Proceedings of the ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering
OMAE2017
June 25-30, 2017, Trondheim, Norway

OMAE2017-61294

CONSTRAINTS IMPLEMENTATION IN THE APPLICATION OF REINFORCEMENT LEARNING TO THE REACTIVE CONTROL OF A POINT ABSORBER

Enrico Anderlini*

Industrial Doctoral Centre for Offshore Renewable Energy
University of Edinburgh
Edinburgh, UK, EH9 3DW
Email: E.Anderlini@ed.ac.uk

David I. M. Forehand

University of Edinburgh
Edinburgh, UK, EH9 3DW

Elva Bannon

Wave Energy Scotland
Inverness, Scotland, IV2 5NA

Mohammad Abusara

University of Exeter
Penryn, UK, TR10 9FE

ABSTRACT

Here, least-squares policy iteration, a reinforcement learning algorithm, is applied to the reactive control of a wave energy converter for the first time. Simulations of a linear point absorber are used for this analysis. The focus of this study is on the implementation of displacement constraints. The use of a penalty term is effective in teaching the controller to avoid the selection of combinations of the damping and stiffness coefficients that would result in excessive displacements in particular sea states. However, the controller can learn that the actions are bad only after trying them, as shown by the simulations. For this reason, a lower-level control scheme is proposed, which changes the sign of the controller force based on the magnitude of the float displacement and sign of its velocity. Its effectiveness is proven in both regular and irregular waves, although greater care is required for the determination of soft constraints.

*Address all correspondence to this author.

INTRODUCTION

With an estimated resource of up to 2.1 TW worldwide [1], wave energy can significantly contribute to power generation in the future, thus helping us to reduce carbon-dioxide emissions associated with the burning of fossil fuels. Despite numerous technologies having been proposed and developed since the 1970s, wave energy converters (WECs) are not financially viable yet. An area that has been identified as important in order to decrease the levellised cost of energy associated with WECs is the design of an effective control strategy. In particular, a good control scheme should try to optimize energy absorption through the operation of actuators, while at the same time preventing large motions in extreme sea conditions [2], where damages to the device are likelier.

Since the first studies on WEC dynamics, different control strategies have been proposed. Detailed reviews on the classical and more recent advances in the field have been published by [3] and [2], respectively. However, most control schemes developed to date rely on models of the system dynamcis in order to determine a suitable control action. This is problematic, since not

only can modelling errors have a negative impact on energy absorption, but the selection of an incorrect action may also result in failure in energetic waves. For this reason, the authors have proposed the application of machine learning algorithms, developed by the computer science industry, in order to remove the model dependence from state-of-the-art control techniques for WECs. In particular, reinforcement learning (RL) has been applied to the resistive and reactive control of a WEC in [4] and [5], respectively. In these studies, the controller learns the optimal damping and stiffness (zero for resistive control) coefficients in each sea state from direct interaction with the environment using Q-learning, a popular algorithm with the robotics industry [6, 7]. The linear model of a point absorber, an established technology consisting of a floating body whose dimensions are small relative to the characteristic wavelength [8], is used for validation. In addition, a more powerful algorithm, least-squares policy iteration (LSPI) [9], is shown to dramatically reduce the learning time for the resistive control of a non-linear point absorber model in [10].

Although force (or torque) constraints are simple to implement with both resistive and reactive control, displacement constraints have been included within the machine learning schemes by providing a penalty term for large motions. Nevertheless, only [4] presents a study on the effectiveness of this technique.

Here, the implementation of displacement constraints is analysed in detail. In particular, the same linear model of a single-degree-of-freedom point absorber used in [4] is employed for validation. LSPI is adopted due to its superior learning performance and applied to reactive control for the first time. Firstly, the controller behaviour is analysed in both regular and irregular waves when no displacement constraints are imposed. Then, the effectiveness of the penalty term is studied in regular waves. After identifying some issues during the initial learning stage, a simple low-level control scheme that controls the actuator force in real time is implemented and analysed in both regular and irregular waves.

REACTIVE CONTROL OF A POINT ABSORBER

System description

The diagram of a general point absorber is shown in Fig. 1. As shown in the figure, energy associated with the motions of the float is extracted through the PTO system, which can be of hydraulic, electromechanical or electrical type. In turn, the PTO is connected to the grid so as to deliver electricity. In order to select optimal control actions, the controller requires knowledge of the body displacement at the PTO, z, and velocity, $\dot{z}$, the generated power P, as well as the wave elevation, ζ. The wave elevation is used to determine the significant wave height, H_s, and energy wave period, T_e. While the former two variables are inferred from on-board accelerometers, the sea state is typically provided by a separate wave buoy for the whole wave farm. Furthermore, the generated power P is obtained from the PTO system. In turn,

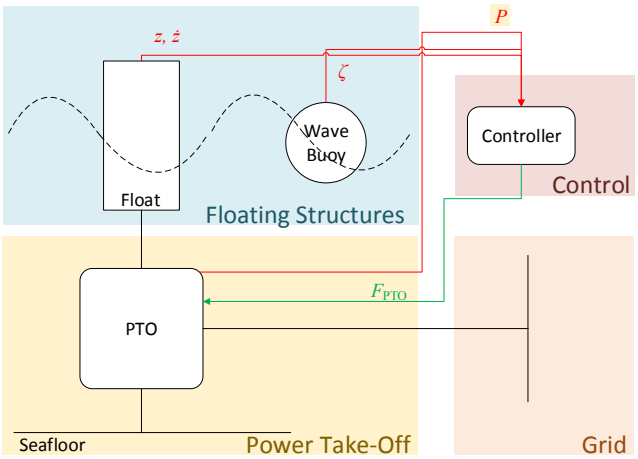

FIGURE 1. Schematic diagram of the WEC.

the controller returns a value of the prescribed PTO force, F_{PTO}, which the PTO must meet.

In the absence of the possibility to perform experimental testing or have access to a prototype, here the system is replaced by a model. For simplicity, a single-body device is analysed, constrained to motions in heave; hence, the problem reduces to a single degree of freedom, indicated by the number 3.

