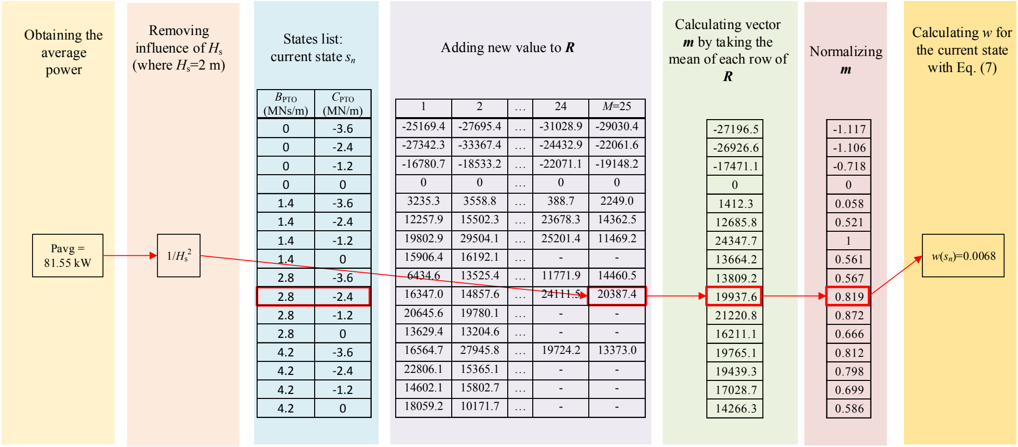


Figure 3: Influence of u on the peakiness of the reward function, based on the example of a normalized Weibull distribution with scale and shape parameters 0.6 and 1.5 respectively.

to raise them to a power, so that a preliminary reward function is given by:

$$w\left(s_n\right) = \begin{cases} \left[\frac{\langle \boldsymbol{m}(s_n)\rangle}{\max_{s=o:p}\langle \boldsymbol{m}(s)\rangle}\right]^u & \text{if } \boldsymbol{m}\left(s_n\right) > 0 \\ \frac{\langle \boldsymbol{m}(s_n)\rangle}{\max_{s=o:p}\langle \boldsymbol{m}(s)\rangle} & \text{if } \boldsymbol{m}\left(s_n\right) \leq 0 \end{cases}. \tag{7}$$

For greater clarity, the calculation of the reward function is shown graphically in Fig. 4 for the final step of the RL algorithm using the simulation in Fig. 10 in Sec. 5.2. Looking at the table $\boldsymbol{R}$, it is possible to make two observations. Firstly, despite an 8-hour-long wave trace being analysed, not all states (i.e. rows of the table) present fully M entries, which means they have been encountered for less than M times (with $M = 25$ in this simulation). Secondly, even for each state, the values of $P_{\mathrm{avg}}/H_{\mathrm{s}}^2$ can present a wide range due to the variation in wave energy for the same discrete sea state. This is the main reason behind selecting a relatively large value of M, which should result in outliers playing a minor role in the calculation of the vector of the mean values, $\boldsymbol{m}$. In Fig. 4, only one sea state is used, so that all entries of $\boldsymbol{m}$ are normalized with respect to the maximum power. However, if more sea states were present, it would be sufficient to update the portion of $\boldsymbol{m}$ corresponding to the current sea state only as defined by indices o and p in Eq. (7). Finally, Fig. 4 shows that the use of a high value of the power u, where $u = 25$ is used in this case, results in a smaller reward being associated


Obtaining the average power
Pavg = 81.55 kW
Removing influence of H_s (where H_s=2 m)
$1/H_s^2$
States list: current state s_n
Adding new value to $\boldsymbol{R}$
Calculating vector $\boldsymbol{m}$ by taking the mean of each row of $\boldsymbol{R}$
Normalizing $\boldsymbol{m}$
Calculating w for the current state with Eq. (7)
$w(s_n)$=0.0068

B_{PTO} (MNs/m)	C_{PTO} (MN/m)
0	-3.6
0	-2.4
0	-1.2
0	0
1.4	-3.6
1.4	-2.4
1.4	-1.2
1.4	0
2.8	-3.6
2.8	-2.4
2.8	-1.2
2.8	0
4.2	-3.6
4.2	-2.4
4.2	-1.2
4.2	0

1	2	…	24	M=25
-25169.4	-27695.4	…	-31028.9	-29030.4
-27342.3	-33367.4	…	-24432.9	-22061.6
-16780.7	-18533.2	…	-22071.1	-19148.2
0	0	…	0	0
3235.3	3558.8	…	388.7	2249.0
12257.9	15502.3	…	23678.3	14362.5
19802.9	29504.1	…	25201.4	11469.2
15906.4	16192.1	…	-	-
6434.6	13525.4	…	11771.9	14460.5
16347.0	14857.6	…	24111.5	20387.4
20645.6	19780.1	…	-	-
13629.4	13204.6	…	-	-
16564.7	27945.8	…	19724.2	13373.0
22806.1	15365.1	…	-	-
14602.1	15802.7	…	-	-
18059.2	10171.7	…	-	-

-27196.5	-1.117
-26926.6	-1.106
-17471.1	-0.718
0	0
1412.3	0.058
12685.8	0.521
24347.7	1
13664.2	0.561
13809.2	0.567
19937.6	0.819
21220.8	0.872
16211.1	0.666
19765.1	0.812
19439.3	0.798
17028.7	0.699
14266.3	0.586

Figure 4: Calculating the reward w (excluding penalties for large motions) at one step of the RL algorithm in irregular waves. This corresponds to the last step in 10 in Sec. 5.2.

with suboptimal combinations of the PTO damping and stiffness coefficients, as expected from Fig. 3.

