Cost of Energy Optimised by Reinforcement Learning

WES Control Systems Stage 2
Public Report

MaxSim

This project has been supported by Wave Energy Scotland

1 Project Introduction

The CEORL project, (Cost of Energy Optimised by Reinforcement Learning) has been investigating the use of reinforcement learning (RL) to develop control policies for wave energy converters (WECs). Stage 2 has demonstrated encouraging results: a step change in energy capture, while at the same time reducing peak loads. It is necessary to improve both in order to reduce the levelised cost of energy (LCOE) of wave power.

The main consortium members are Paul Stansell (MaxSim), Max Carcas (Caelulum), and Alex Hagmüller and Max Ginsburg (Aquaharmonics). Richard Crozier (Reoptimizesystems), Alexandra Price (Wave Conundrums Consulting), David Forehand (University of Edinburgh) and Jos van t' Hoff (Marine Systems Modelling) have been working closely with the consortium. We have also had consultancy advice from David Pizer, Ross Henderson (Quoceant on behalf of the Quantor project), Chris Retzler (Mocean), and Antoine Baudoin, Hervé Gaviglio and Jørgen Hals Todalshaug (CorPower).

2 Description of Project Technology

2.1 Key Features

The four most important features of the CEORL control concept are as follows:

1. RL is used to learn a control policy that specifies the power take off (PTO) impedance of a WEC.
2. The algorithm adjusts the policy with the aim of maximising the long term sum of rewards. Rewards that aim to improve the LCOE, rather than solely maximising energy capture, have been explored.
3. The inputs to the controller are only those sensor data that are available in real-time on a real WEC.
4. The policy allows for continuous fast control of the generator demand signal. Unlike many standard control policies, the PTO impedance is not a fixed value for a given sea-state.

Each of these four key features will now be described in turn.

2.2 Reinforcement Learning

During the training process, the RL agent (see Figure 1) continually updates the control policy, in order to maximise the sum of the rewards over the long term. The input to the policy is known as the *state* and the output is called the *action*. A reward is incurred with every action taken, and can include negative components to represent penalties. Neural networks can be used to learn how an action taken in a given state is likely to affect future rewards.

RL has been successfully applied in many areas, including dynamic control, and has been shown to outperform human-designed solutions. There are open-source resources available for development, and every year the research community adds improved or more specialised algorithms. There is a wide range of methods which can be used to address particular problems. For example, there are risk-adverse training methods which can be used to avoid taking risky actions. Other methods are sample-efficient, enabling fast learning. Work has also been done on the transfer of policies trained in one environment to a slightly different environment. This suggests that RL can be pre-trained on a simulation, and the policy can then be honed on a real device.

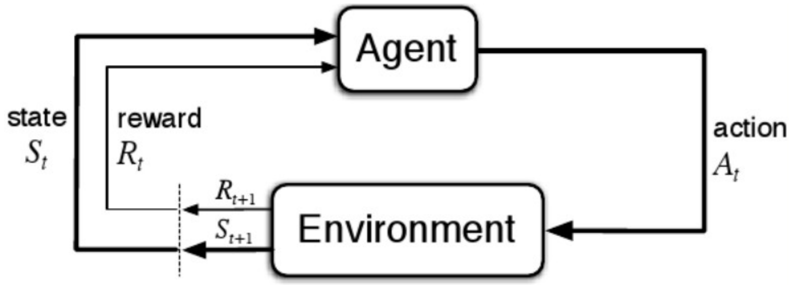

Figure 1: Schematic of the closed-loop agent-environment interaction in reinforcement learning (from Sutton and Barto [2017])

2.3 Rewards Aimed at Improving LCOE

Often the economic viability of a WEC is described in terms of a number of competing metrics, such as capture width and peak loads. WES lists about eighteen such metrics, many of which are tightly coupled, meaning that a control policy that improves one is likely to worsen another. Figure 2 illustrates an example where an increase in peak power capture is accompanied by an increase in peak loads.

The maximum revenue (minimum of LCOE) does not occur at the maximum power capture potential (shown here as energy capture given 100% availability) because once the loads increase beyond their design values, the

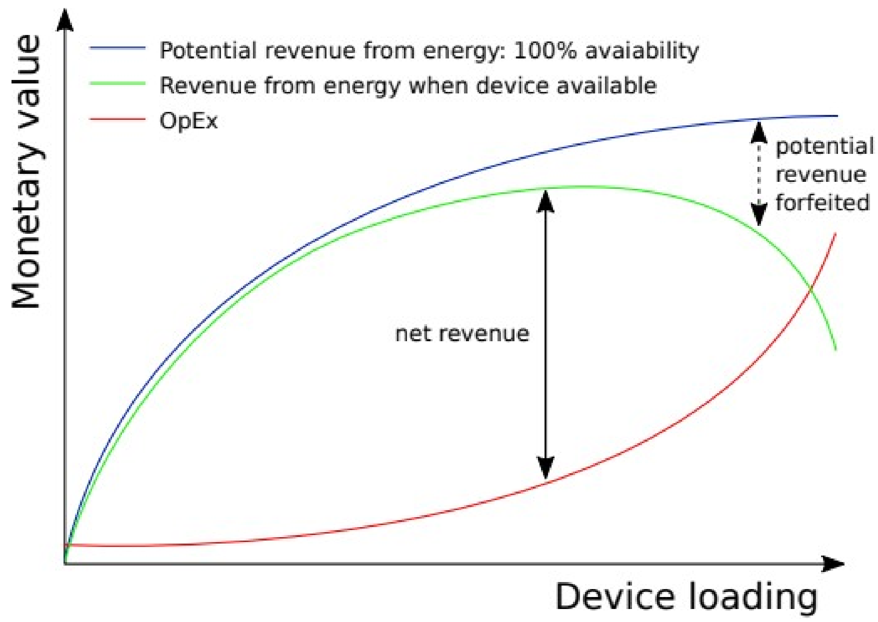

Figure 2 Illustrative graph showing how control actions that increase the power capture potential in an existing WEC may increase device loading. If these loads exceed design loads, this may reduce overall annual energy production (by reducing availability) and increase operational expenditure.

potential revenue forfeited due to the design not being available to operate, and the maintenance costs, increase faster than the power capture potential. In principal, it can be seen that there is an optimal level of peak performance which minimises LCOE.

