Non-causal Control For Wave Energy Conversion Based on the Double Deep Q Network

1st Hanzhen Wang
Imperial College London
London, United Kingdom
hanzhen.wang23@imperial.ac.uk

2nd Vincentius Versandy Wijaya
School of Engineering
University of Southampton
Hampshire, United Kingdom
vvw1n22@soton.ac.uk

3rd Yao Zhang
School of Engineering
University of Southampton
Hampshire, United Kingdom
yao.zhang@soton.ac.uk

4th Tianyi Zeng
Rolls-Royce UTC Manufacturing and On-Wing Technology
University of Nottingham
Nottinghamshire, United Kingdom
tianyi.zeng@nottingham.ac.uk

5th Xin Dong
Rolls-Royce UTC Manufacturing and On-Wing Technology
University of Nottingham
Nottinghamshire, United Kingdom
xin.dong@nottingham.ac.uk

***Abstract*—To harness maximal wave energy, control and optimization for wave energy converters(WECs) have been investigated for decades. It has been long recognized that WEC control is essentially a non-causal control problem, in which future wave determines current control decisions. This paper introduces double deep Q network into the foundation of the non-causal time variant PD control system, enabling real-time parameter adjustments for dynamic control responses. Additionally, this paper delves into a comparative assessment of the influence of different prediction horizons on the efficiency of energy harvesting. The primary objective of this study is to elevate the control performance of wave energy converters, facilitating more efficient capture and conversion of wave energy into usable electrical power. The integration of deep reinforcement learning empowers researchers to adapt swiftly to fluctuating waves and ocean conditions, fine-tuning control parameters to enhance overall system efficiency and stability. Taking the point absorber as an example, the effectiveness of the proposed method has been verified. This method can be straightforwardly applied to other types of WEC, such as Dielectric Elastomer Generators and Dielectric Fluid Generators.**

***Index Terms*—Wave Energy Converter, Double Deep Q Network, Wave Prediction, Robustness**

I. Introduction

As a promising renewable resource, wave energy provides high energy density and continuous power supply [1], [2] and has a great potential of supplying global resources of 146 TWh/yr [3]. However, compared to wind and solar energy, such potential has yet to be fully unrealized due to the high Levelized Cost of Energy (LCOE). Various types of wave energy converters (WECs) have been investigated and developed during the past decades, including point absorbers, overtopping WECs, oscillating water columns, and attenuators [4]. It has been long recognized that control plays an important role in maximizing energy output and enhancing efficiency. More importantly, it has been proven that wave prediction can further improve the control performance [5]. This is called a non-causal control, in which the current control action is determined by not only the current feedback but also the future information.

Recent studies have proposed a large number of prediction-based non-causal control methods that aim at maximizing wave power production under actuator constraints. These studies show a promising energy harvesting performance [6]–[10]. Control methods like MPC based on hydrodynamic principles for WEC control can offer improved performance than traditional control strategy [11]. Another study proposed a fully convex implementation, which trades off the energy absorption, the energy consumed by the actuator, and safe operation [7]. Apart from that, a quadratic programming method gets even better performance [12]. There are also some other relatively effective methods proposed [8], [13], [14]. However, the accuracy of wave prediction is of great importance in the performance of WEC control which cannot be completely insured by wave prediction methods. Inaccurate predictions of the wave will make it difficult to reach the expected performance of the WEC controller [15]. And many works aim at either improving wave forecast precision or making the control algorithm "smarter".

There are a few prediction methods proposed and applied to the WEC non-causal control problem. Some of the prediction approaches are based on statistical methods, like the Auto-Regressive (AR) prediction method [16] and the extended Kalman Filters (EKF) [17]. As a novel model that has been used in multiple fields, Neural networks have also been introduced to forecast short-term wave forces [18], [19]. Other prediction methods rely on the extra sensors that can provide measurements of sea wave elevations at multiple upstream locations with certain distances away from the WEC, such as the deterministic sea wave prediction (DSWP) [20] This type of prediction can give longer and more reliable wave prediction but at the cost of extra more expensive hardware. Generally, a perfect wave prediction is difficult.

So, it usually requires the control method to have a better tolerance of inaccurate prediction or better robustness.

