Multi-Agent Reinforcement Learning Controller to Maximize Energy Efficiency for Multi-Generator Industrial Wave Energy Converter

Soumyendu Sarkar[1*], Vineet Gundeecha[1], Alexander Shmakov[1], Sahand Ghorbanpur[1], Ashwin Ramesh Babu[1], Paolo Faraboschi[1], Mathiew Cocho[2], Alexandre Pichard[2], Jonathan Fievez[2]

[1] Hewlett Packard Enterprise, USA
[2] Carnegie Clean Energy, Australia
{soumyendu.sarkar, vineet.gundecha, alexander.shmakov, sahand.ghorbanpour, ashwin.ramesh-babu, paolo.faraboschi}@hpe.com
{mcocho, jfievez, apichard}@carnegiece.com

Abstract

Waves in the oceans are one of the greatest sources of renewable energy and are a promising resource to tackle climate challenges through decarbonizing energy generation. Lowering the Levelized Cost of Energy (LCOE) for energy generation from ocean waves is key for competitiveness with other forms of clean energy like wind and solar. This requires complex controllers to maximize efficiency for state-of-the-art multi-generator industrial Wave Energy Converters (WEC), which optimizes the reactive forces of the generators on multiple legs of WEC. This paper introduces Multi-Agent Reinforcement Learning controller (MARL) architectures which can handle these multiple objectives for LCOE, helping the increase in energy capture efficiency boosting revenue, reducing structural stress to limit maintenance and operating cost, and adaptively and proactively protect the wave energy converter from catastrophic weather events, preserving investments and lowering effective capital cost. These architectures include 2-agent and 3-agent MARL implementing proximal policy optimization (PPO) with various optimizations to help sustain the training convergence in the complex hyperplane without falling off the cliff. The design for trust is implemented to assure the operation of WEC within a safe zone of mechanical compliance and guarantee mechanical integrity. This is achieved through reward shaping for multiple objectives of energy capture and penalty for harmful motions to minimize stress and lower the cost of maintenance. We achieved double-digit gains in energy capture efficiency across the waves of different principal frequencies over the baseline Spring Damper controller with the proposed MARL controllers.

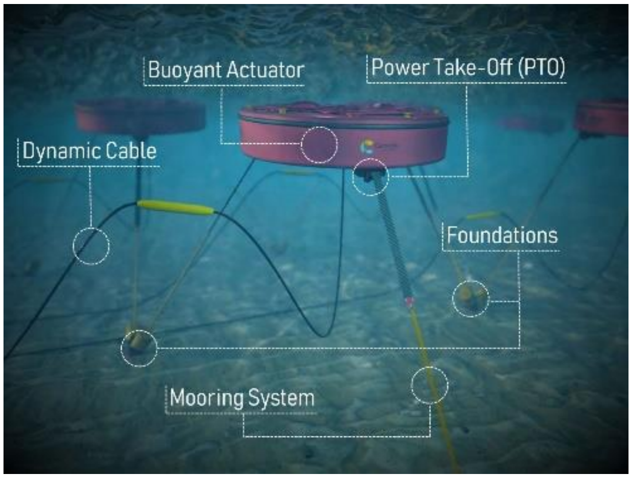


Figure 1: CETO 6 Wave Energy Converter (Ceto Technology, 2020)

Introduction

Alternate energy sources have been gaining much attention from researchers and governments worldwide in recent times. Many countries have pledged to have net-zero emissions by 2050 as carbon dioxide emissions keep increasing year after year from fossil fuels. There is a need to decarbonize electricity generation, and some of the well explored areas to achieve this are through wind and solar energy. Similarly, waves in the ocean are considered one of the more consistent and predictable sources of renewable energy, specifically for countries with large coastlines. For a fact, the worldwide resource of coastal wave energy has been estimated to be over 2 TW, representing about 16% of the world energy consumption (Yusop et al 2020). Several works have been done in the recent past to design and build Wave Energy Converters (WEC) that converts the kinetic and potential energy associated with moving ocean wave into electric energy.

Deployment of WEC needs to achieve Levelized Cost of Energy (LCOE) consistent with other competing sources. A significant challenge that can be observed while deploying WEC is the variable nature of the ocean waves. Waves vary in height and time period, especially in offshore locations, leading to the complexity of capturing energy. Additionally, these devices must withstand extreme wave conditions that rarely occur but could significantly damage the capital investment. Some existing solutions include designing and implementing single-legged to multi-legged WECs with

Spring Damper and electronic controllers to tackle structural and deployment constraints.

This work focuses on a state-of-the-art three-legged Wave Energy Converter system CETO 6 (Ceto Technology, 2020), a successor of single-legged WEC for optimizing power generation, as represented in Figure 3. With such a system, the complexity of the control has increased to a point where it is difficult for traditional engineering approaches to model the WEC structure's variabilities, including interaction with the waves, interaction between the legs and buoy, and so on. Recent advancement in AI has enabled controllers to model and learn complex movements with data collected from previous experiences. Such controllers should react appropriately even for unseen data that traditional learning techniques (supervised, self-supervised) cannot solve. Hence, we propose Multi-Agent Reinforcement Learning Controllers (MARL) to optimally apply reactive forces that control the generator on the PTOs leading to maximizing the power generation while minimizing mechanical stress reducing the yaw motion. We improve the learning policy of MARL architecture by considering multiple rewards from environment interaction using Proximal Policy Optimization. This further helps the training progression in overcoming the challenges of converging to a global optimum for all wave types. We show that our method is robust against disaster events and extreme shifts of wave directionality while significantly improving the power generated over the baseline Spring Damper Model. The main contribution of the paper can be summarized as follows:

1. A novel application of Multi-Agent Reinforcement learning in Wave Energy Converters
2. A multi-objective function that can optimize energy generation efficiency while minimizing mechanical stress, which is critical for deployment.