The CEORL project acknowledges that control which focusses exclusively on improved energy capture may not necessarily improve LCOE. Stage 2 demonstrated that it is possible to optimise for different rewards or combinations of reward. It is possible to set a reward function that rewards energy capture and penalises excessive loads and excursions.

2.4 Realistic Sensor Data

CEORL control policies use inputs (the *states*) that are representative of real sea conditions and constraints. Furthermore, only sensor data that can be collected on-board a device in the ocean will be used, so it does not rely on information from the shore or an off-device instrument for measuring surface elevation. Our goal is a control policy that does not depend on functioning communication links. The policy will not take the sea-state parameters (e.g. significant wave height and period) as inputs, as these are typically measured by a separate buoy.

In Stage 3 the RL-derived policies will be tested in a wave tank. Care will be taken to use only information that would be available in full-scale WECs operating in real seas. Hence the states will not include the wave elevation in the absence of the device (typically measured during the tank calibration) or absolute position of the buoy (as measured by a ground-referenced visual motion capture system) as this won't be available in a commercial full-scale WEC.

In Stage 2, the estimate of power capture used in the reward was calculated with a simple model for PTO losses. In Stage 3, a more accurate model of the mechanical power chain and electrical losses will be used. This will ensure that we are optimising the policy for the electrical power delivered, an essential consideration for a system with reactive control.

2.5 Fast Continuous Control of Demanded PTO Impedance

CEORL control policies provide the demand signal for the generator. The generator's own controller converts this into a demand torque which it attempts to meet using standard techniques such as PID control. Using PTO impedance (e.g. by providing a PTO damping and a PTO spring coefficient), as the demand signal, rather than torque, results in a more stable, robust system. The CEORL system is intended as a sub-system of a standard supervisory controller. By focussing on the algorithm that provides the demand signal, rather than a whole bespoke system, we are avoiding redesign of standard equipment, and ensuring that CEORL can be integrated into existing systems. Many WEC developers already have preferred PTO and supervisory control systems. Switching to new supervisory control systems or PTOs introduces risk, and costs time and money. Offering WEC developers only the demand controller provides the best value for money and gives CEORL access to a larger market.

CEORL's policies can change the PTO impedance as fast as the generator is able to respond. We have observed that the policies tend to make better use of the full range of allowable stroke compared to conventional control that uses fixed-coefficients in each sea-state. In this way, RL-derived policies are able to capture more energy, while attracting lower peak loads.

2.6 Anticipated Benefits

We foresee the following benefits of the system once it is mature:

- Control policies will not be specific to a sea state. There will not be problems with switching between sea states, operating in changing sea states, operating in sea states inadequately represented by spectral parameters, such as bimodal spectra, or adjusting to the natural groupiness of waves.
- Control policies will not depend on functioning communication links for their basic functionality. They will be designed to work well using only data collected on-board. However, in the future, we may consider policies that can be enhanced by use of external data where this is available.
- Control policies will not depend on forecasted data (e.g. calculated from Lidar measurements) for its basic functionality. They will be designed to work well using real-time data. However, in the future, we may consider policies that can be enhanced by the use of external data where this is available.
- Rewards can be chosen to represent the trade-off between volumetric efficiency and reliability.
- Rewards can be chosen so that their long term sum gives the best available proxy for LCOE or other measures of interest.
- The formulation of rewards can be adapted to the technology readiness level; they can be updated as more information becomes available or priorities change.
- Control policies could be updated throughout the WEC's life. Reasons for making updates include more information becoming available about the reliability of components or the impact of reliability on LCOE, and changing behaviour due to wear, faults, biofouling or repairs.
- Control policies could be tailored to specific WEC devices. This can help the integration of separately developed sub-systems into one unit.
- Once the ability of the control policy to keep loads, stroke and power within design limits has been demonstrated in real seas, future design iterations of WECs could include this policy in the optimisation. This will result in a second step-reduction in LCOE. Mechanisms include CAPEX reduction, upscaling, or choosing appropriate relative sub-system ratings.

3 Scope of Work

3.1 Key Activities

The following activities were carried out in Stage 2:

- Built WEC numerical models of increasing complexity.
- Made these simulations usable as environments in OpenAI Gym.
- Investigated different RL algorithms.
- Studied different types of reward.
- Studied robustness with respect to transferability, delay and noise.

- Investigated metrics for Stage 3.
- Conducted FEED studies for Stage 3 hardware-in-the-loop and tank testing.
- Planned Stage 3 objectives, activities and partnerships.

3.2 The Numerical Modelling Activities

The first major challenge for the project was to integrate modelling tools written in different languages. Many device developers, including AquaHarmonics, have models of their device in WECSim, an open-source modelling toolbox for Simulink. However, the toolkit for developing RL policies, OpenAI Gym, makes use of libraries of RL algorithms, written in Python, such as OpenAI Baselines and TensorFlow. OpenAI Gym provides several standard *environments* (simulators) and also allows users to create their own environments. The first challenge was to set up a WEC simulator which would interface with RL algorithms written in Python. In Stage 2, the priority was a simulation that was fast to run and quick to set up. For this reason, the models used for RL were written in Python.