Recently, machine learning techniques have shown amazing performance in conducting complex tasks, especially facing ambiguous inputs like nature language [21], image classification [22], and data-driven modeling. Machine learning comes into the WEC control community mainly in two methods. The first is to use machine learning methods to build a data-based, nonlinear model of the system dynamic for system identification [23]. The second is to optimal paraments [24] of other control theories or to conduct data-driven online control [25]. Different from supervisor learning and non-supervisor learning, reinforcement learning lets the agent learn from the interaction with the environment [26]. Based on data-driven logic, reinforcement learning has shown satisfactory performance in dealing with systems with uncertainty [27], like games [28] and go [29], etc. This makes the reinforcement learning method suitable to deal with inaccurate predictions and new circumstances that the agent has not met before. Therefore, reinforcement has been used in robotics systems [26], which have similar mechanics to the WEC problem and are faced with many uncertain circumstances. A study has shown it has the potential to tackle the inaccuracy of the WEC control problem [30]. However, the research on the application of reinforcement learning in the WEC field is still insufficient. Only a few studies of WEC involve reinforcement learning in WEC control [30]–[32]. The Q-learning algorithm is one of the classical value-based RL algorithms [33]. The research [34] by Anderlini et al. applies Q-Learning in identifying the optimal damping for WECs. Due to the complexity of a WEC system, Deep Q-Network (DQN), a kind of Deep Reinforcement Learning (DRL), is introduced to WEC control later [35]. The DQN algorithm was put forward to play artari [28] at first but shows great competence in improving the performance of the Q-learning algorithm. Usually, a DQN method is used to solve discrete problems, but a recent study applied a time-varying PD control whose gains are adjusted by DQN [30] because the control of a WEC is typically continuous.

This paper investigates a non-causal control strategy using DQN developed for point absorber WEC systems, whose control performance is improved by benefitting from both the wave prediction and the DQN. This paper aims to fill the research gap of incorporating wave prediction into model-free control methods like DQN. Although the future information of waves could be partly reflected by the prediction ability of DQN [30], [35], such prediction only works refer to the future reward. DQN always makes the best decision based on the present reward and the estimate of future reward, but the control of PTO is conducted by the time-variant PD controller. Wave prediction can still improve the performance of the PD controller. Besides, a comparison of the improvement brought by different wave prediction horizons is also necessary to show how prediction influences control performance. To sum up, this paper focuses on the following points:

- A DQN control is proposed and developed in this study which is expected to maximize the energy output.
- Wave prediction is introduced to the time-variant controller whose paraments are decided by the DQN agent
- The comparison of the performance of the DQN control and existing non-data-based controls
- The investigation of the influence brought by different control horizons
- A realistic wave data gathered from the coast of Cornwall, Wales, rizons The UK is used to validate the effectiveness of the proposed control algorithm.
- The proposed control algorithm is generally applicable to other WECs across varied archetypes (e.g., sizes, shapes) in any location of deployment.

The rest of the paper is as follows. Section II introduces the state-space model of the point absorber. The reinforcement learning control method is proposed in Section III, where the basic structure of the agent is introduced. Simulation results for the comparison are shown in Section IV. Section V concludes this paper.

II. State-space model for Environmental Build-up

This section first introduces the dynamical model of a single-point absorber in section II-A. To build the simulation environment, the hydrodynamic model is described in a state-space model form, which introduces modelling uncertainties.

A. Dynamic model

By using Newton's second law, the dynamic equation [37] for the float of the point, the absorber is like function (1).

$$m_s \ddot{z}_v = -f_s - f_r + f_e + f_u \tag{1}$$

where m_s is the float mass. z_w and z_v are the water level and the height of the mid-point of the float respectively. The PTO torque is proportional to the force f_u acting on the piston inside the cylinder. The extracted power is $P = -f_u \dot{z}_v$. The restoring force f_s is given by equation (2).

$$f_s = k_s z_v \tag{2}$$

With the hydrostatic stiffness $k_s = \rho g_s$, ρ as water density, g as standard gravity, and s as the cross-sectional area of the float. f_r is the radiation force determined by equation (3).