Furthermore, some combinations of PTO damping and stiffness coefficients may result in large motions in extreme waves that may lead to failure. For this reason, a penalty, -2, is returned whenever the magnitude of the maximum displacement at the PTO exceeds a set value, $x_{\mathrm{PTO,Max}}$. Hence, the complete reward function is given by:

$$r_{n+1} = \begin{cases} w\left(s_n\right) & \text{if } |\max\left(x_{\mathrm{PTO}}\right)| \leq x_{\mathrm{PTO,Max}} \\ -2 & \text{if } |\max\left(x_{\mathrm{PTO}}\right)| > x_{\mathrm{PTO,Max}} \end{cases}. \tag{8}$$

3.1.4. Exploration Strategy, Learning Rate and Discount Factor

Particularly at the start of the learning process, it is advantageous for the agent to try unseen actions in new states, also known as exploration. As the learning progresses, the controller can shift towards the selection of actions that result in greater reward (exploitation), since there is greater confidence in their values. In this work, this has been achieved through an -greedy exploration strategy, which results in the following action selection at each step of the Q-learning algorithm [28]:

$$a_n = \begin{cases} \arg\max_{a' \in A} \boldsymbol{Q}_n(s_n, a') & \text{with probability } 1 - \epsilon_n \\ \text{random action} & \text{with probability } \epsilon_n \end{cases}. \tag{9}$$

In order to ensure exploration at the start and then shift the focus to exploitation, the exploration rate n is calculated as:

$$\epsilon_n = \begin{cases} \epsilon_0 & \text{if } N \leq 0 \\ \epsilon_0/\sqrt{N} & \text{if } N > 0 \end{cases}, \tag{10}$$

where $N = \sum_{i=1:5} \boldsymbol{N}_n(s_n, a_i) - N_{\min\epsilon}$, with $n_a = 5$ indicating the number of actions. $\boldsymbol{N}$ is the matrix containing the count of the number of visits to each state action pair, $N_{\min\epsilon} = 25$ is the minimum number of visits to each state for an initial random exploration, and $\epsilon_0 = 0.6$ is the initial exploration rate.

Similarly, the learning rate α_n should also decrease as the learning goes on. Nevertheless, a slower decay is sought in order to ensure the controller keeps on updating the Q-table throughout the exploration stage:

$$\alpha_n = \begin{cases} \alpha_0 & \text{if } \boldsymbol{N}_n(s_n, a_n) \leq N_{\min\alpha} \\ \alpha_0/\boldsymbol{N}_n(s_n, a_n) & \text{if } \boldsymbol{N}_n(s_n, a_n) > N_{\min\alpha} \end{cases}. \tag{11}$$

In this article, $\alpha_0 = 0.4$ and $N_{\min\alpha} = 5$. These values have been selected based on previous experience with resistive control [20], with less exploration being allowed to speed up convergence. The learning and exploration rates should be reset on a predefined, regular basis so as to account for changes in the WEC dynamics over time, e.g. due to marine growth or non-critical subsystem failure.

Furthermore, a discount factor $\gamma = 0.95$ is employed. This is used to discount only slightly the future rewards the Q-learning algorithm receives.

3.2. Algorithm

The proposed Q-learning algorithm for the reactive control of WECs can be seen in Fig. 5. The first step consists of the initialization of all variables. $\boldsymbol{Q}$ and $\boldsymbol{N}$ are matrices of dimensions $n_s \times n_a$. $\boldsymbol{R}$ is a vector of vectors, whose dimensions are at most $n_s \times M$, with $M = 10$ in regular waves and $M = 25$ in irregular waves. As in [20], the entries of $\boldsymbol{R}$ are pre-calculated in a run in a similar wave trace, whilst taking random actions. It is expected that $\boldsymbol{R}$ will be pre-initialized using simulations also for the full-scale device: as the WEC begins to operate, $\boldsymbol{R}$ will be updated using actual sensor data. In addition, since some combinations of PTO damping and stiffness coefficients can result in very large motions, it is necessary to initialize the Q-table during a pre-training stage using simulations in order to prevent failure in extreme waves.

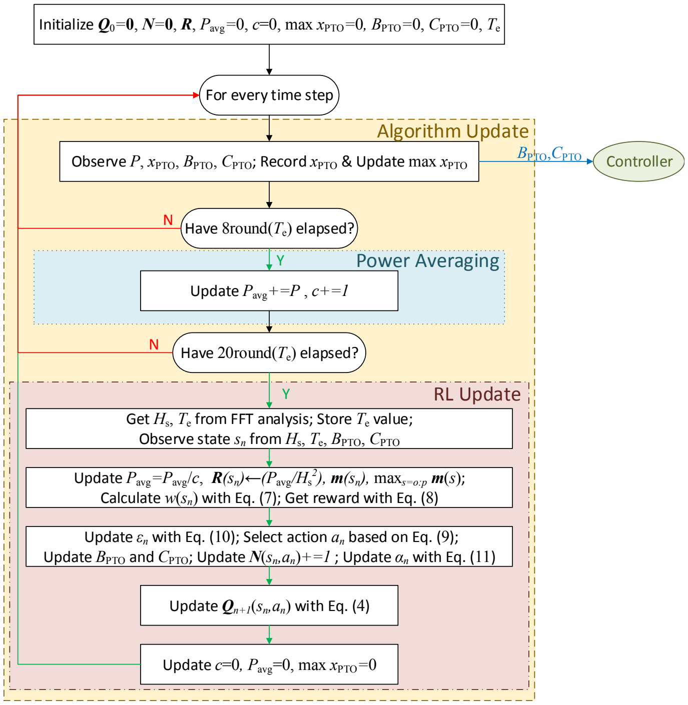

Initialize $\boldsymbol{Q}_0$=**0**, $\boldsymbol{N}$=**0**, $\boldsymbol{R}$, P_{avg}=0, c=0, max x_{PTO}=0, B_{PTO}=0, C_{PTO}=0, T_{e}
For every time step
Algorithm Update
Observe P, x_{PTO}, B_{PTO}, C_{PTO}; Record x_{PTO} & Update max x_{PTO}
B_{PTO},C_{PTO}
Controller
N
Have 8round(T_{e}) elapsed?
Y
Power Averaging
Update P_{avg}+=P , c+=1
N
Have 20round(T_{e}) elapsed?
Y
RL Update
Get H_{s}, T_{e} from FFT analysis; Store T_{e} value;
Observe state s_n from H_{s}, T_{e}, B_{PTO}, C_{PTO}
Update P_{avg}=P_{avg}/c, $\boldsymbol{R}(s_n)\leftarrow(P_{\text{avg}}/H_{\text{s}}^2)$, $\boldsymbol{m}(s_n)$, $\max_{s=o:p} \boldsymbol{m}(s)$;
Calculate $w(s_n)$ with Eq. (7); Get reward with Eq. (8)
Update ε_n with Eq. (10); Select action a_n based on Eq. (9);
Update B_{PTO} and C_{PTO}; Update $\boldsymbol{N}(s_n,a_n)$+=1 ; Update α_n with Eq. (11)
Update $\boldsymbol{Q}_{n+1}(s_n,a_n)$ with Eq. (4)
Update c=0, P_{avg}=0, max x_{PTO}=0

Figure 5: Flowchart of the Q-learning algorithm for the reactive control of WECs.