We started with a simple single degree of freedom, mass-spring-damper model of a heaving buoy, and increased the complexity iteratively. The final modelling iteration (the *Sophisticated Model*) modelled three degrees of freedom (surge, heave and pitch) and used the state-space method to replace the radiation convolution integral. Future modelling improvements will extend the Sophisticated Model.

3.3 The Reinforcement Learning Investigations

Figure 1 shows the general RL training method. In Stage 2 of the CEORL project, the environment was our WEC simulator, and the agent was the RL algorithm that learnt the best actions to take given the state. The state was simulated sensor data, the actions were the demanded PTO damping and spring coefficients, and the reward, in the simplest case, was the sum of the energy and a penalty for excessive relative heave motion.

The Stage 2 program was designed to demonstrate the viability of the CEORL process, and meet the requirements outlined in the WES Stage 2 Application Guidelines.

The first step was to choose a training algorithm that was suitable for our desire for continuous, as opposed to discrete, states and actions, and which would converge for our environment. Figures 3-4 show examples of how the policies that are learnt over a number of *episodes* (iterations) eventually converge on a solution where the objective function, the sum of rewards over each episode, cannot be significantly improved by further learning.

Next, a range of studies were undertaken with the aim of satisfying WES requirements:

- All of the RL work contributed in some way to demonstrating that a step change improvement was possible using a policy learnt using RL algorithms, given suitable rewards.
- The reward studies demonstrated that the improvements in power capture did not have a perverse impact on LCOE by massively increasing loads. Hence it was shown that it was possible to increase power capture without significantly increasing the peak loads.
- We demonstrated policy transferability to provide reassurance that the policy was still applicable in simulators that had differences to the one it was trained on.

- We demonstrated that the method is robust to sensor noise. Indeed, the addition of noise improved the training of the policy, which is a common phenomenon when learning from data.
- We demonstrated that the policy is stable and robust to typical actuator delays, a key problem for traditional control methods.

WES had also emphasised the importance of good metrics. To get the most out of RL, the most relevant metrics should be used as rewards for training. We investigated how LCOE and ROI (return on investment) could be used as rewards. We found the ROI to be more suitable as a reward than LCOE because it was the difference rather than quotient of terms related to annual energy and peak loads.

3.4 Planning of Stage 3 Activities

The work on the FEED studies for the hardware and software of Stage 3, along with a detailed plan of activities, was conducted according to WES guidelines. Other WES-funded projects were consulted: Mocean, Quantor and CorPower. They were asked about design of hardware-in-the-loop (HIL) rigs for WECs, and about what types of technical work they needed to see in order to give them confidence in the technology.

4 Project Achievements

The following advances were demonstrated:

- A step-change increase in energy capture in comparison to fixed-coefficient control that had been optimised for that particular sea-state.
- The ability to limit motions that exceeded a specified threshold.
- In addition to increasing energy capture and avoiding stroke threshold exceedence, lowering PTO forces (peak loads or number of cycles).
- Robust policies that can tolerate noise added to the state and action spaces.
- Stable policies that can tolerate typical actuation delays.
- Good policy transfer.
- Fast training.

Details of these achievements will now be elaborated.

4.1 Increase in Energy Capture while Avoiding Stroke Thresholds

Figure 3 shows a typical convergence plot for a model with simple rewards. Each episode was a 128 s simulation for the same Bretschneider spectrum with T_p = 9.5s, but with different phases.

The red dots show energy capture from an individual episode; the green dots show the *return*, i.e. the sum of the rewards over each episode. In this case the reward had two components: a positive component equal to the energy capture, and a negative component to penalise a stroke threshold. This penalty was zero if the difference between the wave elevation and the heave position was under the threshold. If the difference was greater than the threshold, the penalty was proportional to the amount by which the threshold was exceeded.

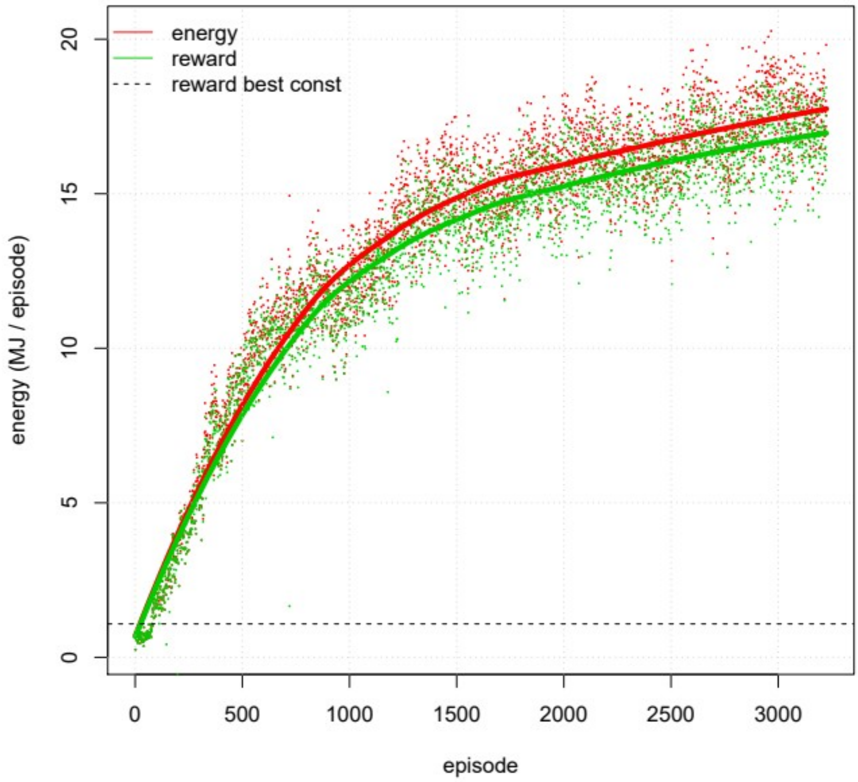

Figure 3: An example of a convergence plot from applying an RL algorithm to a WEC operating in a Bretschneider spectrum with T_p = 9.5 s. This plot shows the scatter of actual values of absorbed energy and reward about the smoothed curve through the points. The dashed black line marks the reward from the best control that is constrained to use constant damping and spring settings.