$$f_r = m_\infty \ddot{z}_v + \int_{-\infty}^{\infty} h_r(\tau)\dot{z}_v(t-\tau)\,d\tau \tag{3}$$

where m_∞ is the added mass, h_r is the kernel of the radiation force that can be computed via hydraulic software packages (e.g. WAMIT [38]). Following [37], the convolutional term in (3) can be approximated by a causal finite-dimensional state-space model.

$$\begin{aligned} \dot{x}_r &= A_r x_r + B_r \dot{z}_v \\ f_R &= C_r x_r \approx \int_{-\infty}^{t} h_r(\tau)\dot{z}_v(t-\tau)\,d\tau \end{aligned} \tag{4}$$

where $(A_r, B_r, C_r, 0)$ and $x_r \in \mathbb{R}^{n_r}$ are the state-space realization and the state respectively. Following [37], the wave excitation force fe can be determined by (5).

$$f_e = \int_{-\infty}^{\infty} h_e(\tau) z_w(t-\tau)\, d\tau \tag{5}$$

where h_e is the kernel of the radiation force and the state-space approximation is given by

$$\begin{aligned} \dot{x}_e &= A_e x_e + B_e z_w \\ f_e &= C_e x_e \approx \int_{-\infty}^{t} h_e(\tau) z_w(t-\tau)\, d\tau \end{aligned} \tag{6}$$

where $(A_e, B_e, C_e, 0)$ and $x_e \in \mathbb{R}^{n_e}$ are the state-space realization and the state respectively.

B. State-space model

With the realizations of (4) and (6), the state-space model of (1) can be represented by

$$\begin{cases} \dot{x} &= A_c x + B_{uc} u + B_{wc} w + \epsilon \\ y &= C_c x \end{cases} \tag{7}$$

Where $w = z_w$ is the wave elevation whose prediction is incorporated into the controller design, $y = z_v$, $y = \dot{z}_v$, $x = [z_v, \dot{z}_v, x_r, x_e]$, $u = f_u$. ϵ represents the modeling uncertainty caused by wave force approximations (4) and (6). And

$$\begin{aligned} A_c &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_s}{m} & 0 & \frac{C_e}{m} & -\frac{C_f}{m} \\ 0 & B_r & A_r & 0 \\ 0 & 0 & 0 & A_e \end{bmatrix} B_{wc} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ B_e \end{bmatrix} B_{uc} = \begin{bmatrix} 0 \\ \frac{1}{m} \\ 0 \\ 0 \end{bmatrix} \\ C_c &= \begin{bmatrix} 0 & 1 & 0_{1\times(n_r+n_e)} \end{bmatrix} \end{aligned} \tag{8}$$

with $m = m_s + m_\infty$.

III. Control with reinforcement learning

A. Deep reinforcement learning control framework

Reinforcement learning could be generally divided into model-based reinforcement learning and model-free reinforcement learning. Unlike traditional model predict control, model-free reinforcement control does not require a precise model for controller design. Instead, a reinforcement learning control system uses the interaction with the environment to learn the policy. Reinforcement learning has shown a strong ability to deal with sequential decision-making [39], like the WEC control with prediction. Deep Q-Learning (DNQ) [28] is now widely in deep reinforcement learning. It is developed from the classical Q-Learning which has been proven to be effective in simple problems. Based on Q-Learning DQN introduced deep neural networks (DNN) to instead the Q-table used in traditional Q-Learning. This leads the RL algorithm to gain the ability to solve more complex problems like chess and video games.

In a typical RL system [40], there are an "agent" and an "environment". Due to the high cost of the mechanical structure, simulation environments are always applied to train the agent. For a DQN agent, there always is a DNN to be trained. In an RL problem, we use s_n to present the current state. Meanwhile a_n stands for the current action the agent does to the environment. The action is selected in an action space according to the policy. Besides, r_n is the reward defined manually to critique the performance of the current step. The calculation of reward always requires professional knowledge. After one step of interaction, the system goes forward and we get s_{n+1}, a_{n+1} and r_{n+1}. In each step, the selection of action is regarded as a Markov decision process. The decision is based on the value function. To avoid the over-estimate phenomena, a method of Double DQN is proposed. There are two neural networks with wights θ and θ^-, which have totally the same structure Q. Thus, the target y_T is expressed as (9).