Wave Energy Converter problem

The WEC considered in this study is composed of a cylindrical Buoyant Actuator (BA), submerged approximately 2 meters under the ocean's surface as shown in Figure 1. The BA is secured to the sea bed through three mooring legs, each of which terminates on one of the three power take-offs (PTOs) located within the BA. The PTOs act like winches -

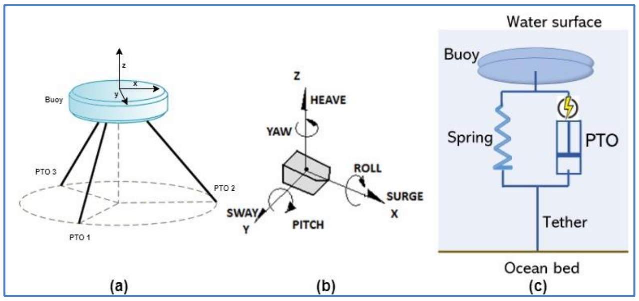


Figure 2: Geometry and parameters of the three-tether wave energy converter: (a) 3D view, (b) PTO and motion with 6 degrees of freedom, (c) WEC spring. (Sergiienko et al. 2020)

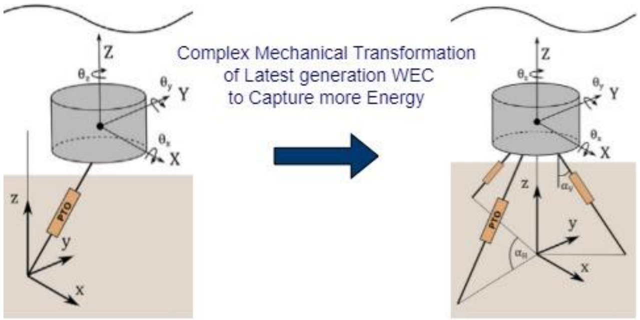


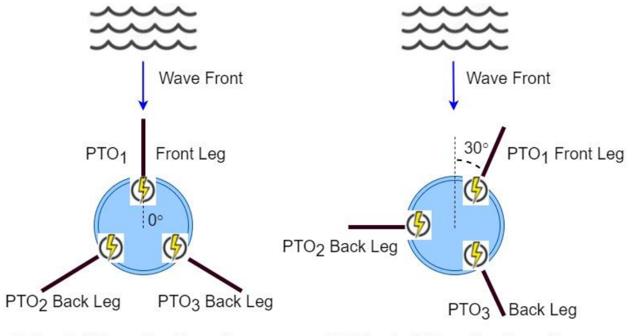
Figure 3: Increase in structural complexity of the 3-tether WEC to capture more power working against the different translational and rotational motions (Rijnsdorp et al. 2018).

they can pay in and out to allow the mooring legs to vary in length (converting the chaotic motions of the BA into linear motions) and also resist the extension of the mooring legs, thereby generating electrical power. The high-level structure of the WEC is represented in Figure 2. Optimal timing of the PTO forces with the wave excitation force is key to maximizing WEC performance. Various control strategies exist, attempting to get as close as possible to the optimal force function with various degrees of success. These include pure damping control, spring damper control, latching control, model predictive control and so on.

Spring Damper Benchmark Controller

The PTO is composed of a mechanical spring and an electrical generator, as represented in Figure 2c. The PTO force and its components are subject to various physical implementation constraints. The damping component is akin to a reactive braking torque reacting against the input shaft, driven by the wave energy source. The captured energy equals the braking mechanical work done by the generator minus losses. The spring component of the generator is tuned to induce resonance at the dominant wave frequency. This is analogous to impedance matching in the mechanical domain, where the impedance is effectively a measure of the opposition to motion when a potential force is applied.

The average mechanical power ($\overline{P_m}$) generated by each PTO is the average of the product of the generator force (F_{gen})



0° Frontal Wave Configuration

30° Angled Wave Configuration

Figure 4: 3-tether WEC configuration with different wave fronts

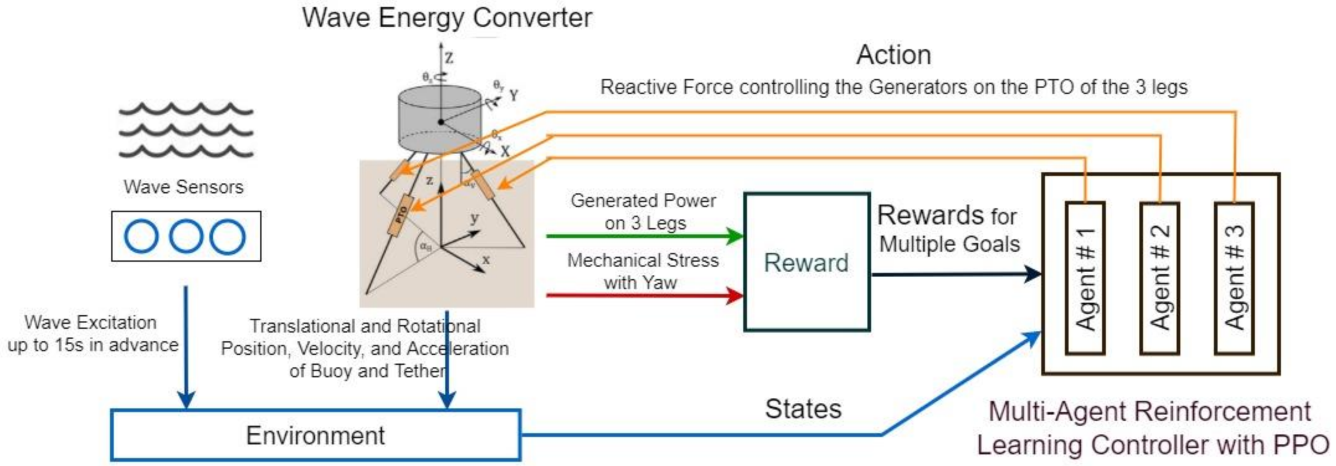

Wave Energy Converter
Z
Y
X
PTO
y
x
z
Action
Reactive Force controlling the Generators on the PTO of the 3 legs
Wave Sensors
Generated Power on 3 Legs
Mechanical Stress with Yaw
Reward
Rewards for Multiple Goals
Agent # 1
Agent # 2
Agent # 3
Wave Excitation up to 15s in advance
Translational and Rotational Position, Velocity, and Acceleration of Buoy and Tether
Environment
States
Multi-Agent Reinforcement Learning Controller with PPO