The trajectory of the solid red line (a running average) shows how updates to the policy during the training process result in more energy capture. It can be seen that for the first few hundred episodes the policy is improving by increasing the energy capture (i.e. reward and energy are the same). Thereafter, the reward is lower than the energy, but the energy still increases with training. This means that the policy is learning how to both avoid excessive excursions and increase energy capture.

The dashed black lines in Figures 3-4 mark the rewards from the best constant control settings. A grid-search was performed over each each control spring and damping setting to find the maximum reward. Each setting was tested for 100 episodes of the spectrum (with different random phases) and the average reward was used to find the optimum values.

4.2 Controlling Maximum Loads

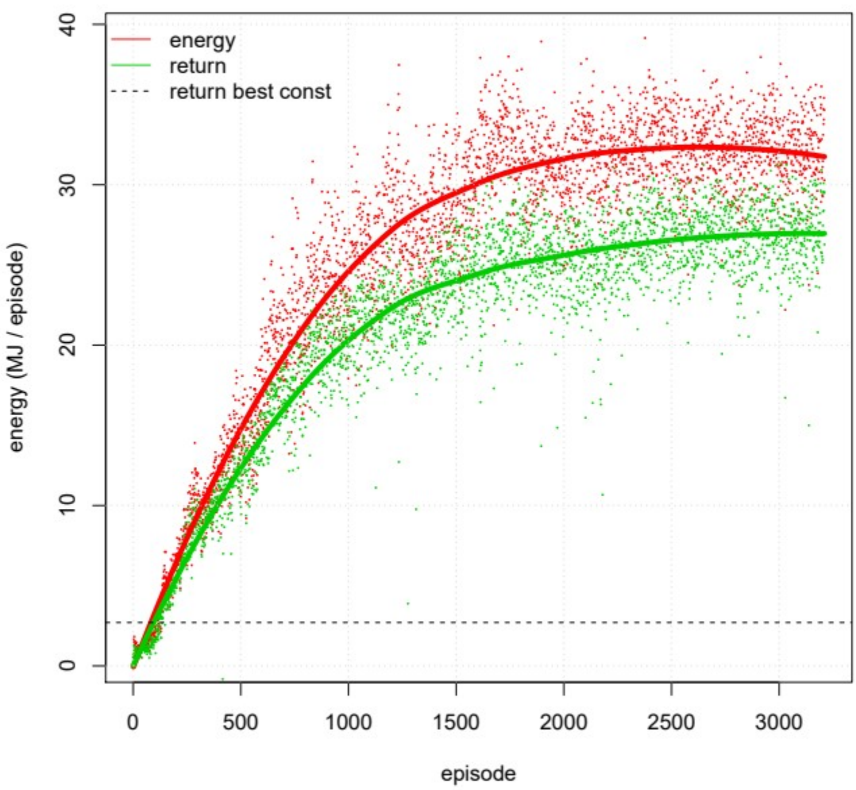

Figure 4: Example convergence plot for a policy learnt by incurring negative rewards when the F_c > 150 kN. The dashed black line marks the return from the best constant control settings without the force constraint.

Figure 4 shows a convergence plot for RL using three reward components: a positive component equal to energy, a penalty for exceeding relative stroke thresholds, and a penalty for exceeding a PTO load threshold. Different force thresholds were applied, e.g. Figure 4 shows policies where a penalty is applied for a control force (F_c) exceeding 150 kN. Comparing Figure 3 to Figure 4, it can be seen that adding a force threshold has increased the gap between energy (red) and return (green). This shows that the RL-derived policies do not apply a hard force limit; the penalty for exceeding this force is activated.

The question then is how well the RL-derived policy limits forces. The best way to answer this is to look at a probability density function. Figure 5 shows the probability density for four cases: the best constant co-efficient control (corresponding to the dashed black line in Figure 4), an RL-derived policy with unlimited forces (a policy of the type shown in Figure 3), an RL-derived policy with the force threshold limited to 150 kN (the case shown in Figure 4), and an RL-derived policy with the force threshold at a higher limit of 500 kN.

There are three important things to come out of this figure.

1. It can be seen that setting a load limit results in a narrowing of the load probability density.
2. The threshold load is a soft limit: the policies penalise exceedence of the force limit, but do not prevent it. The highest loads are around four times the force limit. If we wished to narrow the gap between the highest loads and the force limit, we would increase the weighting of the force penalty with respect to the energy-related component of the reward. This is an interesting direction for future work.
3. For the particular cases examined in the Stage 2 work, the RL-derived policies resulted in higher loads than the baseline case. Our aim initially was to limit loads with respect to other RL-derived policies. However, it was realised that for a more equitable comparison to the baseline case, the loads should be limited to the upper limits experienced by the baseline. For this reason, Stage 3 will consider two measures of success: first, increasing power capture while not exceeding the upper limits of load and stroke, compared to the baseline, and second, decreasing the upper limits of load, while not reducing power capture or increasing (or exceeding) the upper limits of stroke.