$$y_T = r_n + \gamma \max_{a_{n+1}} Q(s_{n+1}, a_{n+1}; \theta_n^-) \tag{9}$$

Where γ is the future reward discount that represents how much we focus on the future reward. The agent is expected to learn the policy to maximize the target y. The weight of Q-network(θ) will be updated based on the weight of the target network (θ^-) with equation (10).

$$\nabla_{\theta_n} L(\theta_n) = E[(y_T - Q(s_n, a_n, \theta_n))\nabla_{\theta_n} Q(s_n, a_n, \theta_n)] \tag{10}$$

Furthermore, a minibatch training (sampled from the stored experience buffer) is also adopted to avoid divergence and smooth the learning. The agent collects r_n and new state s_{n+1} after the last action a_n has been taken. The experience is set $e_n = [s_n, a_n, r_n, s_{n+1}]$. The experience of the agent is saved in a buffer to improve the learning speed. The batch is sampled from the experience buffer to train the deep network with (10). Then the next action a_{n+1} is decided by maximizing the target with (9). The target network θ_n^- is updated after each batch by the Q-network with the function (11).

$$\theta_n^- = \tau\theta_n + (1-\tau)\theta_n^- \tag{11}$$

Where τ is the smoothing factor.

B. DQN problem formulation

Considering the importance of wave prediction in WEC control the WEC control is formulated in the fashion of time-varying PD control with prediction. Figure 1 shows the basic structure of the system. Whilst the TVPD controller can be implemented by (12).

$$F_{PTO} = K_p(n) s_n + K_d(n) x_{pre} \tag{12}$$

where K_p and K_d are adjusted by the action of RL and the x_{pre} is decided by the wave prediction. s_n contains the displacement of float and the velocity of the float. The state can be expressed as equation (13).

$$s_n = [z_v, \dot{z}_v]^T \tag{13}$$

And the length of x_{pre} is decided by the prediction horizon.

$$x_{pre} = [w_1, w_2, ..., w_{horizon}]^T \tag{14}$$

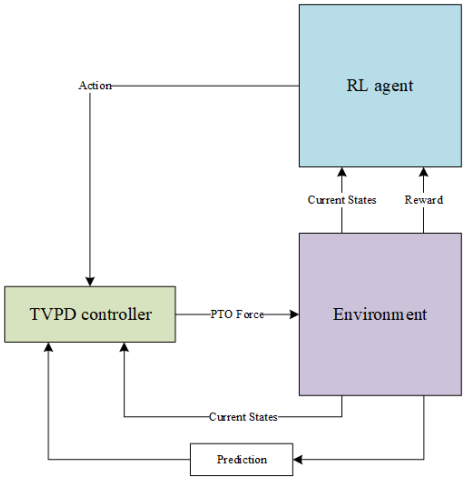

Fig. 1. Structure of the control system

To balance the cost of computing and control performance, the control period of the RL agent and the sampling period are set at the same $T_s = 0.5s$. The mission of the RL agent is to adjust K_x and K_d every RL sample period. The action of the RL algorithm is from the given action space A as follows

$$A = \{a|[(\delta, 0, ..., 0), (-\delta, 0, ..., 0), (0, \delta, ..., 0), (0, -\delta, ..., 0), ...]\} \tag{15}$$

where δ is a small amount. The length of each action choice is decided by the length of the control horizon. The given action space is similar to one-hot code to some degree. In each step, the non-causal PD is updated by the rule (16)

$$[K_p(n+1), K_d(n+1)] = [K_p(n), K_d(n)] + a_n \tag{16}$$

According to a previous study [31], we do not consider more complex actions like $(\delta, \delta, ...)$ and $(-\delta, -\delta, ...)$ to concise the action space. In a WEC control system, we focus on the energy output efficiency most. So it is reasonable to set a reward related to the power P_n. Whilst we need to protect the mechanical structure so we set a punishment notated as r_{punish} when the F_{PTO} output exceed the limitation $F_{\max}$. Therefore the reward r_n is expressed like (17).