Figure 5: Architecture of Multi-agent RL controlling the WEC

and the leg extension/retraction velocity (v_{PTO}), as expressed in equation 1.

$$\overline{P_m} = \sum_{i=1}^{3} \overline{F_{gen_i} \times v_{PTO_i}} \tag{1}$$

$$F_{PTO} = F_{spring} + F_{gen} \tag{2}$$

where F*spring* is the spring force, *Fgen is the generator force*

Related Work

Reinforcement Learning (RL) has been applied to continuous control tasks for different applications (Lillicrap et al. 2019) (Duan et al. 2016). Recent works have applied RL to control simple one-legged WECs in different academic settings. (Anderlini et al. 2016) uses RL to control the PTO damping and stiffness coefficients for discrete sea states. (Anderlini et al. 2017a) applies least-squares policy iteration for resistive control of a nonlinear model of a wave energy converter. (Anderlini et al. 2018) uses RL to obtain optimal reactive force for a two-body heaving point absorber with one degree of freedom. A non-RL technique that utilizes artificial neural networks has also been proposed (Anderlini et al. 2017b) to generate power through WECs. (Anderlini et al. 2020) makes use of Deep Reinforcement Learning for real-time control of a WEC in continuous action space. However, most of these techniques have been applied to one degree of freedom motion simple point absorber. To our knowledge, **RL has not been used to control multi-legged and multi- generator WECs** with six degrees of freedom of motion.

Reinforcement Learning Environment

As shown in Figure 5, the environment depicted on the right consists of the 3-legged wave energy convertor (WEC) and the wave sensors. The environment feeds the buoy's translational and rotational motion, along with the velocity and acceleration into the RL controller, along with values related to the leg extension and tension. It also feeds the oceanic waves of different heights and principal frequencies. Based on these inputs, the RL controller directs actions using the reactive forces on the three generators for the wave energy converter legs. The projected power generated in the three legs of the WEC, along with the safety estimate, is fed back to the RL controller as rewards. This feedback helps the RL controller assess the effects of its action based on inputs from the environment to take further actions based on the altered state.

Multi-Agent RL Design for WEC

Reinforcement Learning architectures, environment state design, and reward shaping have been explored. We limit the policy optimization algorithms to Proximal Policy Optimization (PPO) for this study. We empirically found that this algorithm performed better for our use case than other RL algorithms that we tried, like the DQN, Soft Actor-Critic, and Asynchronous Advantage Actor-Critic (A3C). We made several modifications and tuning to the PPO implementation to overcome the challenges for this highly sensitive multi-agent dynamic control task with partial observability, continuous action space, long-term dependence, preferred zones of operation, and equipment stress limitations. These PPO augmentations improved long-term training and stability over A3C, prevented policy degeneration, and prevented "falling off the cliff" during training, ensuring convergence to better optima. Some of these techniques show how RL for real complex systems varies from the Open AI prototype models and can be applied to controlling other real-world cases.

Ocean waves have a complex spectrum but have a range of principal frequencies. We considered the standard wave periods from 6s to 16s, which resulted in variability in performance as the mechanical structure of WEC is optimized for the middle of the frequency spectrum.

Architecture choices

The heterogeneity and complexity of WEC require a versatile controller like Multi-Agent Reinforcement Learning (MARL). The three legs and the generators mounted for each of the legs act differently, as they tend to generate different amounts of energy based on the orientation of the mechanical structure and wave directionality. Simpler one agent RL with multiple actions failed to control the WEC effectively, which resulted in poor performance. Hence, separate agents of MARL were used to control the reactive force of the generators on one of the legs to learn the environment and policy better. The three-agent MARL has each agent controlling one of the three generators on the three legs, as represented in Figure 5. However, for simpler frontal waves, a simpler two-agent MARL was leveraged for faster training and bootstrapping of 3 agent. The objective for using two agents was to exploit the symmetry of the two back legs for frontal waves(Figure 4), which were each 60 degrees apart from the axis of symmetry and duplicate the agent for the back legs while keeping a separate agent for the front leg, which is aligned to the axis of symmetry. However, the default 3-agent MARL was chosen for its versatility to handle wavefront at different angles. This leads to optimal energy capture, with the challenge of convergence to better control policies.

Environment state design: We validated the inclusion of states in successive steps and evaluated the impact of the choices based on total rewards. The states included are represented in table 1. During training, the state information is provided as a vector represented by s in equation 3.

$$s = [\ e\ \dot{e}\ \ddot{e}\ g\ \dot{g}\ z\ \dot{z}\ ]^T \tag{3}$$

Where e represents the buoy position, g represents the tether extension, and z represents wave excitation. All RL agents share the continuous observation space of position and wave. The continuous action space for the individual RL agent is defined by the reactive force $f_{gen(i)}$ for the controlled generator, where "*i*" represents the index for the agent.

Further, specifying an appropriate reward function is fundamental to have the agent learn the desired multi-objective behavior. Hence the reward is defined as,

$$Reward_i = \alpha.(P_{own(i)} + \eta_i.P_{others}) + (1-\alpha)\ yaw \tag{4}$$

Where P represents the generated power defined by, $-f_{gen} * \dot{e}$. η is the hyperparameter for the team coefficient, and α is the hyperparameter for yaw minimization of individual legs.