Modelling of the system dynamics

Hydrodynamics The hydrodynamic model of this simple device has been obtained in [4]. Using linear wave theory, it is possible to express the equation of motion of the WEC as shown graphically in Fig. 2. In the figure, M is the float mass, $C_{3,3}$ the hydrostatic stiffness coefficient, $A_{3,3}(\infty)$ the added mass at infinite wave frequency, F_{PTO} the PTO force and F_3 the wave excitation force. The hydrodynamic coefficients are calculated using the panel code WAMIT. In order to speed up simulations, the radiation convolution is approximated with a state-space system using frequency-domain identification as explained in [11].

Excitation force The wave excitation force is computed from the convolution of the wave elevation and the diffraction impulse response function [12]. The latter is computed from WAMIT.

In the simulations, the wave elevation presents an initial ramp function up to a time of 50 s in order to prevent divergence. In irregular waves, ζ is computed as the superposition of multiple individual wave components, whose amplitude is derived from the specified wave spectrum [13]. The circular wave frequency step has been set to 0.005 rad/s, ranging from 0 to 5 rad/s, since this value is smaller than the Nyquist frequency for a 15-minute window so as to prevent a repetition of the wave trace [14]. Thus,

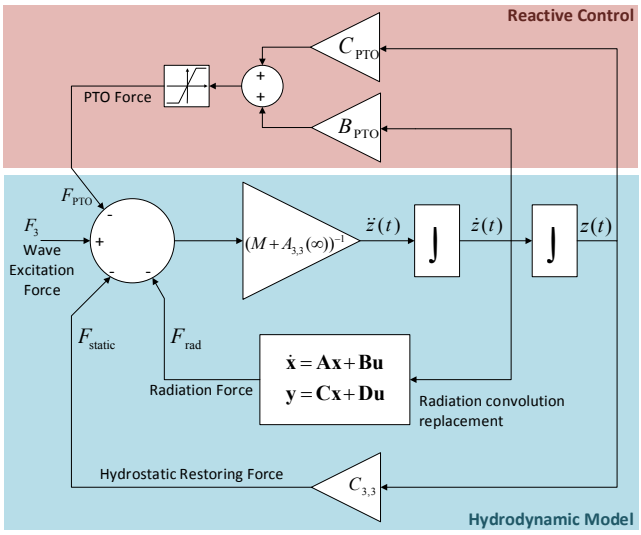

FIGURE 2. Block diagram representing the hydrodynamic model of the point absorber.

each trace of irregular waves is generated as the combination of 15-minute-long time series, where the random number generator is initialized with a different seed for each component. In order to smooth the connection between the separate traces, a 20-point filter is employed over the last and first of each consecutive time series.

Reactive control As can be seen in Fig. 2, the PTO force F_{PTO} is given by the sum of a damping and stiffness term, with B_{PTO} and C_{PTO} being the damping and stiffness coefficients, respectively. Additionally, the force is saturated at $\pm F_{\mathrm{Max}}$ due to the physical limits of the PTO system.

The generated power can be computed as [15]:

$$P(t) = \begin{cases} \eta F_{\mathrm{PTO}}(t)\dot{z}(t) & \text{if } F_{\mathrm{PTO}}(t)\dot{z}(t) > 0, \\ \frac{1}{\eta} F_{\mathrm{PTO}}(t)\dot{z}(t) & \text{otherwise,} \end{cases} \tag{1}$$

where t indicates time and η is the overall PTO efficiency.

State-of-the-art reactive control consists in the calculation of the optimal PTO damping and stiffness coefficients, $B_{\mathrm{PTO,opt}}$ and $C_{\mathrm{PTO,opt}}$, respectively, such that they maximize the energy absorption in each sea state. The sea state is given by the significant wave height, H_{s}, and the energy wave period, T_{e}, in irregular waves and the wave height, H, and period, T, in regular ones. Furthermore, in the optimization, the float displacement must be constrained to safe limits $|z| < z_{\mathrm{Max}}$ so as to avoid structural damage in extreme sea conditions. At the moment, simulations are used to compute $B_{\mathrm{PTO,opt}}$ and $C_{\mathrm{PTO,opt}}$ for a set of discrete sea states, generating a matrix. Once the WEC is operational at sea, the controller selects the coefficients corresponding to the current sea state. However, this technique suffers from modelling errors, and it is not adaptive to changes in the device dynamics.

REINFORCEMENT LEARNING

Within the RL framework [6], an agent, which in a state s, takes an action a. As a result, it receives a reward, r, and lands into a new state, s'. The action selection, modelled as a Markov decision process, is based on the value function $Q(s,a)$, which represents an estimate of the future reward. The behaviour of the controller is referred to as policy, π. The agent learns an optimal policy, π^*, with time for the maximization of the total reward.

Least-Squares Policy Iteration

LSPI is a powerful, off-line, off-policy RL algorithm, originally developed by [9]. An application to resistive control of a WEC is described in [10], and the reader is referred to that text for a more detailed description.

With LSPI, a linear architecture is employed for the approximation of the Q function:

$$Q(s,a) \approx \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}, \tag{2}$$

where $\boldsymbol{\Theta}$ is the weight matrix and $\boldsymbol{\phi}$ is the vector of arbitrary, linearly independent features. The subscript $(:,a)$ indicates the a^{th} column of the matrix, with $\boldsymbol{\Theta}$ having $|\mathscr{A}|$ columns, where $\mathscr{A}$ is the action space. $\boldsymbol{\Theta}$ and $\boldsymbol{\phi}$ usually have $J \ll |\mathscr{S}|$ rows, with $\mathscr{S}$ indicating the state space. Equation (2) indicates that the Q-function, which represents an estimate of the future reward for a particular state s and action a, is approximated by the sum of J products of weights and linearly independent basis functions of the state for each action. Thus, least-squares fixed-point approximation can be used to find the weights as shown in Fig. 3 [9].

Here, tabular features or basis functions are employed for simplicity after analysing their good performance in [10] for resistive control. These result in an exact representation if discrete states and actions are used [9], since each discrete state-action pair is assigned an individual weight (with all other weights being deactivated). In this case, $J = |\mathscr{S}|$.