$$r_n = \begin{cases} P_n, & |F_{PTO}| \leq F_{max} \\ r_{punish}, & |F_{PTO}| \geq F_{max} \end{cases} \tag{17}$$

In this case, r_punish is always set to a small plural number to let the agent learn to avoid the behavior that will damage the system. Also, it worth mentioning that the period of RL sampling and the period of simulation is different. So the P_n is actually calculate as equation (18).

$$P_n = \frac{P\delta t}{\delta t_{RL}} \tag{18}$$

Through the Q-function, the DRL control method can take all the future discount rewards which consist of past rewards and future rewards. This is similar to $\sum r_n$ but uses an estimate of the future.

C. DQN Agent structure and training

The structure of the neural network is shown in Figure 2. Each fully connected layer contains 24 cells and the optimizer is SGD. The agent needs to be trained previously. Parameters are set as table I.

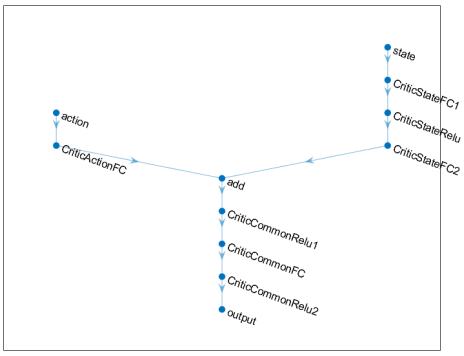

Fig. 2. Structure of the neural network

TABLE I
TRAINING PARAMETERS

Hyperparameters	Values
Learning Rate	0.01
Batch Size	64
Replay Memory Size	10^4
Discount Factor	0.999
Target smoothing Factor	0.99
Number of Neurons Each Hidden Layer	24
Target Update period	0.5s

Besides, we train three sets of agents facing different forecast horizons. For the agent in each set, the initial value of the TVPD controller is set by linear optimal non-causal control method proposed in reference [10].

IV. Simulation Results

This section shows the results of the simulation. The simulation and agent are both built with MATLAB. We built three sets of simulations, the prediction horizon is set 1, 3, and 5 steps. In the simulation episode, we set 20001 steps which are equal to 200 seconds in realistic. The training of the agent will stop when the total rewards come to a rather stable value. The three sets of simulations take 35, 73 and 21 episodes. To train the agent it takes around 6 hours, 12 hours and 5 hours. The hardware we use to train is a GPU Nvidia GTX1650 and CPU Intel Core i7-9750H. The average reward, which is filtered with a moving average filter, is like the figure 3. For a WEC system, we mostly focus on the energy converted. To show the performance, we compare it to the casual control method. The energy output by the classical control methods and by our learning-based control method can be found in Figure 4. Also, we compare different outcomes by the variant of prediction horizons to find out the best prediction horizon. Final energy output in

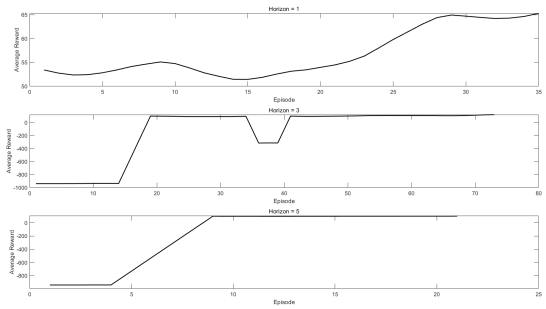

Fig. 3. Average reward during training

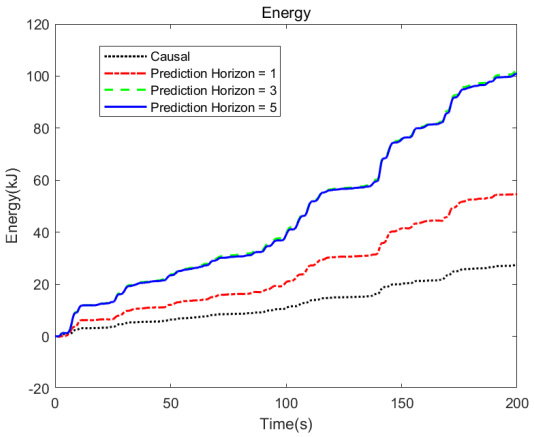

Fig. 4. Energy output by different horizons

the 200s simulation is Horizon 3: 102kJ; Horizon 5: 100.9kJ; Horizon 1: 54.7kJ; PD controller: 27.64kJ. In Figure 4, we can find that it gets similar performance between horizon 3 and horizon 5. But it is worth mentioning that we use less cost to train the agent with Horizon 5 as shown in Figure 3.