Table 1: Environment states for RL.

position	position of the buoy with velocity and acceleration for the translational and rotational motion
yaw	rotational yaw motion to monitor stress
tether	extension and velocity of tether
wave	wave elevation and rate of change for present and 10s ahead in time from sensors

The following sections discuss the intuition behind formulating the reward function.

Cooperation vs. competition for agents

As the wave energy converter has a generator on each of the three-leg extensions and different RL agents controlling individual generators, it is important to include the power contribution from all the generators in the reward. Though it looks like a cooperative MARL problem on the surface, the disparity in the power generated by individual legs and the trade-off by one leg to get additional power in other legs makes the solution a combination of cooperation and competition. So we need flexibility in determining the extent to which we add the power generated by the other legs to rewards of individual legs. Also, we needed an option for adversarial contributions of power from the other legs in the reward. We termed this multiplier **team coefficient** "η", where positive value adds power generated by the other legs in the reward and vice versa, represented in equation 5.

$$Reward = P_{own} + \eta\ .P_{others} \tag{5}$$

Here, η = team coefficient, P_{own} is the generator's power being controlled, and P_{other} is power from other generators. We implemented a combined Bayesian hyperparameter search of the optimum 'team coefficient' of the individual agents for the back and front legs. We achieved best performance with 'team coefficient' of +0.8 for the agent for the back legs and -0.6 for the agent for the front leg, as shown in Figure 6.

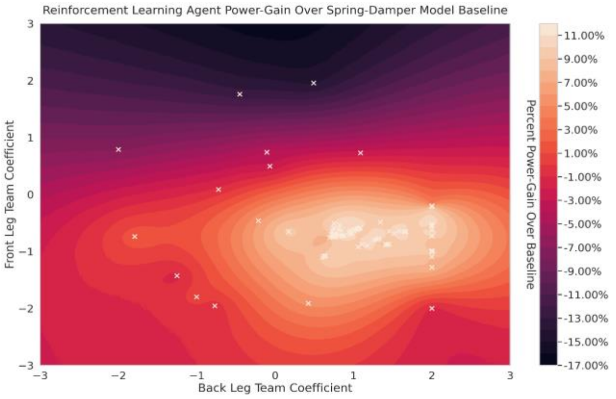


Figure 6: Team coefficient hyperparameter optimization for the front and back legs. The brighter region indicates optimal team coefficients.Bayesian optimization helps faster convergence.

Refinements to PPO for training stability and optimization

This subsection describes the implementation details of refinements, the effect of those modifications and explains the intuition behind the results.

One major challenge for effective multi-agent reinforcement learning is training stability and difficulty in convergence (Yu et al. 2021). Additionally, the design of the neural network architecture of the policy and critic network to accurately model the dynamics of the system is essential. A primary symptom of this instability during our initial experiments on the WEC environment was the inability to sustain good performance after achieving an initial peak. The average reward over episodes rapidly deteriorated as training progressed and never recovered. We mitigated this "falling off the cliff" with effective design choices, data transformation, and tuning as described below, which will also help tackle similar control problems for other use cases.

- **Clipped Surrogate Objective Function:** Proximal Policy Optimization (PPO) uses a clipped surrogate loss function in order to prevent the policy from taking large steps during training, avoiding "falling off the cliff" (Schulman et al. 2017). The ratio between the update policy output and the old policy output is given by the expression:

  $$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (6)$$

  Hence, the central objective function of PPO with clipping is defined as,

  $$L^{CLIP}(\theta) = E[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t)] \quad (7)$$

  where the second term $clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t$ clips the default policy gradient $r_t(\theta)\hat{A}_t$ to the range $[1 - \varepsilon, 1 + \varepsilon]$. Clipping serves as a regularizer by limiting the policy to change dramatically. The hyperparameter ε corresponds to how far the new policy can go from the old while still leveraging the objective.

  This is very effective at stabilizing multi-agent reinforcement learning and improving maximal performance on a variety of tasks (Yu et al. 2021; Brockman et al. 2016; Vinyals et al. 2017). We explore the effect of this clipping parameter on the WEC environment in Figure 7. We find that smaller clipping values helped stabilize training significantly, unlike the suggested ε=0.2, allowing for more monotonic performance improvement. This is because, for WEC, the action is highly sensitive to future rewards causing the probability density function (pdf) of actions to be dense. In addition to this aggressive clipping, we stabilize RL training with several other modifications to PPO.

- **Recurrent Neural Network**: As our observation for the environment state primarily includes information about the current state of the buoy and wave, we needed to use recurrent networks to estimate the time series of the incoming wave and adjust accordingly. Usage of recurrent network architectures in reinforcement learning enables estimation of the hidden state of a partially observable environment. We built and tuned a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) network instead of a basic fully connected neural network for both the policy and critic (larger network) to leverage the temporal hidden states for more robust policy. These LSTMs also had access to partial future wave elevation data from the wave sensors placed further into the ocean with a 10s anticipation time.