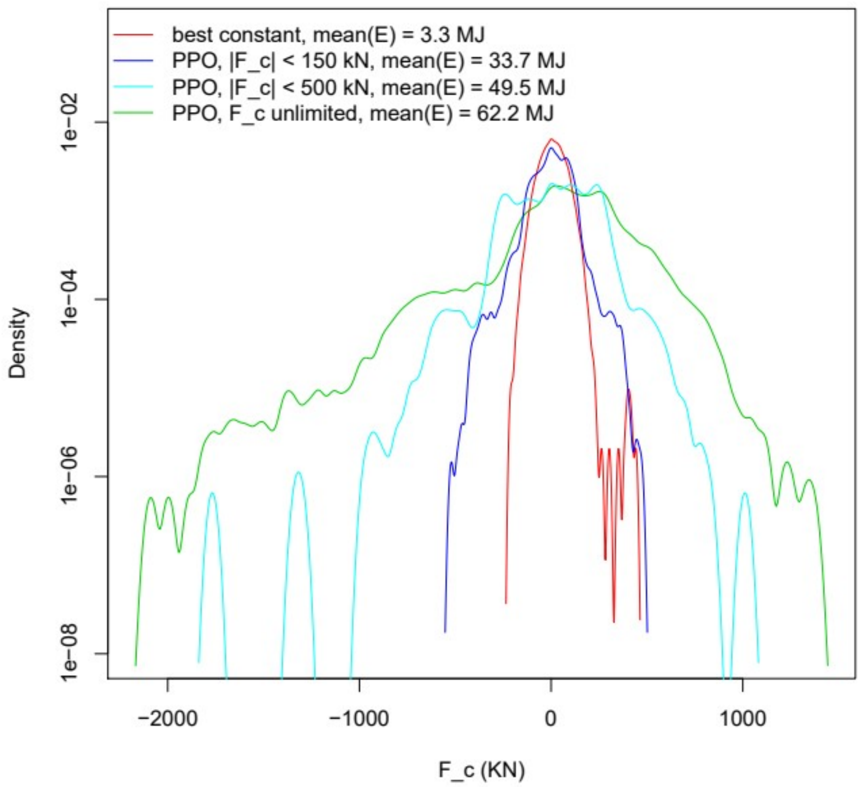

Figure 5: Probability densities of control force, F_c, for different threshold control forces.

4.3 Controlling Force Reversals

It was observed that the RL-derived control policies switched the control parameters on and off at a high frequency. There was a concern about whether this was implementable, which was explored by adding a delay to the numerical model (the RL 'environment') (see Section 4.5). There was also a concern that this could have an impact on fatigue.

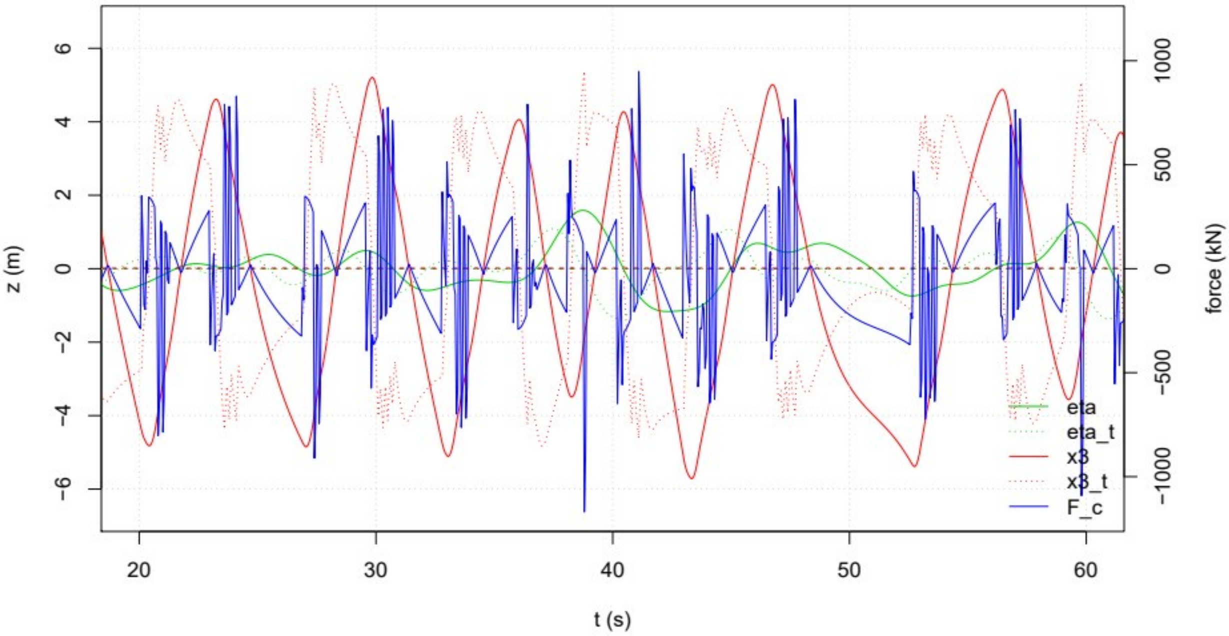

Figure 6: An example time-series plot for a control policy that does not penalise force reversals, showing the wave elevation (eta), the heave displacement (x3), the heave velocity (x3_t), and the control force (F_c). Despite being in the legend, eta_t is not plotted.

Figure 6 shows a close-up of wave elevation, position, velocity, and control (PTO) force for a typical RL policy. It can be seen that the control force contains pulses under a second long. It can also be seen that the velocity is closer to a square wave than a sinusoid. The mechanism for 'squaring out' velocity appears to be the short pulses of control force. Every time there is a velocity zero-crossing, there is burst of control force pulses, alternating between positive and negative. These pulses end at the next position zero-crossing.