V. CONCLUSION

In light of the unique characteristics of Wave Energy Converters (WECs), we have introduced a reinforcement learning methodology to augment the non-causal TVPD (Time Variant PD Control) controller. After subjecting our approach to rigorous simulation tests using the collected dataset, we have achieved energy conversion results that outperform traditional control methods by a substantial margin. Furthermore, we conducted a comparative analysis of the control effects stemming from different prediction time steps.

Our findings unequivocally demonstrate that longer-term predictions, specifically, those involving three steps at 0.03 seconds and five steps at 0.05 seconds, significantly outperform the results obtained from shorter-term predictions, such as one step at 0.01 seconds. This indicates that extending the prediction horizon has a profound impact on improving the controller's efficacy. Additionally, it's noteworthy that even though a five-step prediction has a negligible effect on the ultimate total energy output, it does alleviate the training burden on the reinforcement learning agent. This observation suggests that the increased predictive horizon enhances the controller's performance, consequently lightening the workload for the reinforcement learning agent.

Future work will focus on applying this method to Dielectric Elastomer Generator WECs and Dielectric Fluid Generator WECs, which has larger model uncertainties.

ACKNOWLEDGEMENT

This work was funded by Wave Energy Scottland Direct Generation Competition and the UK Royal Society IEC-NSFC (223485).

REFERENCES

[1] A. Clément, P. McCullen, A. Falcão, A. Fiorentino, F. Gardner, K. Hammarlund, G. Lemonis, T. Lewis, K. Nielsen, S. Petroncini, M.-T. Pontes, P. Schild, B.-O. Sjöström, H. C. Sørensen, and T. Thorpe, "Wave energy in europe: current status and perspectives," *Renewable and Sustainable Energy Reviews*, vol. 6, no. 5, pp. 405–431, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032102000096

[2] B. Drew, A. R. Plummer, and M. N. Sahinkaya, "A review of wave energy converter technology," *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, vol. 223, no. 8, pp. 887–902, 2009. [Online]. Available: https://doi.org/10.1243/09576509JPE782

[3] R. Kempener and F. Neumann, "Wave energy technology brief," *International Renewable Energy Agency (IRENA)*, 2014.

[4] R. Baños, F. Manzano-Agugliaro, F. Montoya, C. Gil, A. Alcayde, and J. Gómez, "Optimization methods applied to renewable and sustainable energy: A review," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 4, pp. 1753–1766, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032110004430

[5] J. Falnes and A. Kurniawan, *Ocean waves and oscillating systems: linear interactions including wave-energy extraction*. Cambridge university press, 2020, vol. 8.

[6] J. Hals, J. Falnes, and T. Moan, "Constrained Optimal Control of a Heaving Buoy Wave-Energy Converter," *Journal of Offshore Mechanics and Arctic Engineering*, vol. 133, no. 1, p. 011401, 11 2010. [Online]. Available: https://doi.org/10.1115/1.4001431

[7] G. Li and 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*, vol. 69, pp. 453–463, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148114002456

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

[9] R. Genest and J. V. Ringwood, "A critical comparison of model-predictive and pseudospectral control for wave energy devices," *Journal of Ocean Engineering and Marine Energy*, vol. 2, pp. 485–499, 2016.

[10] S. Zhan and G. Li, "Linear optimal noncausal control of wave energy converters," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 4, pp. 1526–1536, 2019.

[11] N. Faedo, S. Olaya, and J. V. Ringwood, "Optimal control, mpc and mpc-like algorithms for wave energy systems: An overview," *IFAC Journal of Systems and Control*, vol. 1, pp. 37–56, 2017.