With LSPI, it is possible to identify two main stages: policy evaluation (the critic) and policy improvement (the actor) [9]. The controller finds the optimal policy by iterating between these two stages. The scheme is trained off-line using samples of a format (s,a,r,s') that have been previously recorded from observations of the environment. The algorithm is summarized in Fig. 3, with the discount factor being set here to $\gamma_{\mathrm{d}} = 0.95$. The development of the scheme is described in [9] to which the reader is referred.

input: $\boldsymbol{W}$: set of samples (s,a,r,s')
γ_d: discount factor
$\delta = 10^{-3}$: stopping criterion
π_0: initial policy, given as $\boldsymbol{\Theta}_0 = \mathbf{0}$
π: policy, or exploration strategy

- $\boldsymbol{\Theta}' \leftarrow \boldsymbol{\Theta}_0$
- **while** $||\boldsymbol{\Theta} - \boldsymbol{\Theta}'|| \geq \delta$:
  - $\Theta \leftarrow \boldsymbol{\Theta}'$
  - $\widetilde{\boldsymbol{A}} \leftarrow \mathbf{0}$ ($J \times J$ matrix)
  - $\widetilde{\boldsymbol{B}} \leftarrow \mathbf{0}$ (J vector)
  - **for** each $(s,a) \in \boldsymbol{W}$:
    - $\widetilde{\boldsymbol{A}} \leftarrow \widetilde{\boldsymbol{A}} + \boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma_d \boldsymbol{\phi}(s', \pi(s'))\right)^T$
    - $\widetilde{\boldsymbol{b}} \leftarrow \widetilde{\boldsymbol{b}} + \boldsymbol{\phi}(s) r$
    - $\boldsymbol{\Theta}'_{:,a} \leftarrow \widetilde{\boldsymbol{A}}^{-1}\widetilde{\boldsymbol{b}}$
- **return** $\boldsymbol{\Theta}$

FIGURE 3. LSPI algorithm, taken from [10].

Application of LSPI to the reactive control of a WEC

A time-averaged approach is used, with the duration of the time-averaging period, or time horizon, being indicated by H_{RL}. The RL states are determined from the significant wave height, the energy wave period, and the PTO damping and stiffness coefficients. At the start of each time horizon h, an action, which consists in a step change in B_{PTO} and C_{PTO}, is chosen following the current policy. The new value of the coefficients is held constant throughout the duration of the time horizon. At its end, the reward is obtained as a function of the mean generated power, P_{avg}. Conversely, a penalty is returned instead if the displacement constraints have been exceeded during the time horizon. As a result of the action selection, the agent lands in a new state, and the sample (s,a,s',r) is added to the sample set $\boldsymbol{W}$. After the collection of N_{s} samples, the policy is updated using the LSPI algorithm shown in Fig. 3. In the following sections it is possible to find a detailed description of the RL state and action spaces, the reward function and the exploration strategy.

State Space As in [5], the discrete RL state-space is expressed as:

$$\mathscr{S} = \left\{ s | s_{i,l,m,n} = \left(H_{\mathrm{s},i}, T_{\mathrm{e},l}, B_{\mathrm{PTO},m}, C_{\mathrm{PTO},n}\right), \begin{matrix} i = 1:I, \\ l = 1:L, \\ m = 1:M \\ n = 1:N \end{matrix} \right\}. \tag{3}$$

Because discrete states are employed, $J = ILMN$. I and L are determined from the wave data at the deployment site, with steps of 1 m and 1 s being common for H_{s} and T_{e}, respectively [13].

Action Space As in [5], only 5 actions are selected in order to reduce the convergence time of the RL algorithm. In particular, the RL action space is given by:

$$\mathscr{A} = \{a | [(-\Delta B, 0), (0, -\Delta C), (0,0), (0, +\Delta C), (+\Delta B, 0)]\}, \tag{4}$$

where $\Delta B = B_{\mathrm{PTO},m+1} - B_{\mathrm{PTO},m}$ and $\Delta C = C_{\mathrm{PTO},n+1} - C_{\mathrm{PTO},n}$. Furthermore, the states corresponding to the maximum and minimum values of the coefficients are limited to a smaller set of actions so as to prevent the actuator from going beyond the state space limits. For instance, if $B_{\mathrm{PTO}} = 0$ Ns/m, the first action in Eq. (4) is not allowed.

Reward Function The reward function is described thoroughly in [5], to which the reader is referred for a detailed explanation. For the purposes of this study, it is important to notice that a penalty term $p = -1$ is returned whenever the magnitude of the maximum displacement exceeds the limit, i.e. $\max|z| > z_{\mathrm{Max}}$. As a result, the algorithm learns to avoid actions that may result in structural damages. However, the reader will notice that learning occurs *after* the event. Hence, it is clear that in order to prevent failure, the value of z_{Max} should be set using a safety factor, i.e. z_{Max} needs to be designed as a soft constraint. This way the controller understands it is moving towards a dangerous region in the $B_{\mathrm{PTO}} - C_{\mathrm{PTO}}$ space and will take evasive actions before failure actually occurs. Simulations will be required to assess the sensitivity of the floater displacement to different PTO coefficient values in each sea state in order to select a suitable safety factor. However, this goes beyond the scope of this work.

The reward function can be summarized as:

$$r = \begin{cases} \left[\frac{\boldsymbol{m}(s_h)}{\max_{s''=o:p} \boldsymbol{m}(s'')}\right]^u & \text{if constraints met} \\ p & \text{otherwise} \end{cases}. \tag{5}$$

The entries of the vector $\boldsymbol{m}$, whose length is equal to the total number of discrete states $|\mathscr{S}|$, correspond to the average of up to n_{v} values of $P_{\mathrm{avg}}/H_{\mathrm{s}}^2$ that are stored for each state. Older values are overwritten by new ones once n_{v} values are registered. In regular waves, $n_{\mathrm{v}} = 10$, whereas in irregular waves $n_{\mathrm{v}} = 20$. The indices o and p ensure that the maximization in Eq. (5) is performed only over the values of $\boldsymbol{m}$ corresponding to the current sea state, as given by H_{s} and T_{e}. The value of the power has been set to $u = 25$ in this work.