Figure 7 shows a close-up of a simulation using the same WEC and wave elevation, but a policy that penalises control force reversals. It can be seen that the number of control force zero-crossings has indeed decreased. The burst of pulses that starts at each velocity zero-crossing and ends at each position zero-crossing in Figure 6 is replaced in Figure 7 by something akin to a running average. Despite the absence of the pulses in the policy show in Figure 7, the velocity has a similar form to that shown in Figure 6: it is much closer to a square wave than a sinusoid.

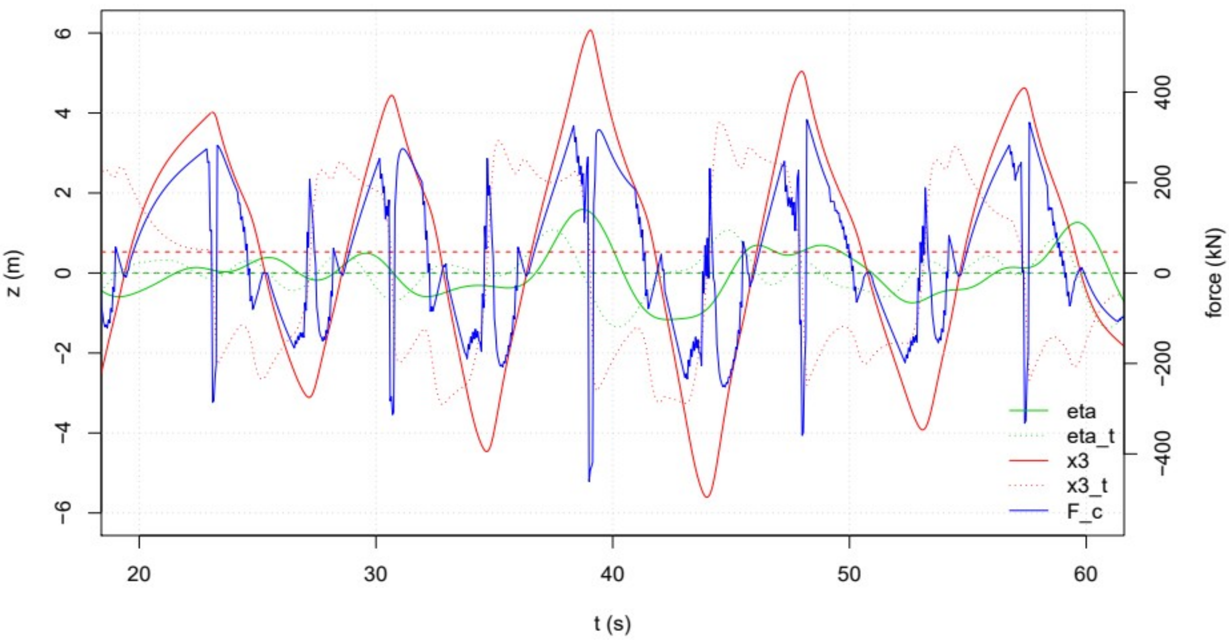

Figure 7: An example time-series plot for a control policy that penalises force reversals, showing the wave elevation (eta), the heave displacement (x3), the heave velocity (x3_t), and the control force (F_c). Despite being in the legend, eta_t is not plotted.

4.4 Customised Rewards

Previously the rewards for training RL had simply been the sum of weighted components, encouraging energy capture or penalising behaviour that had an impact on cost. It was not clear how the resulting sum related to levelised cost of energy (LCOE). In these trials we explored rewards that were an intentional proxy for LCOE. As RL aims to maximise rewards, the proxy was for the inverse of LCOE.

The intension was not to model LCOE exactly, but to build a reward function that used the elements of LCOE that were relevant to control. The first concession was to use a fixed value of CAPEX, i.e. to consider a device that has already been designed and built. The goal, then, is to optimise operation of that device, rather than optimise the device itself. As well as being a simpler case, it is also realistic in terms of a first step for WEC developers who might be potential customers. We assumed a discount rate of 0% to avoid unnecessary complexity at this initial round of modelling.

The next step was to define the probability of the device not being available to operate due to a fault. By assuming a fixed daily operations and maintenance charge, it was then possible to model both the OPEX and device availability as functions of the probability of unavailability.