After the initialization stage, the algorithm is run indefinitely until maintenance is due. At every time step, the selected PTO damping and stiffness coefficients are implemented by the controller through the PTO system. Furthermore, the generated power and the displacement at the PTO are sampled in order to update respectively the mean absorbed power and the maximum displacement value in each time horizon. In particular, the power averaging is performed only after 8round(T_e) have passed in order to remove transient effects due to change in B_PTO or C_PTO. A longer time is required than for resistive control in [20], since a change in PTO stiffness coefficient can cause large motions. Additionally, the time horizon lasts 20round(T_e) in both regular and irregular waves. This results in a speed-up in convergence as compared with 30round(T_z) in [20], whilst still ensuring the algorithm is stable.

From Fig. 5, the Q-learning update at the end of each episode can be seen. The values of the significant wave height and energy wave period are computed using spectral analysis and Fast Fourier Transforms (FFT) from the record of the wave elevation with a unidirectional wave spectrum for simplicity [29].

4. Simulation System

4.1. Hydrodynamic Model

A representation of the WEC analysed in this work can be found in [21, 22, 23]. Assuming linear wave theory and small body motions, the response of the device can be obtained from the superposition of the inertial, hydrostatic, viscous, radiation, diffraction and incident forces in addition to the control force [30]. The two-body problem can be thus modelled with a twelve-degree-of-freedom model. However, by considering only planar motion, i.e. surge, heave and pitch, and the axisymmetric geometry of both float and reaction plate, which means heave is decoupled from surge and pitch [30], it is possible to simplify the model to the coupled heave degrees of freedom of the two bodies. Therefore, using Cummins' formulation for the radiation force [31], it is possible to express the equations of motion of the device in the time domain with the following matrix notation:

$$(\boldsymbol{M} + \boldsymbol{A}(\infty))\,\ddot{\boldsymbol{x}}(t) + \int_0^t \boldsymbol{K}(t-\tau)\dot{\boldsymbol{x}}(\tau)\mathrm{d}\tau + \boldsymbol{C}\boldsymbol{x}(t) = \boldsymbol{f}_\mathrm{ex}(t) + \boldsymbol{f}_\mathrm{PTO}(t) + \boldsymbol{f}_\mathrm{v}(t). \tag{12}$$

$\boldsymbol{M}$ is the inertia matrix, which can be obtained using the data in [22], and $\boldsymbol{C}$ the stiffness matrix. The calculation of the heave hydrostatic stiffness for the float is standard [30], whose dimensions can be found in [22], with the sea water density $\rho = 1025$ kg/m^3 and the gravitational acceleration $g = 9.81$ m/s^2. The reaction plate and spar do not present any hydrostatic stiffness because they are fully submerged. Nevertheless, a stiffness term of 10 MN/m, which is likely to be provided by the mooring system, is specified in order to prevent an unstable behaviour with reactive control.

In Eq. (12), $\boldsymbol{K}$ is the radiation impulse response function matrix, and $\boldsymbol{A}(\infty)$ the added mass matrix at infinite wave frequency. These variables can be computed using the commercial program WAMIT, where the geometry is created following the dimensions in [22]. In particular, panels are included at the waterline within the float contour so as to remove the effects of irregular frequencies [32]. Furthermore, the bottom is left with a hole where the top of the spar fits. Similarly, the top of the spar is left without panels. Care has been taken in ensuring there is a match in the position of the points lying on the inner border of the bottom of the float and on the outer border of the top of the spar to prevent errors in the solution. This arrangement results in incorrect volume and hydrostatic calculations, but in an accurate computation of the radiation and diffraction coefficients. In addition, the use of dipoles on the reaction plate has been found to result in instabilities in the radiation approximation, described hereafter. For this reason, it has been preferred to model the reaction plate as a thicker plate, with a thickness of 3 m (1/10th of the diameter [22]). This approximation has been found to have only a minor effect on the radiation coefficients in heave.

In Eq. (12), $\boldsymbol{f}_{\text{ex}}$ is the excitation force vector, which is calculated from the convolution of the excitation impulse response function, also obtained via WAMIT, and the wave elevation as described in [30]. According to Newton's 3$^{\text{rd}}$ law of action and reaction, $\boldsymbol{f}_{\text{PTO}} = [F_{\text{PTO}}, -F_{\text{PTO}}]^T$ is the controller force vector, with F_{PTO} being computed as in Eq. (1). For this simple case, the displacement and velocity at the PTO can be obtained from the difference in the displacement and velocity of the two bodies: $x_{\text{PTO}}(t) = \boldsymbol{x}_3(t) - \boldsymbol{x}_9(t)$, where 3 and 9 indicate the heave degree of freedom of the float and reaction plate respectively according to standard practice. The viscous drag force, $\boldsymbol{f}_{\text{v}}$, can be calculated with Morison's equation [33]. While no drag force has been modelled on the water-piercing float, its contribution is expected to be non-negligible on the motions of the reaction plate. Since the magnitude of the velocity of the reaction plate is relatively small in all sea states analysed