Exploration Strategy At the start of each time horizon h, a new action is chosen with an ε-greedy policy [6]:

$$a = \begin{cases} \arg\max_{a' \in \mathscr{A}} Q(s_h, a') & \text{with probability } 1 - \varepsilon_h \\ \text{random action} & \text{with probability } \varepsilon_h \end{cases}, \tag{6}$$

where ε_h is the exploration rate. This means that with probability $1-\varepsilon_h$ the greedy action is selected, i.e. the action that corresponds to the maximum Q-value for the current state. As the Q-function is a measure of the expected total reward, this corresponds to maximizing the future reward.

Initially, greater exploration is desired, whereas as the learning progresses the shift will move to the exploitative, or greedy, action. Hence, the exploration rate is expressed as:

$$\varepsilon_h = \begin{cases} \varepsilon_0 & \text{if } \boldsymbol{N}(s_h) \leq N_\varepsilon \\ \varepsilon_0 / \sqrt{\boldsymbol{N}(s_h) - N_\varepsilon} & \text{if } \boldsymbol{N}(s_h) > N_\varepsilon \end{cases}, \tag{7}$$

with $\boldsymbol{N}(s_h)$ indicating the number of visits to the current state. $N_\varepsilon = 5$ is the minimum number of encounters for random exploration, and the initial exploration rate is set to $\varepsilon_0 = 0.5$.

Algorithm

The algorithm for the reactive control of the point absorber using LSPI is shown graphically in Fig. 4. During each time horizon h, the encountered sea state, mean generated power and maximum float displacement are measured and used to obtain the state and reward. Additionally, at the end of the horizon, a new action is chosen following the selected policy, with the controller landing in a new state. The current state, action, new state and reward are stored to memory as a sample in a list $\boldsymbol{W}$. Physical limitations of the computer memory mean that only a predefined number of samples can be stored. Hence, new samples are stored only if they are not a repeat, with a difference greater than 10^{-3} being acceptable for the reward. Once the memory limit is reached, older values will be overwritten. During this process, it is fundamental to ensure that the range of samples remains broad.

In Fig. 4, it can be seen that the policy is improved every $N_h = 40$ time horizons using the LSPI algorithm described in Fig. 3. In order to ensure this is performed in real-time, the update will need to be performed off-line exploiting parallel processing.

The duration of the time horizon has been set to $H_{\mathrm{RL}} = 10T$ in regular waves and $H_{\mathrm{RL}} = 200$ s in the analysed irregular waves. The longer time is necessary due to the stochastic nature of random seas. As described later, a JONSWAP wave spectrum is used, which presents most of the wave energy concentrated in the area near the peak frequency [13]. If a wider-banded or even double-peaked wave spectrum were adopted instead, the time horizon duration needs to be increased in order to fully characterize the spectrum with fast-Fourier transforms [13]. In addition, the averaging process is started only after $H_{\mathrm{RL},1} = 0.5H_{\mathrm{RL}}$ in regular waves and $H_{\mathrm{RL},1} = 0.4H_{\mathrm{RL}}$ in irregular waves to remove the influence of the transient effects associated with the change in PTO damping and stiffness coefficients.

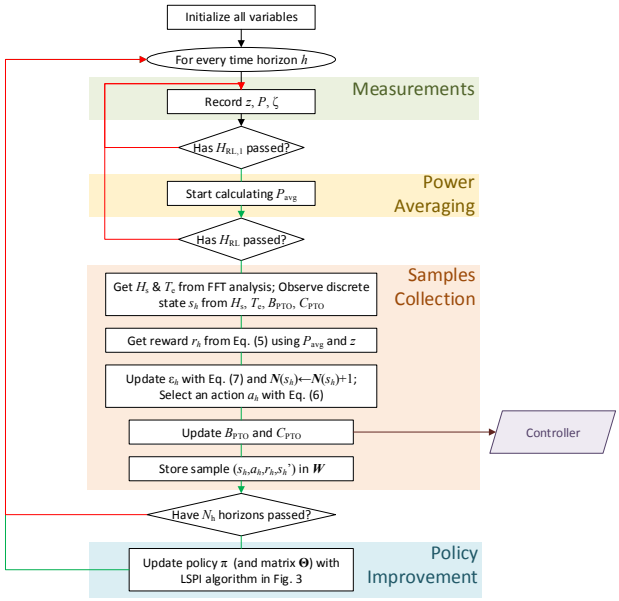

FIGURE 4. Flowchart of the LSPI algorithm for the reactive control of the point absorber.

LOW-LEVEL REAL-TIME CONTROLLER

Although the penalty term is effective in teaching the controller to avoid selecting combinations of the PTO coefficients that result in large motions, it does not prevent it from taking them. In fact, the agent needs to take those actions first in order to learn that they are bad. Including a safety factor in the displacement constraints reduces this risk considerably. Similarly, simulations can be used to pre-train the controller within a safe environment. When the control scheme is then applied to the actual device, the controller is expected to move only about the optimum point. Nevertheless, these approaches do not remove the risk of exceeding the actual displacement constraints completely due to the random element in the ε-greedy exploration strategy in Eq. (6).

For this reason, we propose the use of a lower-level, real-time control scheme within reactive control. Its aim is to ensure that displacement constraints are met. It is assumed that it is possible to directly control the PTO as is the case when model predictive control is applied. Whether this is feasible in practice or how it is implemented in reality is not treated within this work.

A very simple scheme is considered based on the magnitude of the instantaneous magnitude of the displacement and sign of the velocity of the float. If the magnitude is greater than a certain value that corresponds to a soft constraint, say $z_{\mathrm{lim}} = 90\% z_{\mathrm{Max}}$, and the sign of the velocity is either rising in a wave peak or decreasing in a wave trough, the applied PTO force is changed in

sign:

$$F_{\mathrm{PTO,rt}}(t) = \begin{cases} -F_{\mathrm{PTO}}(t) & \text{if } z > z_{\mathrm{lim}} \;\&\; \dot{z} > 0 \text{ or } z < -z_{\mathrm{lim}} \;\&\; \dot{z} < 0 \\ F_{\mathrm{PTO}}(t) & \text{otherwise} \end{cases}, \tag{8}$$

where F_{PTO} is obtained as in Fig. 2. Even though this behaviour will pose an extra burden on the PTO and it is likely to decrease the generated power, it is expected to limit the displacement when motions are large. It should be noted that z_{Max} is obtained from the sensitivity analysis of the displacement on the PTO coefficients in each sea state using simulations and it already includes a safety factor, as aforementioned.