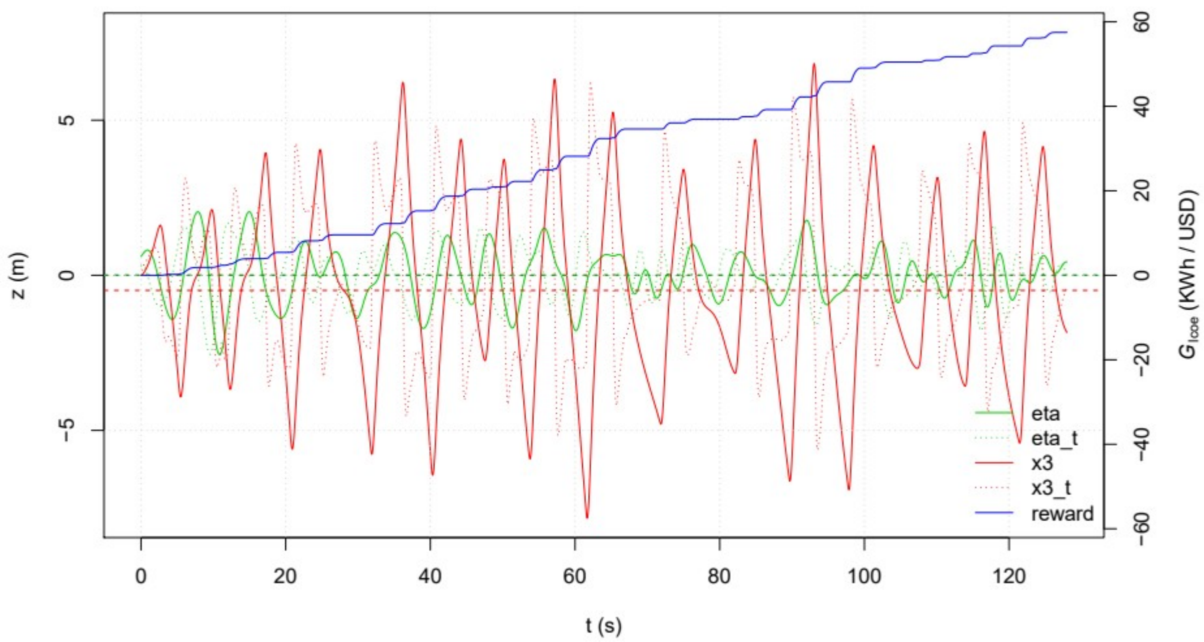

Figure 8: A threshold control force of 500 kN: example time-series plots for a single episode run using a control policy that optimised G_{lcoe}, which is a proxy of 1/LCOE. The plot shows the wave elevation (eta), the heave displacement (x3), the heave velocity (x3_t), and the control force (F_c). Despite being in the legend, eta_t is not plotted.

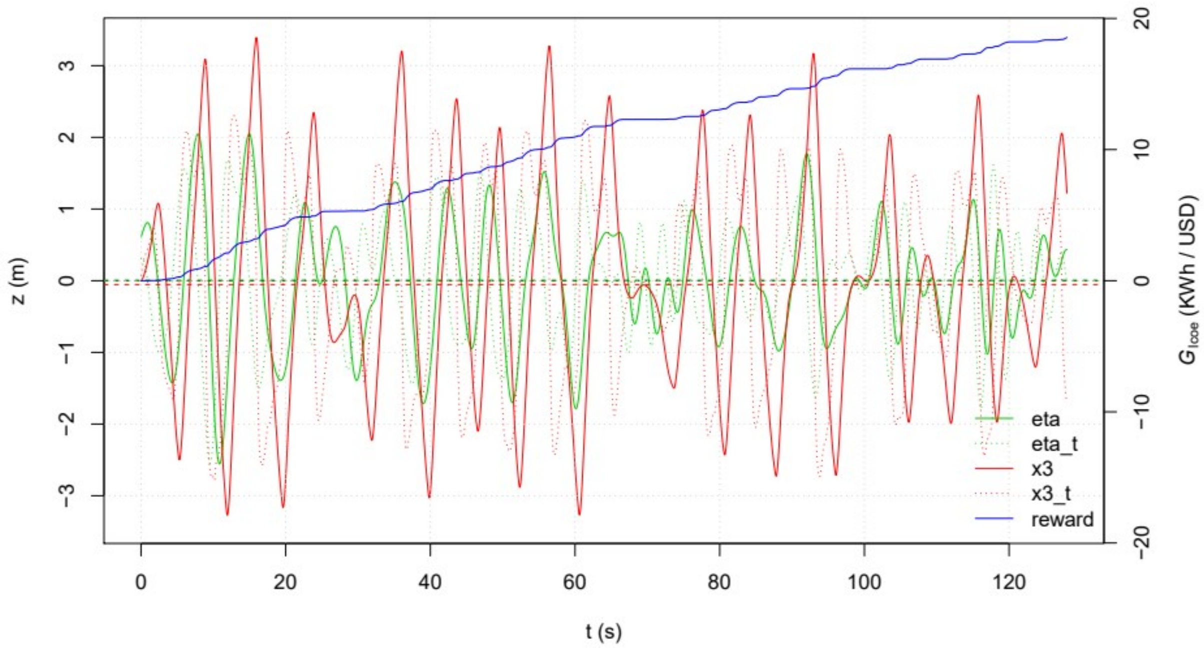

Figure 9: A threshold control force of 150 kN: example time-series plots for a single episode run using a control policy that optimised G_{lcoe}.

Next, we used a typical S-N fatigue curve to relate the probability of device unavailability to the ratio of the control force to a design maximum for the control force. Figure 8 shows a policy assuming a design control force of 500 kN and Figure 9 shows a policy assuming a design control force of 150 kN. The different levels of design control force would have had an impact on CAPEX. However, this has not been included in the reward here. The parameters of interest when comparing these two plots are displacement and velocity: it is clear that reducing the design force results in a policy with smaller motions.

The aim of this study was to produce custom rewards that were proxies for LCOE. It demonstrated one approach to transforming the complexity and long timescales of the LCOE calculation to a practical reward that could be used with RL.

4.5 Learning with Delay and Noise

The WES Control landscaping document identified that delay and noise were two common sources of problems when control algorithms which had been shown to be successful in simulations were subsequently tested on real devices.