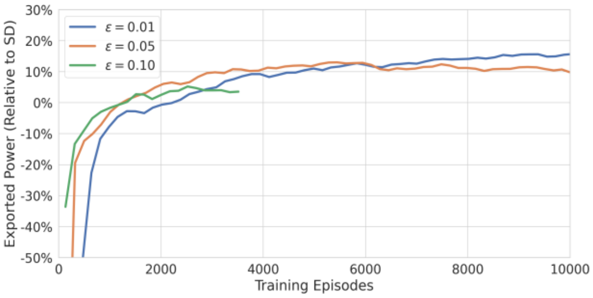


Figure 7: A comparison of power production with respect to the choice of PPO clipping parameter (ϵ)

- **Adaptive Normalization**: Due to the interacting reward terms, the value and policy loss may vary wildly with different parameters and configurations. To accommodate the variability in the reward term, we use an exponential moving average to normalize the reward from the environment by dividing by the standard deviation of the discount return (Andrychowicz et al. 2020). We also normalize the critic's outputs during training with the mean and standard deviation of the value function targets (Andrychowicz et al. 2020).
- **Finite horizon fix**: To accommodate different progressions of complex ocean waves with a variety of different principal frequencies, wave spreads, and heights, we train on many initial wave configurations. This necessitates limiting the episodes to a finite horizon and shorter lengths. However, this creates a discontinuity at the terminal state, as the terminal state is artificially chosen to optimize memory usage. In order to prevent introducing time-dependence to the return due to this finite horizon, we use the critic to estimate the future return of the terminal states instead of assuming the return will always be zero. This is especially necessary since the RL controller requires a very long rollout length. At 0.1 second resolution, the simulator requires 160 timesteps to cover one power cycle with a principal wave frequency of 16s.
- **Rollout length**: Due to the periodic nature of the environment, the current action tends to significantly affect states and rewards far into the future. We find that longer rollout lengths and smaller discount factors improved power generation, necessitating accurate, critical estimates at the terminal state.

```
Algorithm 1: Reinforcement Learning Training
Input:
Environment state: buoy position (6 degrees of freedom), tether extension + preprocessing for 1st and 2nd derivates
Excitation: ocean wave episodes from JonSwap spectrum
PPO Parameters: clipping parameter $\varepsilon$, rollout
Reward Parameters: yaw penalty $\alpha$, team coefficient $\eta$
    $Reward = \alpha * power + (1 - \alpha) * yaw$
Initialization: policy parameter $\theta_0$, value parameter $\phi_0$
Excitation parameters: wave height, freq, direction
Output: Optimized Policy and value DNNs
Let K= 0… N
  • Collect set of trajectories (state, action) by running policy $\pi_k = \pi(\theta_k)$ in the environment
  • Calculate reward $R_t$ and advantage $\hat{A}_t$ based on current value function. Calculate $r_t(\theta)$, which is the ratio between the updated policy and old policy output
  • Update the PPO policy with clipping
      $L^{CLIP}(\theta) = E\left[min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t\right)\right]$
  • Exponential moving average to normalize the reward
  • Critic to estimate the future return of the terminal state to accommodate finite horizon
  • Calculate Critic Loss and update value function
  • Action clipping for preferred operational limits
end
```

Design for Trust

The target wave energy converter is adversely affected by the rotational yaw motion. This spinning motion of the voluminous buoy causes the tether connections to wear out faster and has potential maintenance implications. The yaw motion is most significant when wavefronts hit the WEC at angles away from the axis of symmetry of the WEC, with 30 degrees presenting an extreme case based on deployments in various oceans. We account for this yaw movement using three independent agents of MARL, one for each tether, and by including an additional term in the reward proportional to the instantaneous yaw movement of the buoy.

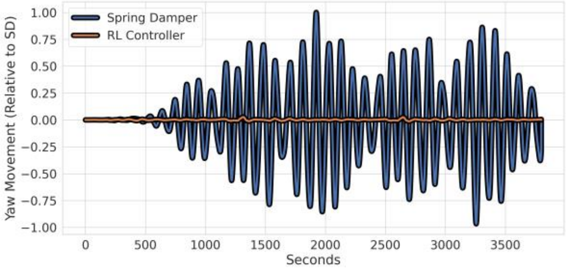


Figure 8: Comparison of Yaw movement between RL and the spring damper (SD) controllers for an episode with wave height of 2m and principal wave period of 12s. Values are relative to maximum SD yaw.

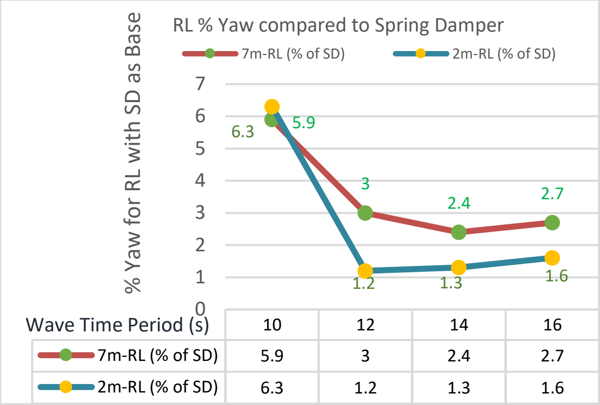


Figure 9: ***Yaw for RL as a percentage of Spring Damper's yaw (SD)*** *for normal (2m median height) and survival condition (7m height) waves at 30°*

The total reward is a weighted mixture of the power and yaw reward terms:

$$Reward = (\alpha)\ power + (1 - \alpha)\ yaw \quad (8)$$

$$power = P_{own(i)} + \eta_i . P_{others} \quad (9)$$

where α is a tunable yaw penalty hyper-parameter (lower the stronger), "i" represents the agent instance. This led to significant improvements in yaw reduction resulting in much less displacement than what is produced by the currently deployed spring damper controller, as shown in Figure 8 and Figure 9.

Although we expected the power generation to decrease with yaw control, we found that adding the penalty for yaw in the reward function improved power generation, as observed from Figure 10. We hypothesize that this is because yaw control is generally an easier task to perform in isolation. This allows the RL agent to quickly enter a stable regime, from which it can focus on improving power generation. A comparison of different values for α is presented in Figure 10. This combined reward serves the dual purpose of

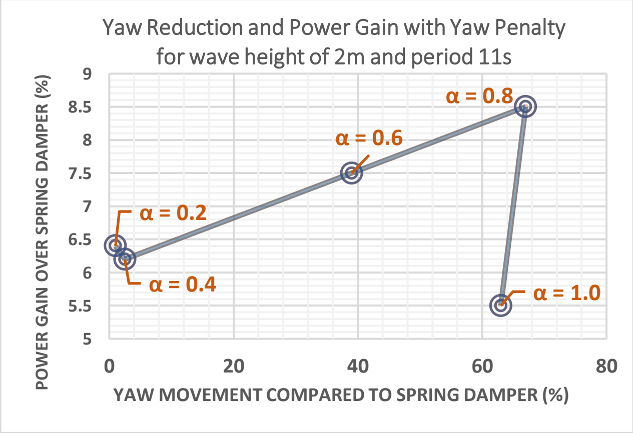


Figure 10: Tradeoff between Yaw and Power improvement for various values of the yaw penalty hyperparameter α (lower α puts higher emphasis on correction).