[12] Q. Zhong and R. W. Yeung, "An efficient convex formulation for model-predictive control on wave-energy converters," *Journal of Offshore Mechanics and Arctic Engineering*, vol. 140, no. 3, p. 031901, 2018.

[13] U. A. Korde and J. Ringwood, *Hydrodynamic control of wave energy devices*. Cambridge University Press, 2016.

[14] F. Fusco and J. V. Ringwood, "A simple and effective real-time controller for wave energy converters," *IEEE Transactions on sustainable energy*, vol. 4, no. 1, pp. 21–30, 2012.

[15] L. Li, Z. Yuan, Y. Gao, and X. Zhang, "Wave force prediction effect on the energy absorption of a wave energy converter with real-time control," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 615–624, 2018.

[16] Y. Zhang, T. Zeng, and G. Li, "Robust excitation force estimation and prediction for wave energy converter m4 based on adaptive sliding-mode observer," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1163–1171, 2019.

[17] F. Fusco and J. V. Ringwood, "Short-term wave forecasting for real-time control of wave energy converters," *IEEE Transactions on sustainable energy*, vol. 1, no. 2, pp. 99–106, 2010.

[18] L. Li, Z. Gao, and Z.-M. Yuan, "On the sensitivity and uncertainty of wave energy conversion with an artificial neural-network-based controller," *Ocean Engineering*, vol. 183, pp. 282–293, 2019.

[19] L. Li, Z. Yuan, and Y. Gao, "Maximization of energy absorption for a wave energy converter using the deep machine learning," *Energy*, vol. 165, pp. 340–349, 2018.

[20] L. Abusedra and M. Belmont, "Prediction diagrams for deterministic sea wave prediction and the introduction of the data extension prediction method," *International shipbuilding progress*, vol. 58, no. 1, pp. 59–81, 2011.

[21] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 160–167. [Online]. Available: https://doi.org/10.1145/1390156.1390177

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[23] J. F. Gaspar, M. Kamarlouei, A. Sinha, H. Xu, M. Calvário, F.-X. Faÿ, E. Robles, and C. G. Soares, "Speed control of oil-hydraulic power take-off system for oscillating body type wave energy converters," *Renewable Energy*, vol. 97, pp. 769–783, 2016.

[24] E. Anderlini, D. Forehand, E. Bannon, and M. Abusara, "Reactive control of a wave energy converter using artificial neural networks," *International journal of marine energy*, vol. 19, pp. 207–220, 2017.

[25] N. M. Tri, D. Q. Truong, P. C. Binh, D. T. Dung, S. Lee, H. G. Park, K. K. Ahn *et al.*, "A novel control method to maximize the energy-harvesting capability of an adjustable slope angle wave energy converter," *Renewable Energy*, vol. 97, pp. 518–531, 2016.

[26] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[27] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[29] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[30] S. Zou, X. Zhou, I. Khan, W. W. Weaver, and S. Rahman, "Optimization of the electricity generation of a wave energy converter using deep reinforcement learning," *Ocean Engineering*, vol. 244, p. 110363, 2022.

[31] E. Anderlini, D. Forehand, E. Bannon, Q. Xiao, and M. Abusara, "Reactive control of a two-body point absorber using reinforcement learning," *Ocean Engineering*, vol. 148, pp. 650–658, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801817304699

[32] L. Bruzzone, P. Fanghella, and G. Berselli, "Reinforcement learning control of an onshore oscillating arm wave energy converter," *Ocean Engineering*, vol. 206, p. 107346, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801820303784

[33] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

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

[35] E. Anderlini, S. Husain, G. G. Parker, M. Abusara, and G. Thomas, "Towards real-time reinforcement learning control of a wave energy converter," *Journal of Marine Science and Engineering*, vol. 8, no. 11, 2020. [Online]. Available: https://www.mdpi.com/2077-1312/8/11/845

[36] G. Weiss, G. Li, M. Mueller, S. Townley, and M. Belmont, "Optimal control of wave energy converters using deterministic sea wave prediction," *Fuelling the Future: Advances in Science and Technologies for Energy Generation, Transmission and Storage*, vol. 396, 2012.

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

[38] C. Lee, "Wamit theory manual, department of ocean engineering," 1995.

[39] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[40] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.