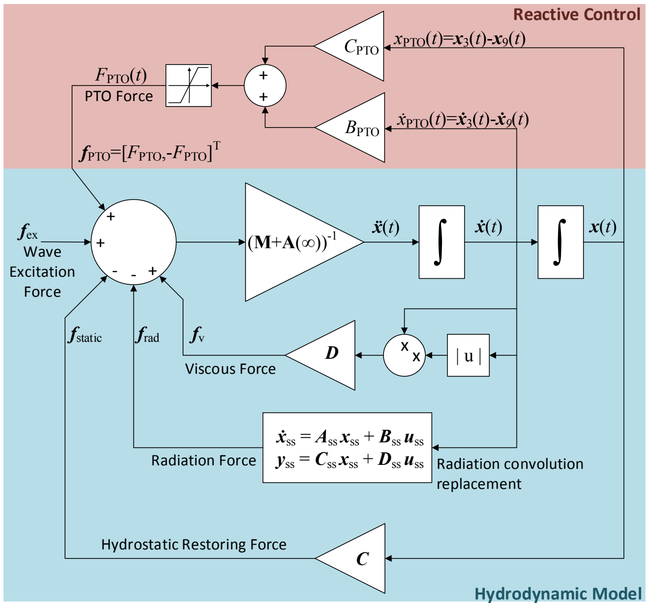

Reactive Control
C_{PTO}
$x_{\mathrm{PTO}}(t)=\boldsymbol{x}_3(t)-\boldsymbol{x}_9(t)$
$F_{\mathrm{PTO}}(t)$
PTO Force
+
+
B_{PTO}
$\dot{x}_{\mathrm{PTO}}(t)=\dot{\boldsymbol{x}}_3(t)-\dot{\boldsymbol{x}}_9(t)$
$\boldsymbol{f}_{\mathrm{PTO}}=[F_{\mathrm{PTO}},-F_{\mathrm{PTO}}]^{\mathrm{T}}$
$\boldsymbol{f}_{\mathrm{ex}}$
Wave
Excitation
Force
+
+
-
-
+
$(\mathbf{M}+\mathbf{A}(\infty))^{-1}$
$\ddot{\boldsymbol{x}}(t)$
$\int$
$\dot{\boldsymbol{x}}(t)$
$\int$
$\boldsymbol{x}(t)$
$\boldsymbol{f}_{\mathrm{static}}$
$\boldsymbol{f}_{\mathrm{rad}}$
$\boldsymbol{f}_{\mathrm{v}}$
Viscous Force
$\boldsymbol{D}$
x
x
| u |
$\dot{\boldsymbol{x}}_{\mathrm{ss}} = \boldsymbol{A}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}} + \boldsymbol{B}_{\mathrm{ss}}\boldsymbol{u}_{\mathrm{ss}}$
$\boldsymbol{y}_{\mathrm{ss}} = \boldsymbol{C}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}} + \boldsymbol{D}_{\mathrm{ss}}\boldsymbol{u}_{\mathrm{ss}}$
Radiation Force
Radiation convolution
replacement
Hydrostatic Restoring Force
$\boldsymbol{C}$
Hydrodynamic Model

Figure 6: Block diagram used for the calculation of the motions of the float and reaction plate.

in this article, a constant drag coefficient $C_{\mathrm{D}} = 5$ is employed, taken from [22].

Fig. 6 shows the expression of Eq. (12) in a block diagram. In order to reduce the computational requirements of the hydrodynamic model, the radiation convolution integral is approximated by a state-space formulation as in [20]. Frequency-domain system identification is employed so as to obtain state-space matrices $\boldsymbol{A}_{\mathrm{ss}}$, $\boldsymbol{B}_{\mathrm{ss}}$, $\boldsymbol{C}_{\mathrm{ss}}$, and $\boldsymbol{D}_{\mathrm{ss}}$ according to the procedure described by [26], with $\boldsymbol{D}_{\mathrm{ss}} = 0$. The matrix $\boldsymbol{D}$ is used to calculate the viscous drag force. All its entries are zero, except for $\boldsymbol{D}_{9,9} = 0.5 C_{\mathrm{D}} \rho \pi R_{\mathrm{plate}}^2$, where $R_{\mathrm{plate}} = 15$ m is the radius of the reaction plate [22]. In addition, the hydrodynamic model in Fig. 6 has been expressed in a discrete state-space format through a first-order hold [34] in order to reduce the computational cost of the solution. The sampling time has been set to $\Delta t = 0.1$ s.

4.2. Simulation Model

Numerical simulations have been run for the Reference Model 3 two-body point absorber, whose dimensions can be found in [22]. The maximum PTO

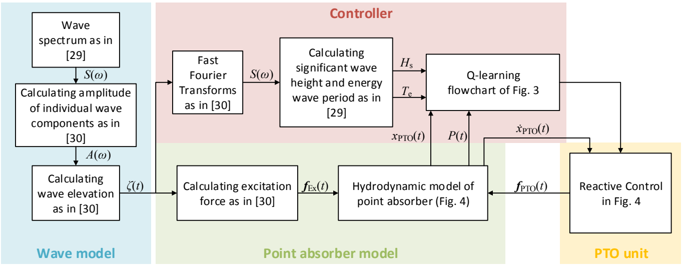

Controller
Wave spectrum as in [29]
$S(\omega)$
Calculating amplitude of individual wave components as in [30]
$A(\omega)$
Calculating wave elevation as in [30]
$\zeta(t)$
Fast Fourier Transforms as in [30]
$S(\omega)$
Calculating significant wave height and energy wave period as in [29]
H_s
T_e
Q-learning flowchart of Fig. 3
$x_{PTO}(t)$
$P(t)$
$\dot{x}_{PTO}(t)$
Calculating excitation force as in [30]
$f_{Ex}(t)$
Hydrodynamic model of point absorber (Fig. 4)
$f_{PTO}(t)$
Reactive Control in Fig. 4
Wave model
Point absorber model
PTO unit

Figure 7: Workflow diagram of the program used to simulate the point absorber.

force that can be exerted due to the generator rating has been assumed to be $F_{\text{Max}} = 1$ MN, while the magnitude of the maximum displacement at the PTO has been limited to $x_{\text{PTO,Max}} = 5$ m.