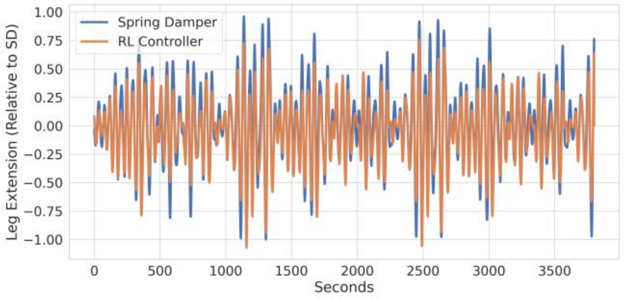


Figure 11: Tether extension across an episode of wave height 2m and wave period 12s after minimizing Yaw with reward shaping. Values are relative to maximum tether extension of the SD for the episode.

energy capture maximization while limiting the stress on the WEC to avoid costly maintenance in the open sea with these submerged structures. Also, Figure 11 shows that minimizing yaw with RL reduces the tether extension relative to SD, adding to the stability. This also impacts design, and minimizing mechanical stress can lower manufacturing costs, as in the case of wind energy, where lowering stress on blades has a similar effect on capital cost.

Assured ML: The target wave energy converter has limitations on the maximum tension in the spring extensions of the three legs anchored to the ocean bed. Also, there are limits placed on the maximum reactive force on the generator, based on the generator ratings. There are other limitations like the minimum allowed tension in the mooring tethers to ensure mechanical integrity. In this design, as we explored the maximization of the energy capture, we made RL adapt to the hard limits by implementing clipping on the RL action, which is the reactive force of the generator, as a safeguard guaranteeing stable operation of the WEC.

RL control during survival conditions

In addition to reducing typical maintenance, we find that the attributes of long episode horizons, a low discount factor, and the effective long-term planning empowered by the LSTM model significantly improved the RL controller's behavior in extreme and potentially dangerous conditions. We evaluated the controllers' behavior for the extreme wave height of 7m to analyze disaster events. Figure 12 shows that for 7m waves with a median time period of 12 seconds, the RL controller minimizes the yaw and reduces yaw to a level much lower than even the yaw for a 2m high wave with spring damper control. Additionally, we notice that the RL controller can better react to changing conditions within a wave episode and correct the yaw displacement, as seen in Figure 12, between times of 500 and 700 seconds into the episode (of similar nature). We notice that both for heights 7m and 2m, the spring damper was unable to successfully correct the yaw displacement because it was already in a suboptimal position from handling previous several wave peaks. But RL was able to successfully mitigate the yaw build up. Additionally, the RL controller can perform these corrective maneuvers while maintaining positive power production and avoiding sudden spikes in power consumption (motor mode) like the spring damper at height 7, further reducing strain on the system during these storm-like conditions.

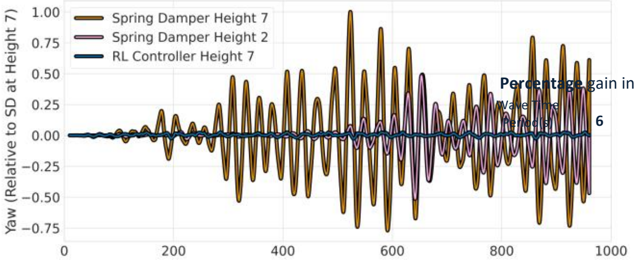


Figure 12: A comparison of Yaw for the RL and spring damper controllers on an extreme wave height of 7m and period 12s. The SD yaw for a wave of height 2m is also included as a reference.

The analysis also showed that compared to the default spring damper controller, the RL controller tends to influence the reactive force of the generator to pull in the tether extensions in extreme wave conditions. We notice an average pull on the tether extension of 0.379 meters compared to the spring-damper for 7m high waves. However, the RMS of the tether extension with respect to this mean remains consistent with the spring damper values.

Results

The CETO 6 wave energy converter (WEC) platform simulator was used for this work, which accurately models the mechanical structure, the mechanical response, the electro-mechanical conversion efficiency with losses for generator and motor modes, and the fluid dynamical elements of the wave excitation.

Wave data such as the distribution of principal time periods, height and spectrum were collected from Albany in Western Australia, Armintza in Spain (Biscay Marine Energy Platform: BiMEP), and Wave Hub on the north coast of Cornwall in United Kingdom. The wave generator model used in simulation uses well-es-tablished ocean wave spectrum like Jonswap which accu-rately models the heterogeneous components in ocean waves, letting the simulator sample the waves for training and evaluation. For evaluation, we used 1000 episodes for each principal wave period and height, where each episode covers 2000 sec of continuous wave

Table 2: Results of Power Gain for PPO MARL controller for various wave time periods with spring damper energy capture as base

Percentage gain in energy capture of RL over Spring Damper							
Wave Time Period(s)	**6**	**8**	**10**	**12**	**14**	**16**	**Avg**
0 °	41.4	26.9	16.9	13.4	11.7	9.2	**19.9**
30 °	15	19.8	9.3	11.8	16.5	11.1	**13.9**