SIMULATION RESULTS

Simulations have been run in both regular and irregular waves. In regular waves, the wave height and period have been set to $H = 2$ m and $T = 8$ s, respectively. As a single sea state is considered, $I = L = 1$ and $J = MN$. In irregular waves, a JONSWAP wave spectrum has been adopted [13]. In order to demonstrate the ability of RL to switch between different sea states and pick up learning from where the controller left off the last time it encountered a particular sea state, two alternating sea states are considered. The first sea state presents $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 7$ s, while the latter $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s. The two sea states alternate every 2 hours, which is a realistic duration for a sea state, with typical values ranging from 0.5 to 6 hours [13]. Therefore, $I = 1$, $J = 2$ for the simulations in irregular waves.

A vertical cylinder with a radius of 5 m and a draught of 8 m has been selected for the float geometry, as in [4]. Similarly, the same fifth-order state-space model as in [4] has been used to approximate the radiation convolution. Due to the well-behaved nature of the model of the system dynamics, a first-order-accurate Euler solver has been used for its solution, with a time step of 0.1 s. The overall PTO efficiency has been set to $\eta = 0.75$ and the force constraint to $F_{\mathrm{Max}} = 0.5$ MN.

Using the LSPI algorithm, the PTO damping and stiffness coefficients are each discretized with 7 values, ranging from 0 to 300 kNs/m and -300 to 0 kN/m with steps of 50 kNs/m and 50 kN/m, respectively. The resulting relatively fine discretization seems realistic for practical applications. In irregular waves, only 4 values are used for the PTO damping coefficient (hence, in steps of 100 kNs/m) in order to speed up convergence, since the mean generated power is more affected by the PTO stiffness coefficient. As a result, the total number of states for each sea state is 49 and 56 in regular and irregular waves, respectively.

Additionally, a Nelder-Mead Simplex optimization algorithm has been used to find the optimal coefficients for each sea states in the analysis as described in [15]. This has been used in the presentation of the results as a benchmark for the reinforcement learning solution.

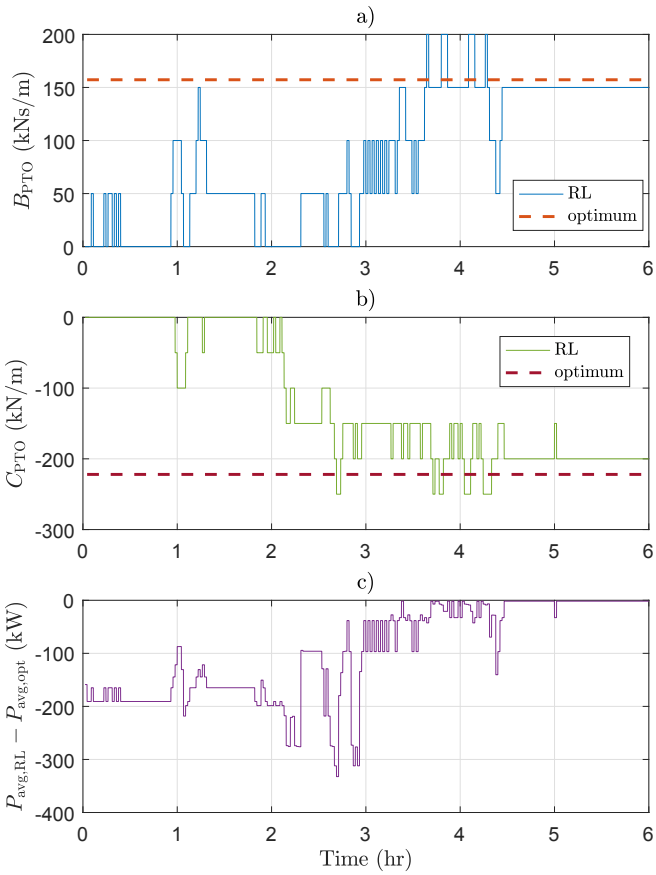

FIGURE 5. Selection of the PTO damping (a) and stiffness (b) coefficients by the RL control as compared with the respective optimal values in regular waves with $H = 2$ m and $T = 8$ s and a maximum allowable displacement of 5 m. The difference in the corresponding mean generated power can be seen in (c).

RL solution with no displacement constraints

Initially, the simulations are run with the float displacement limit set at $z_{\mathrm{Max}} = 5$ m, which is never exceeded in either regular or irregular waves for the selected sea states.

Regular waves A wave trace lasting 6 hours is generated. The time variation of the PTO damping and stiffness coefficients selected by the LSPI algorithm can be seen in Fig. 5a and Fig. 5b, respectively. Figure 5c shows the difference between the generated power and the power generated using the optimal coefficients, used as a benchmark.

Irregular waves A 24-hour long time series is employed, with the sea states alternating every 2 hours. The RL control action and the corresponding generated power can be seen in Fig. 6. It can be seen that the first sea state ($H_{\mathrm{s}} = 2$ m and