The program employed for the simulations is summarized in Fig. 7. While a buoy will record the wave elevation in practice as shown in Fig. 1, a wave model is used to generate the wave elevation time series in Fig. 7. On the one hand, the wave elevation is used to obtain H_s and T_e. On the other hand, it is required for the calculation of the wave excitation force through the diffraction convolution integral [30].

In order to generate the wave elevation in irregular waves, the amplitude wave spectrum $S(\omega)$ needs to be specified for a number of circular wave frequencies [29], ω. The individual wave components are superimposed to calculate ζ, each having a wave amplitude $A(\omega) = \sqrt{2S(\omega)\Delta\omega}$, where $\Delta\omega$ is the circular frequency step [30]. $\Delta\omega$ should be selected smaller than the Nyquist frequency in order to prevent a repetition of the wave trace [34]. This is particularly problematic, since it is evident from Sec. 5.2 that very long wave traces are required for the RL algorithm to converge. For this reason, it has been preferred to generate the wave trace as the combination of 15-minute long wave traces, where a different seed for the random number generator is used for each one. Furthermore, a 20-point filter is used over the last and first 20 s of each trace in order to smoothen the connection. Therefore, $\Delta\omega = 0.005$ rad/s has been used, since it meets the Nyquist criterion [34], which has been possible by fitting the diffraction coefficients generated by WAMIT with a high-order polynomial.

For simplicity, the PTO damping coefficient is assumed to range from 0 to 4.2 MNs/m in steps of 1.4 MNs/m, so that $K = 4$. Similarly, the PTO stiffness coefficient is taken to range from -3.6 MN/m to 0 MN/m in steps of 1.2 MN/m, so that $L = 4$. These values have been selected as they fully enclose the optimal coefficients for the analysed sea states. As a result of the choice of PTO damping and stiffness coefficients, 16 RL states are used when a single sea state, as given by H_s and T_e, is considered. Nevertheless, for a more realistic implementation a finer resolution and a wider range are expected.

5. Simulation Results

The learning capabilities of the algorithm are assessed in both regular and irregular waves. Since the same time horizon length has been selected in both cases, the same wave trace length of 8 hours has been employed as opposed to [20], where a longer time series was required in irregular waves. The hydrodynamic model is initialised for 15 minutes to prevent numerical instabilities, although this trace is not reported in the plots. Additionally, the RL response is validated against optimal reactive control, whose coefficients are obtained from Nelder-Mead simplex optimizations [19] in 20-minute-long wave traces.

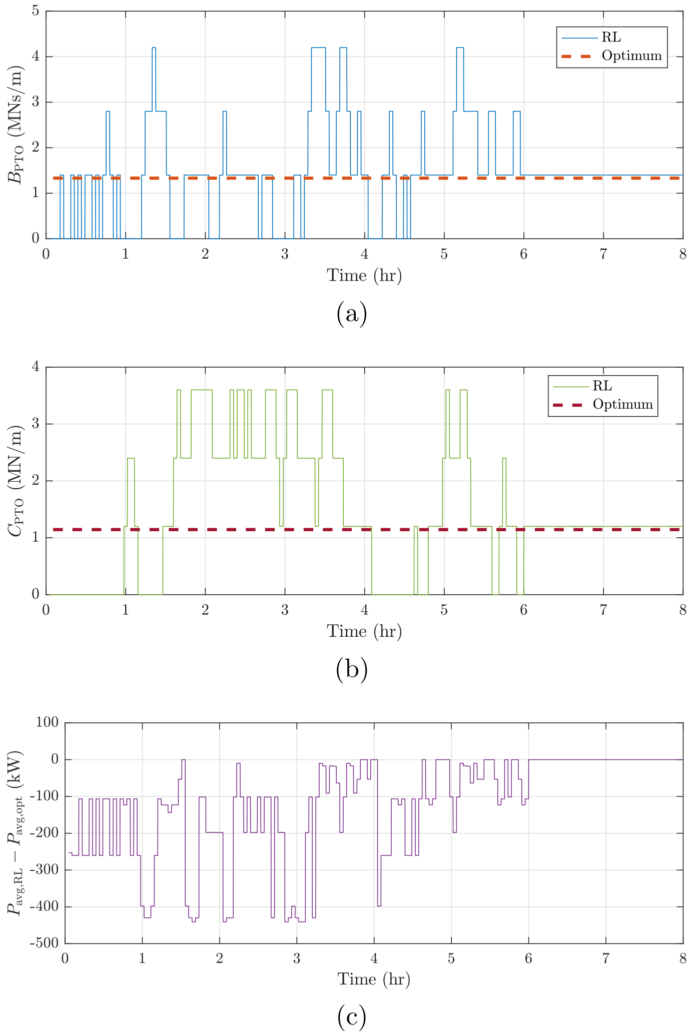
5.1. Regular Waves

A single sea state, i.e. $I = J = 1$, has been analysed in regular waves, with unit amplitude and a wave period of 8 s. Fig. 8a and Fig. 8b compare the curves of the PTO damping and stiffness coefficients respectively with time as selected by the Q-learning algorithm against the optimal values. The difference in the corresponding mean absorbed power and the optimal mean generated power of 260.5 kW can be seen in Fig. 8c.

In addition, the reward function is plotted against the PTO damping and stiffness coefficients in Fig. 9 for the same sea state. In particular, two values have been used for u, the power of the normalized power in Eq. (7). Note that because the displacement limit is not reached, $r = w$ in this case (Eq. (7) and Eq. (8)). The case of $u = 1$ corresponds to purely the normalized mean generated power values, while $u = 25$ is used in the actual cost function in this article.

5.2. Irregular Waves

Similarly, a single sea state, with a significant wave height of 2 m and a peak wave period of 9.25 s is considered in irregular waves as a proof of concept. From the FFT analysis, the energy wave period for the generated



(a)

(b)

(c)

Figure 8: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the RL control as compared with the respective optimal values in regular waves of unit amplitude and a wave period of 8 s. (c) shows the difference between the corresponding mean generated power and the optimal mean generated power.

wave trace is 8 s. As per the regular waves case, $I = J = 1$ so that the RL problem reduces to 16 states.