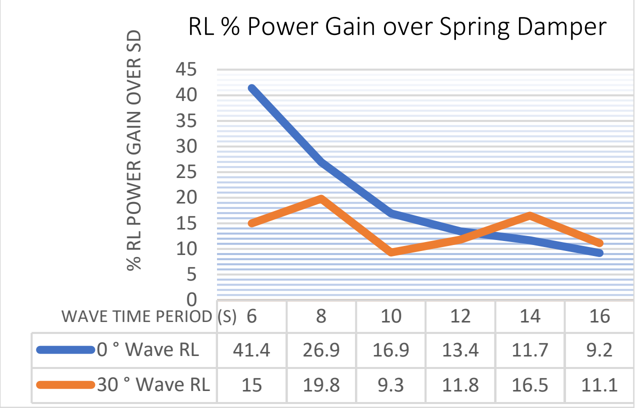


Figure 13: ***Percentage power gain for RL*** *controller over Spring Damper for 2m high waves of different time periods*

data in steps of 0.2 sec for Reinforcement Learning (RL) loop and 0.05 sec (4x) for simulation response. For each training run, there are roughly 50 million steps for convergence, with 2 thousand such training runs required for hyper-parameter optimization and model search with early stops. For regular operation, we show results of median wave height of 2m for the entire wave frequency spectrum spanning time periods of 6s to 16s.

The power generated by the baseline spring damper controller with resonant spring constant and damping constant for a wave time period and height is used as a reference for evaluation to estimate the gain of energy capture by RL controllers as a percentage improvement. A direction of 0° indicates frontal waves with the wavefront aligned with the front leg, as shown in Figure 4. For evaluation, we used the same seed for sampling waves for episodes between RL and SD.

Table 2 and Figure 13 shows a significant improvement in captured power with RL controller over baseline spring damper (SD) controller for the entire frequency spectrum of ocean waves. The MARL performs better for frontal waves at (0°) than the waves at an angle of 30°. However, the 3-agent RL performs better than spring damper controller for all angled waves, including an extreme angle of 30°. We also observed that there was variations in gains by RL controller with different time periods when compared to spring damper model that is resonantly tuned to the mechanical structure of the WEC for a certain frequency band.

Table 3: Results of Yaw minimization for RL over Spring Damper (SpgDp)

Yaw (RMS) for wave height of 2m					
Wave Time Period(s)	**10**	**12**	**14**	**16**	**Avg**
Normal wave height of 2m (maintenance focus)					
RL (% of SD)	6.3	1.2	1.3	1.6	2.6
Extreme wave height of 7m (survival conditions)					
RL (% of SD)	5.9	3	2.4	2.7	3.5

Table 3 shows that 3-agent MARL almost eliminated the yaw, which causes mechanical stress, while still making significant energy capture gains over baseline spring damper, as shown in Table 2. Table 3 also shows that for natural disasters with surging waves of 7m height, the 3-agent MARL can almost eliminate yaw, just like it did for waves of normal height.

Ablation study

Table 4 compares the results of different RL algorithms (SAC, A3C) with PPO for frontal waves. As DQN has issues with using continuous action space, it was not suitable for this environment setting.

One of the main observed drawbacks of SAC and A3C was the "falling off cliff" problem. During training, the power gains for the RL controller for different wave periods peaked at different points for SAC and A3C and eventually fell off the cliff making them hard for deployment. On the other hand, for PPO, with the clipped surrogate objective function as highlighted in the paper, the training stability was maintained for all wave time periods which led to better convergence and one single deployable checkpoint. Finally, PPO has good convergence for all angled waves, while A3C failed to have uniform convergence.

Table 4: Results of Power gain for various RL Agents over Spring Damper for Frontal 0° Waves

Percentage gain in energy capture of RL over Spring Damper							
Wave Time Period(s)	**6**	**8**	**10**	**12**	**14**	**16**	**Avg**
SAC	17.2	13.7	0.5	-1.4	1.1	3.2	3.2
A3C	35.5	27.3	7.0	8.3	7.2	3.6	14.8
PPO	41.4	26.9	16.9	13.4	11.7	9.2	**19.9**

Intuition behind the performance improvement for RL

The normal intuition is that the reactive forces for the generator on the legs will be proportional to the velocity of the tether as energy is captured working against this motion. However, the RL controller is fuzzier about it as shown in Figure 14 and Figure 15, implying that it takes a long-term view and compromises short-term objectives for greater

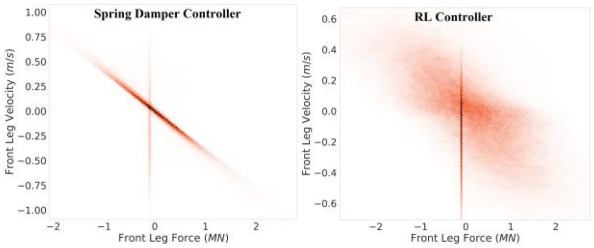


Figure 14: Generator Reactive force vs Velocity of tether: Front legs for 0° wave front

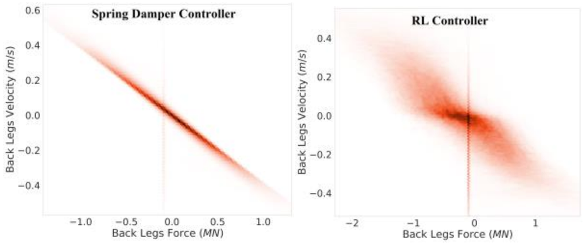


Figure 15: Generator Reactive force vs Velocity of tether: Back legs for 0° wave front

gains on energy capture at the more opportune segments of the wave cycles.

Computation Complexity

The RL inference model accelerated with Intel Math Kernel Library and Streaming SIMD Extensions takes 100% of one hyperthreaded core of Xeon Gold 6246R at 3.4GHz to run which has 32 hyperthreaded cores and comes to 2.2 GOps for floating point math. For deployment, the targeted embedded platforms will require a rating of around 1.8x 2.2GOps given the slimmed down features of the CPU.