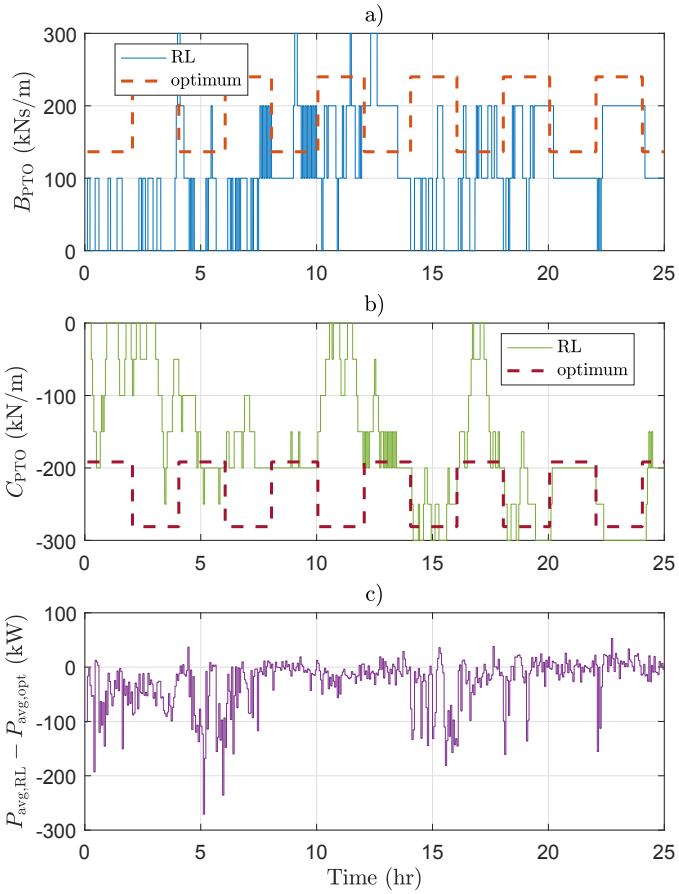

FIGURE 6. Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the RL control as compared with the respective optimal values in irregular waves with two, alternating sea states. (c) shows the difference between the corresponding and the optimal mean generated power.

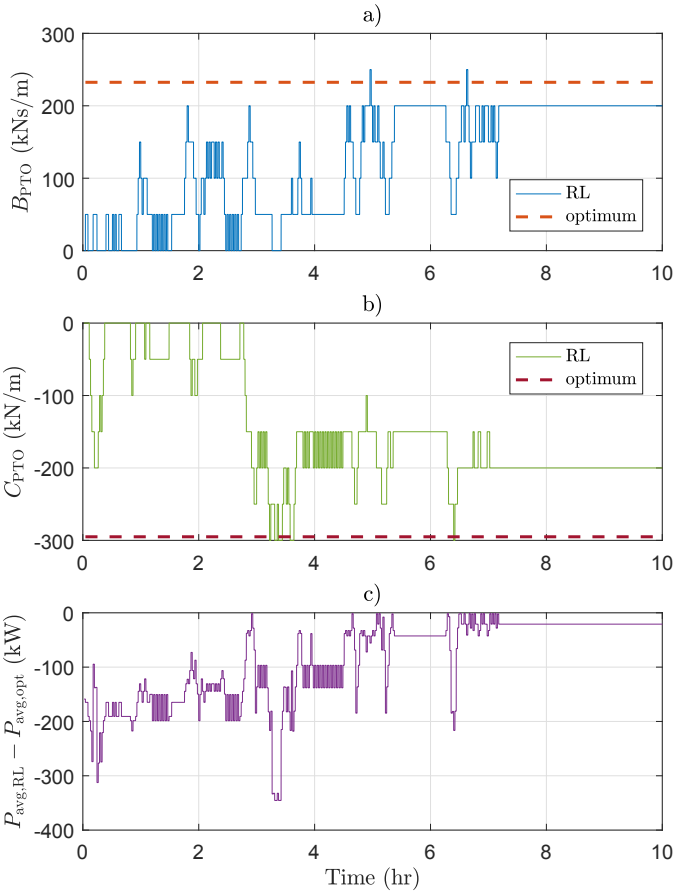

FIGURE 7. Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the RL control as compared with the respective optimal values in regular waves with $H = 2$ m and $T = 8$ s and a maximum allowable displacement of 2 m. (c) shows the difference between the corresponding and the optimal mean generated power.

$T_{\mathrm{e}} = 7$ s) is run for an extra hour to show that the wiggle in the B_{PTO} value just before $t = 22$ hr is due to random actions being selected by the ε-greedy exploration strategy.

RL solution with displacement constraints active

In order to assess the efficacy of the penalty formulation, the displacement constraint has been lowered to ± 2 m, which is lower than the amplitude of the response achieved with the optimal PTO setting. This is preferred over an increase in the energy content of waves because a linear model is used for the hydrodynamics, whose validity is void for large motions. Only regular waves have been analysed for this study, considering a 10-hour-long wave trace.

The control action selection of RL can be seen in Fig. 7 as compared with the optimal coefficients. These are calculated by modifying the cost function of the Nelder-Mead optimization to return a power value of 0 if the displacement constraint is exceeded. The difference in the generated power between the RL response and the optimal solution is shown in Fig. 7c. Note that for the combinations $B_{\mathrm{PTO}} = 200$ kNs/m and $C_{\mathrm{PTO}} = -250$ kN/m, and $B_{\mathrm{PTO}} = 200$ kNs/m and $C_{\mathrm{PTO}} = -300$ kN/m the constraint is exceeded.

Real-time controller for soft displacement constraints

The performance of the proposed low-level controller is assessed with simulations in both regular and irregular waves. In particular, the coefficients learned by RL in the unconstrained runs are employed in conjunction with a soft constraint $z_{\mathrm{lim}} = 0.9 z_{\mathrm{Max}} = 1.8$ m. For clarity, the constrained response of the point absorber is plotted against the unconstrained response.