In Fig. 10a and Fig. 10b, it is possible to see the PTO damping and stiffness coefficients respectively adopted by the RL control scheme as compared with the optimal values in this sea state. Fig. 10c shows the difference

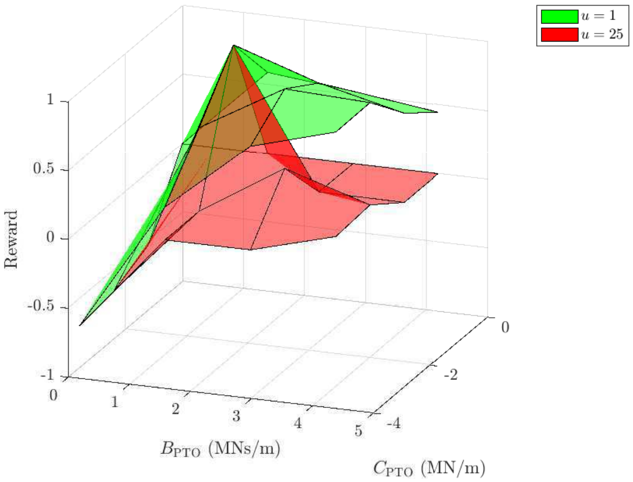


Figure 9: Reward function for all possible configurations of the PTO damping and stiffness coefficient for the device in regular waves with $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s using two values for u.

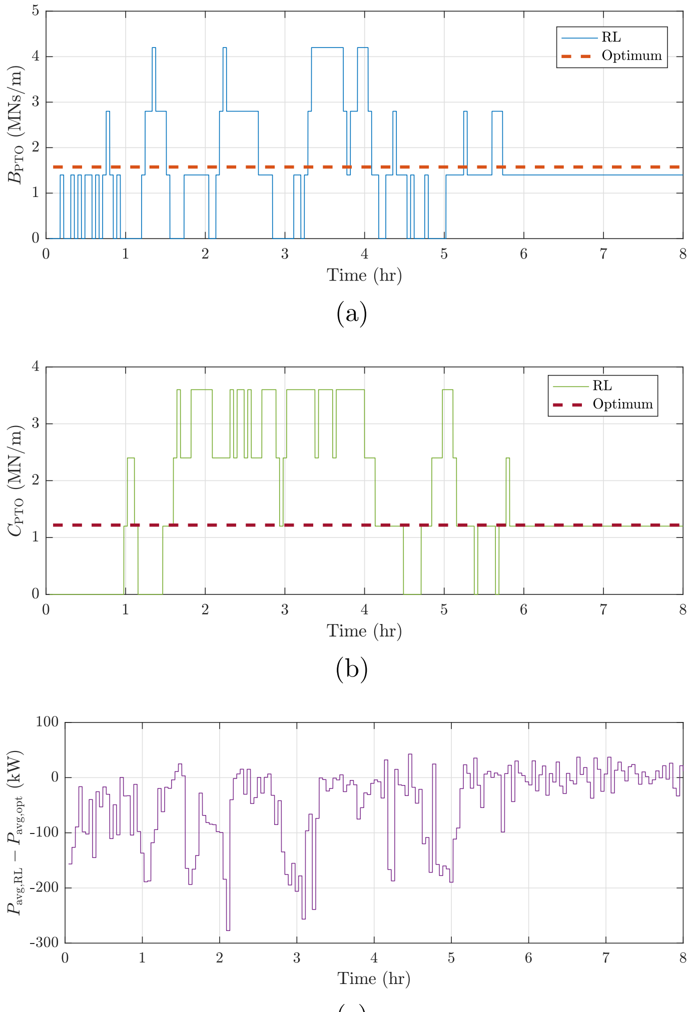
in the corresponding mean absorbed power, with the mean generated power obtained by using the optimal coefficients being 90.582 kW.

6. Discussion

6.1. Regular Waves

As is clear from Fig. 8, in regular waves the Q-learning algorithm learns the optimal PTO coefficients in approximately six hours from a random start ($\boldsymbol{Q} = \mathbf{0}$). This is almost double the time required by the control scheme for resistive control in [20] mainly due to the longer time horizon employed: $20T_{\mathrm{e}}$ as opposed to $10T_{\mathrm{z}}$, with the energy wave period being typically greater than the zero-crossing mean wave period. In fact, a shorter time horizon may be used considering the deterministic nature of regular waves. Additionally, the convergence time is strongly dependent on the number of discrete B_{PTO} and C_{PTO} values employed, with only 16 states currently being used.

In Fig. 8, it is also interesting to notice the random initial behaviour of the controller due to the selected exploration strategy, which enables the agent to visit most states. As the learning progresses, the exploration rate tends to zero and the algorithm chooses the optimal, exploitative actions.



(a)

(b)

(c)

Figure 10: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the RL control as compared with the respective optimal values in irregular waves with $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s. (c) shows the difference between the corresponding mean generated power and the optimal mean generated power.

In order to meet the requirements of the linear wave theory assumption of the hydrodynamic model, a short wave height has been chosen. As a result, the prescribed maximum PTO displacement is never exceeded. Hence, the penalty term in Eq. (8) is not applied. If it were, the controller would be expected to select a higher PTO damping coefficient, as in [20] for resistive control. Conversely, a PTO stiffness coefficient with a smaller, if not zero, magnitude is forecast, as the controller tries to move away from resonance. On the other hand, the force reaches the saturation limit even in this mild sea state. However, a bang-bang behaviour similar to the one in [20] is not observed with reactive control.

In Fig. 9, it is interesting to compare the developed cost function ($u = 25$) with the original normalized generated mean power surface ($u = 1$) for the same sea state. As can be seen, raising the non-dimensional power values to a high power results in a much peakier reward function, similar to what is observed in Fig. 3 for an example function. This is highly desirable, since it enables the controller to learn more quickly what the optimal action is, as there is a more significant gain associated with it as compared with suboptimal solutions. This results in a considerable speed-up in the convergence time as opposed to the case of $u = 1$. Even higher values of the power u may be required for a finer mesh of PTO damping and stiffness coefficients, since this can present a flatter region around the optimum. As aforementioned, this approach is necessary because the actual position of the optimum is unknown, with the best reward function in terms of convergence time being the one that presenting a value of +1 at the optimum and 0 everywhere else.