Each full training takes about 1-3 days on a server with 2x Xeon Gold 6246R at 3.4GHz (2x71.5GOps) with 8x Nvidia Volta 100 GPUs.

Conclusions

The proposed MARL controller yielded double-digit gains over the entire spectrum of waves boosting higher energy production. At the same time, it helped reduce mechanical stress, which impacts maintenance and operating costs, and actively mitigated adverse effects of high waves characteristic of disaster events, helping to preserve capital investment and cost of manufacturing.

This RL controller can be extended in its scope from an individual wave energy converter to a cluster of wave energy converters to address the “cross-talk” issue that is a well-established problem in wind farms. MARL architectures, refinements, and optimizations described in this paper, with multiple objectives mentioned above, are applicable to many other clean energy problems like wind energy.

Further, the proposed MARL architecture and the PPO refinements to stabilize training for global optima described in this paper can be used in many other complex control applications with multiple actors or entities to control.

References

Anderlini, E.; Forehand, D.; Bannon, E.; and Abusara, M. 2017a. Reactive control of a wave energy converter using artificial neural networks. International Journal of Marine Energy, 19: 207–220.

Anderlini, E.; Forehand, D.; Bannon, E.; Xiao, Q.; and Abusara, M. 2018. Reactive control of a two-body point absorber using reinforcement learning. Ocean Engineering, 148: 650–658.

Anderlini, E.; Forehand, D. I. M.; Bannon, E.; and Abusara, M. 2017b. Control of a Realistic Wave Energy Converter Model Using Least-Squares Policy Iteration. IEEE Transactions on Sustainable Energy, 8(4): 1618–1628

Anderlini, E.; Forehand, D. I. M.; Stansell, P.; Xiao, Q.; and Abusara, M. 2016. Control of a Point Absorber Using Reinforcement Learning. IEEE Transactions on Sustainable Energy, 7(4): 1681–1690.

Anderlini, E.; Husain, S.; Parker, G. G.; Abusara, M.; and Thomas, G. 2020. Towards Real-Time Reinforcement Learning Control of a Wave Energy Converter. Journal of Marine Science and Engineering, 8(11).

Bouville, M. 2008. Crime and punishment in scientific research. arXiv:0803.4058.

Clancey, W. J. 1979. Transfer of Rule-Based Expertise through a Tutorial Dialogue. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Clancey, W. J. 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI83), 556–560. Menlo Park, Calif: IJCAI Organization.

Clancey, W. J. 1984. Classification Problem Solving. In Proceedings of the Fourth National Conference on Artificial Intelligence, 45–54. Menlo Park, Calif.: AAAI Press.

Clancey, W. J. 2021. The Engineering of Qualitative Models. Forthcoming.

Duan, Y.; Chen, X.; Houthooft, R.; Schulman, J.; and Abbeel, P. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control. arXiv:1604.06778.

Engelmore, R.; and Morgan, A., eds. 1986. Blackboard Systems. Reading, Mass.: Addison-Wesley.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. 1984. Strategic explanations for a diagnostic consultation system. International Journal of Man-Machine Studies, 20(1): 3–19.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2019. Continuous control with deep reinforcement learning. arXiv:1509.02971.

NASA. 2015. Pluto: The' Other' Red Planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.

Rice, J. 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ.

Robinson, A. L. 1980. New Ways to Make Microcircuits Smaller. Science, 208(4447): 1019–1022.

Yusop, Z. M.; Ibrahim, M. Z.; Jusoh, M. A.; Albani, A.; and Rahman, S. J. A. 2020. Wave-Activated-Body Energy Converters Technologies: A Review. Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 76(1): 76–104.

Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.; and Wu, Y. 2021. The surprising effectiveness of mappo in cooperative, multi-agent games. arXiv preprint arXiv:2103.01955.

Schulman, J.; Wolski, F.; Dhaliwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms.arXiv:1707.06347.

Vinyals, O.; Ewalds, T.; Bartunov, S.; Georgiev, P.; Vezhnevets, A. S.; Yeo, M.; Makhzani, A.; K¨uttler, H.; Agapiou, J.; Schrittwieser, J.; et al. 2017. Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:1708.04782.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.;Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. arXiv preprint arXiv:1606.01540.

Tan, M. 1993. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In Proceedings of the Tenth International Conference on Machine Learning, 330–337. Morgan Kaufmann.

Sergiienko, N. Y.; Neshat, M.; da Silva, L. S.; Alexander, B.; and Wagner, M. 2020. Design optimization of a multi-mode wave energy converter. In International Conference on Offshore Mechanics and Arctic Engineering, volume 84416, V009T09A039. American Society of Mechanical Engineers.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term-memory.Neural computation, 9(8): 1735–1780.

Andrychowicz, M.; Raichuk, A.; Sta´nczyk, P.; Orsini, M.;Girgin, S.; Marinier, R.; Hussenot, L.; Geist, M.; Pietquin,O.; Michalski, M.; et al. 2020. What matters in on-policy reinforcement learning? a large-scale empirical study.arXivpreprint arXiv:2006.05990.

Rijnsdorp, D.P., Hansen, J.E. and Lowe, R.J., 2018. Simulating the wave-induced response of a submerged wave-energy converter using a non-hydrostatic wave-flow model. *Coastal Engineering*, *140*, pp.189-204.

“Ceto Technology.” Carnegie, 14 Dec. 2020, https://www.carnegiece.com/ceto-technology/.

Shi, W., Feng, Y., Huang, H., Liu, Z., Huang, J. and Cheng, G., 2021. Efficient hierarchical policy network with fuzzy rules. *International Journal of Machine Learning and Cybernetics*, pp.1-13.

Lim, H.K., Kim, J.B., Heo, J.S. and Han, Y.H., 2020. Federated reinforcement learning for training control policies on multiple IoT devices. *Sensors*, *20*(5), p.1359.