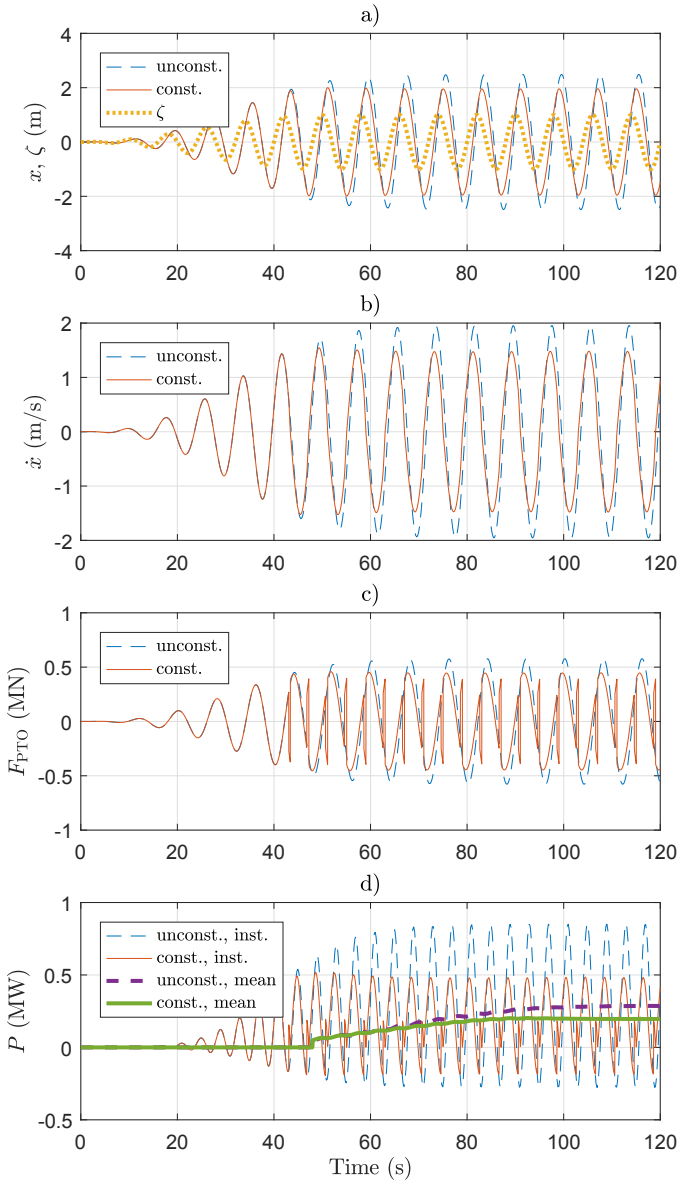

FIGURE 8. Response of the device in both constrained and unconstrained conditions in regular waves with $H = 2$ m and $T = 8$ s, including plots of the displacement, velocity, PTO force and generated power.

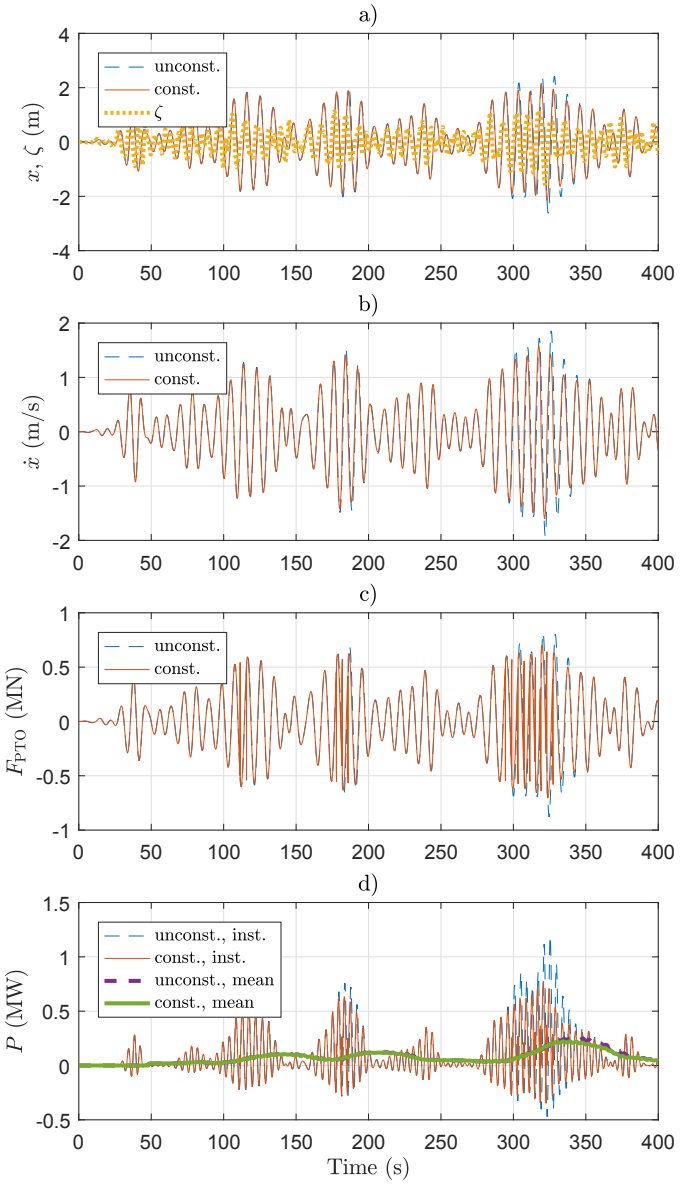

FIGURE 9. Response of the device in both constrained and unconstrained conditions in irregular waves with $H_s = 2$ m and $T_e = 8$ s, including plots of the displacement, velocity, PTO force and generated power.

Regular waves In regular waves with $H = 2$ m and $T = 8$ s, the PTO coefficients are set to $B_{PTO} = 150$ kNs/m and $C_{PTO} = -200$ kN/m. A 120-s-long time series is sufficient to get a fully-developed response, as shown in Fig. 8 for both constrained (continuous line) and unconstrained (dotted line) cases. In particular, Fig. 8a presents the float displacement and the wave elevation ζ, Fig. 8b the float velocity, Fig. 8c the PTO displacement, and Fig. 8d the instantaneous and mean generated power.

Irregular waves Considering only one sea state with $H_s = 2$ m and $T_e = 8$ s and a JONSWAP spectrum, the PTO coefficients are set to $B_{PTO} = 200$ kNs/m and $C_{PTO} = -300$ kN/m. Figure 9 shows the response of the point absorber for a portion of the wave trace with a higher energy content. In particular, Fig. 8a presents the float displacement and the wave elevation ζ, Fig. 8b the float velocity, Fig. 8c the PTO displacement, and Fig. 8d the instantaneous and mean generated power.

DISCUSSION

RL Analysis

By looking at Fig. 5 and Fig. 6, LSPI with discrete states and actions is found to learn the optimal coefficients in both regular and irregular waves when no displacement constraints are active. However, the selected large number of states results in very slow learning time. This is particularly evident in irregular waves, where up to 12 hours are required for convergence per sea state. Nevertheless, this figure shows that RL is able to recognize the change in sea state and pick up learning from where it left off the last time the controller encountered those wave conditions. This is fundamental for a practical implementation of LSPI control of a WEC. The learning time strongly depends on the number of states and thus on the discretization of H_s, T_e, B_{PTO} and C_{PTO}. For this reason, alternative machine learning schemes that provide a continuous regression, such as artificial neural networks, may be superior.