It is important to notice that raising negative normalized mean generated power values to a high value of u is strongly undesirable. This would have the effect of decreasing the magnitude of the reward as for positive power values, but in this case it would actually mean increasing the reward associated with suboptimal points. It would be even worse to use an even value for u, since it would turn negative mean generated power values into positive ones, thus teaching the controller a completely wrong policy. Hence, positive values of u for negative normalized mean generated power values must be avoided at all costs.

6.2. Irregular Waves

From Fig. 10, it is evident that the developed statistical reward function is effective in ensuring convergence in irregular waves as well, despite their stochastic nature. Furthermore, since the same horizon time length is employed as per the regular waves run, the learning time is no greater as opposed to the study by [20]. Nevertheless, the challenge that irregular waves pose to the convergence of the correct action selection can be understood by comparing Fig. 8c and Fig. 10c, where the much more oscillatory nature of the mean absorbed power in irregular waves is clear.

A typical sea state has a duration that ranges between 30 minutes and 6 hours [29]. Hence, even though the learning time is smaller than in [20] despite the larger number of states, convergence is still unlikely to be achieved before there is a variation in the significant wave height and energy wave period. However, as shown in [20] for irregular waves with multiple sea states, the Q-learning algorithm applied to the control of WECs is able to pick up the learning process from where it left off the last time it encountered a particular sea state. This represents the main advantage of reinforcement learning over traditional optimization algorithms, which would be unable to identify whether a change in the cost function is due to a change in the PTO damping or stiffness coefficients or due to noise in the wave energy.

In a realistic application, a finer grid of B_{PTO} and C_{PTO} values would be desired in order to deal with a large range of sea states. Nevertheless, this may increase the learning time excessively. The Q-table is expected to be pre-initialized through numerical simulations in order to prevent selecting PTO settings that result in excessive motions in energetic sea states, which could be a real problem with reactive control. In addition, the exploration and learning rates should be reset every season so as to check if there have been variations in the device response over time, e.g. due to slow marine growth or abrupt non-critical subsystems failure. Since the operational life of WEC technologies is envisioned to be 20 to 25 years, a relatively poor performance during the initial stages of operation should be more than offset by increases in the absorbed wave power throughout a devices operating life through the removal of modelling errors.

Finally, it is important to understand that RL is proposed as a method to remove the dependence of existing WEC control strategies from hydrodynamic models. Therefore, the overall controller performance is only as good as the control scheme itself, with reactive control representing a significant improvement over resistive control treated in the previous study [20].

7. Conclusions

The authors have presented an on-line, model free strategy for the reactive control of WECs using RL, building on a previous study on resistive control. The algorithm has been validated through a numerical model of a two-body point absorber which assumes linear wave theory. In both regular and irregular waves the controller is shown to learn the optimal PTO damping and stiffness coefficients that result in maximum energy absorption. In order to achieve convergence in irregular waves, a statistical reward function has been developed, which averages over multiple mean absorbed power values in each sea state. As the control scheme is independent of internal models of the device response, it is simple to implement on a real, full-scale WEC. Additionally, it can adapt to variations in the machine conditions over time, e.g. due to ageing or marine biofouling. Although the Q-table has been randomly initialized, in a real application it is expected to be pre-calculated through simulations in order to prevent the adoption of actions that may cause failures in extreme waves. The action-values will then be slowly substituted by the actual measured data during operation with corresponding necessary adjustments. Finally, this method, which has already been generalised to the application to multi-body devices in this work since a previous study [20], can be further extended to the treatment of arrays of devices.

Acknowledgements

The authors would like to thank the Energy Technologies Institute (ETI) and the Research Councils Energy Programme (RCEP) for funding this research as part of the IDCORE programme (EP/J500847/), as well as the Engineering and Physical Sciences Research Council (EPSRC) (grant EP/J500847/1). In addition, Mr. Anderlini would like to thank Wave Energy Scotland (WES) for sponsoring his Eng.D. research project.

WES is taking an innovative approach to supporting the development of wave energy technology by managing the most extensive technology programme of its kind in the sector, concentrating on key areas which have been identified as having the most potential impact on the long-term levellised cost of energy and improved commercial viability.

References

[1] K. Gunn, C. Stock-Williams, Quantifying the Potential Global Market for Wave Power, Proceedings of the 4th International Conference on Ocean Engineering (ICOE 2012) (2012) 1–7.

[2] A. F. D. O. Falcão, Wave energy utilization: A review of the technologies, Renewable and Sustainable Energy Reviews 14 (3) (2010) 899–918. doi:10.1016/j.rser.2009.11.003.

[3] S. H. Salter, J. R. M. Taylor, N. J. Caldwell, Power conversion mechanisms for wave energy, Proceedings of the I MECH E Part M 216 (1) (2002) 1–27. doi:10.1243/147509002320382112.

[4] J. V. Ringwood, G. Bacelli, F. Fusco, Energy-Maximizing Control of Wave-Energy Converters: The Development of Control System Technology to Optimize Their Operation, IEEE Control Systems Magazine 34 (5) (2014) 30–55.

[5] K. Budal, J. Falnes, Optimum Operation of Wave Power Converter, Marine Science Communications 3 (2) (1977) 133–150.

[6] A. Babarit, G. Duclos, A. H. Clément, Comparison of latching control strategies for a heaving wave energy device in random sea, Applied Ocean Research 26 (5) (2004) 227–238. doi:10.1016/j.apor.2005.05.003.

[7] A. Babarit, A. H. Clément, Optimal latching control of a wave energy device in regular and irregular waves, Applied Ocean Research 28 (2) (2006) 77–91. doi:10.1016/j.apor.2006.05.002.

[8] A. Babarit, M. Guglielmi, A. H. Clément, Declutching control of a wave energy converter, Ocean Engineering 36 (12-13) (2009) 1015–1024. doi:10.1016/j.oceaneng.2009.05.006.