From Fig. 7, when the displacement constraints are active, LSPI seems to be unable to converge towards the optimal coefficients. In fact, it does learn the otpimal coefficients, since the penalty term is active for both combinations $B_{PTO} = 200$ kNs/m and $C_{PTO} = -250$ kN/m, and $B_{PTO} = 200$ kNs/m and $C_{PTO} = -300$ kN/m. Hence, the response of RL is affected by the inability to select a PTO damping coefficient closer to 220 kNs/m, which results in a power loss of about 30 kW as compared with the optimal response, as can be seen in Fig. 7c. The sensitivity of the mean generated power on the PTO coefficients with reactive control is another indicator that a continuous optimization method, such as the one based on neural networks in [10], is superior for this application.

Another worrying feature that can be seen in Fig. 7 is the selection that result in the exceedance of the displacement constraint during the exploration stage of the RL algorithms. It is clear that the controller selects combinations of low damping and high stiffness that result in extreme displacement and even negative power. Although the algorithm does learn to avoid these states because of the penalty term, the fact that they are encountered at all would result in failure in practice. A solution would be to pre-train RL with simulations, so that it learns to avoid some extreme actions. Once applied to the actual device, the controller would then correct the rewards it obtained from the simulations from those observed in reality. Nevertheless, this does not completely remove the possibility of selecting catastrophic actions. Hence, the proposed low-level controller is likely to be required as a fall-back option in any case.

Real-time controller for soft displacement constraints

As is clear from Fig. 8a, the proposed real-time controller is able to limit the float displacement within ± 2 m in regular waves despite the use of soft rather than hard constraints. The action of the controller is evident in Fig. 8c from the comparison between the constrained and unconstrained cases. Whether such a response is practically feasible is another problem that will need addressing. Furthermore, as expected the application of the constraints results in a drop in mean generated power in Fig. 8d.

In irregular waves, a particularly challenging situation has been analysed, with energetic waves relative to the selected sea state. As can be seen in Fig. 9, the soft constraints are not able to prevent the float from exceeding the limits, despite the magnitude of the displacement is only just greater than 2 m. This is because of the steepness of the response, which means the effect of the PTO force of opposite sign comes too late. This shows that the selected controller is too simplistic, and more accurate studies are required in order to prevent exceedance of displacement constraints in realistic wave conditions. In particular, the value of 90% use to determine the soft constraints will need to be adjusted based on the device dynamics and the sea states of interest.

CONCLUSIONS

For the first time, LSPI has been applied to the reactive control of a WEC, with the simulation of a point absorber being used for validation. The algorithm has been found to be able to learn the optimal PTO damping and stiffness coefficients in each sea state in both regular and irregular waves, although the performance is affected by the discretization of the coefficients. In particular, in irregular waves, two alternating sea states are employed to show that the controller is able to pick up learning in each sea state from where it left off the last time it encountered those wave conditions. However, a long learning time is required due to the stochastic nature of random sea waves (> 12 hours). Machine learning algorithms for regression studies should be investigated as an alternative. RL itself if more suitable for bang-bang control, such as latching or declutching control.

A detailed study on the implementation of displacement constraints with RL control for WECs has been performed here for the first time as well. In particular, the use of a penalty term is efficient in teaching the controller to prevent the choice of combinations of the coefficients that result in excessive motions. However, it does not prevent the controller from selecting them in the first place, since the agent cannot know they are bad actions until it tries them. For this reason, an additional lower-level, real-time control algorithm has been proposed, which modifies the sign of the PTO force in order to avoid exceeding the displacement limits. It has been assessed in both regular and irregular waves, and found to be effective particularly in regular waves. Nevertheless, the selection of the soft constraints is likely to require further study, as the hard constraint has been exceeded in irregular waves for a wave group with steep wavelets. The development of a real-time controller based on machine learning may alleviate this problem by including constraints within its design as is the case with model predictive control.

ACKNOWLEDGMENT

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

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

[3] Salter, S. H., Taylor, J. R. M., and Caldwell, N. J., 2002. "Power conversion mechanisms for wave energy". *Proceedings of the I MECH E Part M,* ***216***(1), pp. 1–27.

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

[5] Anderlini, E., Forehand, D. I. M., Bannon, E., Xiao, Q., and Abusara, M., 2017. "Reactive Control of a Two-Body Point Absorber using Reinforcement Learning". *Ocean Engineering*(IDCORE Special Issue), p. under review.

[6] Sutton, R. S., and Barto, A. G., 1998. *Reinforcement Learning*, hardcover ed. MIT Press.

[7] Khan, S. G., Herrmann, G., Lewis, F. L., Pipe, T., and Melhuish, C., 2012. "Reinforcement learning and optimal adaptive control: An overview and implementation examples". *Annual Reviews in Control,* ***36***(1), pp. 42–59.

[8] Falcão, A. F. D. O., 2010. "Wave energy utilization: A review of the technologies". *Renewable and Sustainable Energy Reviews,* ***14***(3), pp. 899–918.

[9] Lagoudakis, M. G., and Parr, R., 2003. "Least-squares policy iteration". *The Journal of Machine Learning Research,* ***4***, pp. 1107–1149.

[10] Anderlini, E., Forehand, D. I. M., Bannon, E., and Abusara, M., 2017. "Control of a Realistic Wave Energy Converter Model using Least-Squares Policy Iteration". *IEEE Transactions on Sustainable Energy*, p. under review.

[11] Forehand, D., Kiprakis, A. E., Nambiar, A., and Wallace, R., 2016. "A Bi-directional Wave-to-Wire Model of an Array of Wave Energy Converters". *IEEE Transactions on Sustainable Energy,* ***7***(1), pp. 118–128.

[12] Falnes, J., 2005. *Ocean waves and Oscillating systems*, paperback ed. Cambridge University Press.

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

[14] Franklin, G. F., Powell, J. D., and Emami-Naeini, A., 2008. *Feedback Control of Dynamic Systems*, 6th editio ed. Pearson.

[15] Nambiar, A. J., Forehand, D. I. M., Kramer, M. M., Hansen, R. H., and Ingram, D. M., 2015. "Effects of hydrodynamic interactions and control within a point absorber array on electrical output". *International Journal of Marine Energy,* ***9***, pp. 20–40.