[9] T. K. A. Brekken, On Model Predictive Control for a point absorber Wave Energy Converter, Proceedings of the IEEE Trondheim PowerTech (2011) 1–8doi:10.1109/PTC.2011.6019367.

[10] J. Hals, J. Falnes, T. Moan, Constrained Optimal Control of a Heaving Buoy Wave-Energy Converter, Journal of Offshore Mechanics and Arctic Engineering 133 (1) (2011) 011401. doi:10.1115/1.4001431.

[11] G. Li, M. R. Belmont, Model predictive control of sea wave energy converters - Part I: A convex approach for the case of a single device, Renewable Energy 69 (2014) 453–463. doi:10.1016/j.renene.2014.03.070.

[12] M. Richter, O. Sawodny, M. E. Magaña, T. K. a. Brekken, Power optimisation of a point absorber wave energy converter by means of linear model predictive control, IET Renewable Power Generation 8 (2) (2014) 203–215. doi:10.1049/iet-rpg.2012.0214.

[13] F. Fusco, J. V. Ringwood, A simple and effective real-time controller for wave energy converters, IEEE Transactions on Sustainable Energy 4 (1) (2013) 21–30. doi:10.1109/TSTE.2012.2196717.

[14] M. Richter, M. E. Magana, O. Sawodny, T. K. a. Brekken, Nonlinear Model Predictive Control of a Point Absorber Wave Energy Converter, Sustainable Energy, IEEE Transactions on 4 (1) (2013) 118–126. doi:10.1109/TSTE.2012.2202929.

[15] D. Oetinger, M. E. Magaña, S. Member, O. Sawodny, Decentralized Model Predictive Control for Wave Energy Converter Arrays, Sustainable Energy, IEEE Transactions on 5 (4) (2014) 1099–1107.

[16] G. Li, M. R. Belmont, Model predictive control of sea wave energy converters - Part II: The case of an array of devices, Renewable Energy 68 (2014) 540–549. doi:10.1016/j.renene.2014.02.028.

[17] K. U. Amann, M. E. Magaña, S. Member, O. Sawodny, Model Predictive Control of a Nonlinear 2-Body Point Absorber Wave Energy Converter With Estimated State Feedback, Sustainable Energy, IEEE Transactions on 6 (2) (2015) 336–345.

[18] O. Sawodny, D. Oetinger, M. E. Magaña, Centralised model predictive controller design for wave energy converter arrays, IET Renewable Power Generation 9 (2) (2015) 142–153. doi:10.1049/iet-rpg.2013.0300.

[19] A. J. Nambiar, D. I. M. Forehand, M. M. Kramer, R. H. Hansen, D. M. Ingram, Effects of hydrodynamic interactions and control within a point absorber array on electrical output, International Journal of Marine Energy 9 (2015) 20–40. doi:10.1016/j.ijome.2014.11.002.

[20] E. Anderlini, D. I. M. Forehand, P. Stansell, Q. Xiao, M. Abusara, Control of a Point Absorber using Reinforcement Learning, Transactions on Sustainable Energy 7 (4) (2016) 1681–1690.

[21] V. S. Neary, M. Previsic, R. a. Jepsen, M. J. Lawson, Y.-H. Yu, A. E. Copping, A. a. Fontaine, K. C. Hallett, Methodology for Design and Economic Analysis of Marine Energy Conversion (MEC) Technologies, Tech. Rep. March, Sandia National Laboratories (2014). doi:SAND2014-9040.

[22] M. Previsic, K. Shoele, J. Epler, Validation of Theoretical Performance Results using Wave Tank Testing of Heaving Point Absorber Wave Energy Conversion Device working against a Subsea Reaction Plate, 2nd Marine Energy Technology Symposium (2014) 1–8.

[23] Y. Yu, M. Lawson, Y. Li, M. Previsic, J. Epler, J. Lou, Experimental Wave Tank Test for Reference Model 3 Floating- Point Absorber Wave Energy Converter Project, Tech. Rep. January, National Renewable Energy Laboratory (2015). doi:NREL/TP-5000-62951.

[24] R. Henderson, Design, simulation, and testing of a novel hydraulic power take-off system for the Pelamis wave energy converter, Renewable Energy 31 (2) (2006) 271–283. doi:10.1016/j.renene.2005.08.021.

[25] A. F. D. O. Falcão, Modelling and control of oscillating-body wave energy converters with hydraulic power take-off and gas accumulator, Ocean Engineering 34 (14-15) (2007) 2021–2032. doi:10.1016/j.oceaneng.2007.02.006.

[26] D. Forehand, A. E. Kiprakis, A. Nambiar, R. Wallace, A Bi-directional Wave-to-Wire Model of an Array of Wave Energy Converters, IEEE Transactions on Sustainable Energy 7 (1) (2016) 118–128. doi:10.1109/TSTE.2015.2476960.

[27] E. Tedeschi, M. Carraro, M. Molinas, P. Mattavelli, Effect of control strategies and power take-off efficiency on the power capture from sea waves, IEEE Transactions on Energy Conversion 26 (4) (2011) 1088–1098. doi:10.1109/TEC.2011.2164798.

[28] R. S. Sutton, A. G. Barto, Reinforcement Learning, hardcover Edition, MIT Press, 1998.

[29] L. H. Holthuijsen, Waves in Oceanic and Coastal Waters, Cambridge University Press, 2007.

[30] J. Falnes, Ocean waves and Oscillating systems, paperback Edition, Cambridge University Press, 2005. doi:10.1016/S0029-8018(02)00070-7.

[31] W. E. Cummins, The impulse response function and ship motions, Schiffstechnik 47 (9) (1962) 101–109.

[32] WAMIT, User Manual: Version 7.0, 2013. arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.

[33] J. Morison, J. Johnson, S. Schaaf, The Force Exerted by Surface Waves on Piles, Journal of Petroleum Technology 2 (5) (1950) 149–154. doi:10.2118/950149-G.

[34] G. F. Franklin, J. D. Powell, A. Emami-Naeini, Feedback Control of Dynamic Systems, 6th Edition, Pearson, 2008.