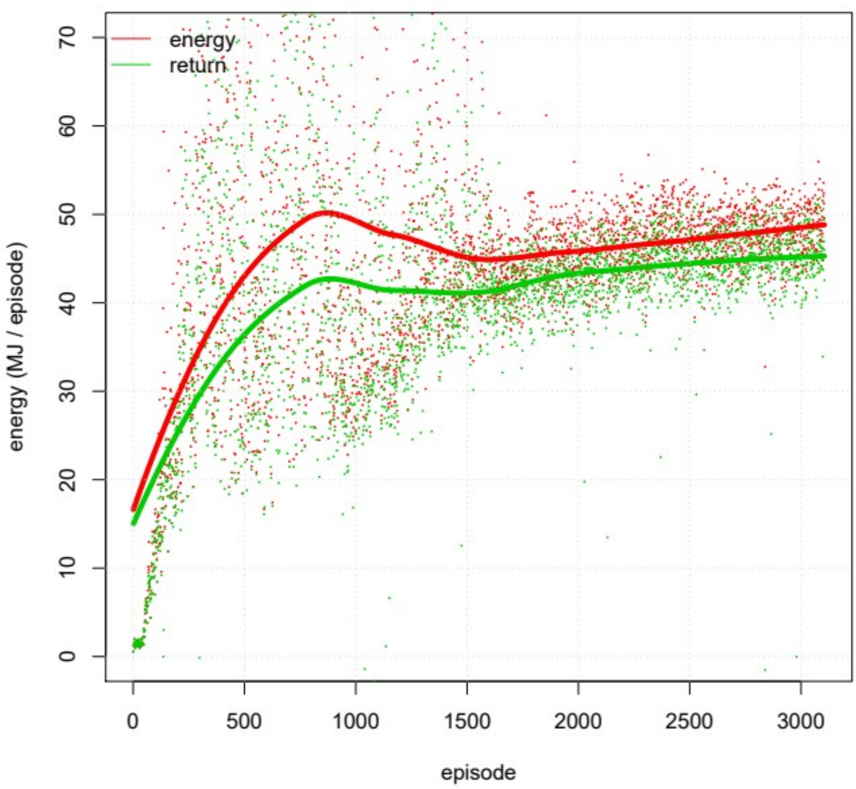

Figure 10: Convergence plot for RL training session with no noise and no time delay.

Figure 10 shows the convergence plot for a baseline case of no noise and no delay, and Figure 11 shows a typical plot for a typical policy.

In Stage 2, a policy was learnt on a simulation where noise had been added to the control inputs (*states*) and actuator outputs (*actions*). Another policy was trained on a simulation with delays added to the states and actions. Values of noise and delay were chosen that were typical for real WECs. Figure 12 shows the convergence plot for a policy where both noise and delays were added to the states and actions. Figure 13 shows a typical time-series plot for the converged policy.

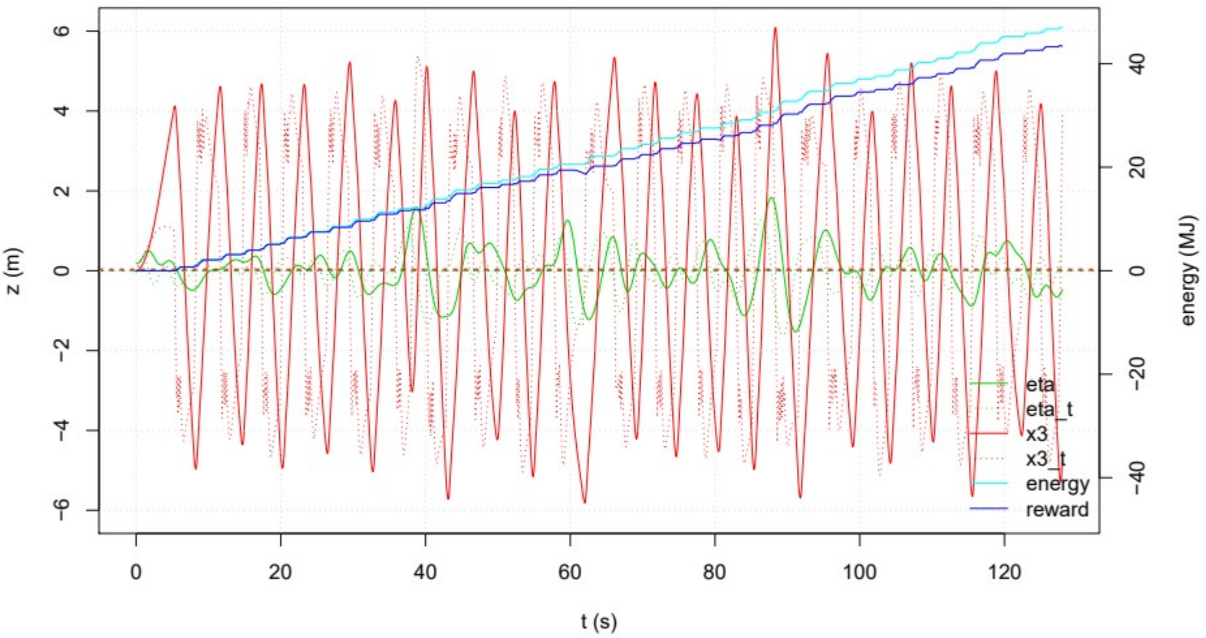

Figure 11: Baseline case: successful learning with no noise and no delay. Example time-series plot showing the wave elevation (eta), the heave displacement (x3), the heave velocity (x3_t), the cumulative energy absorbed and the cumulative reward. Despite being in the legend, eta_t is not plotted.

4.6 Policy Transfer

In conventional control, noise and delay are not a problem if the amount of noise or delay are known and included in the models. Difficulties arise when the values of noise and delay are not known. We wanted to explore how policies transferred between models with different levels of noise and delay. Figure 14 shows how a policy that has been trained on a model without noise or delay (e.g. in the training session shown in Figure 10) performs on a simulation which includes noise and delay (such as the one used in the training session shown in Figure 12). Comparing Figure 14 with Figure 11, it can be seen that the energy capture is very similar, but the reward has dropped slightly.

Overall, the policy transfer is good. The transferred policy is not as good at avoiding the penalties for relative amplitude threshold exceedence, but the difference is so small that the difference in behaviour cannot be spotted by comparing displacement plots. While this was a relatively simple study, it nevertheless gives some confidence that policies trained on simulations have a good chance of operating in a robust manner on real devices.

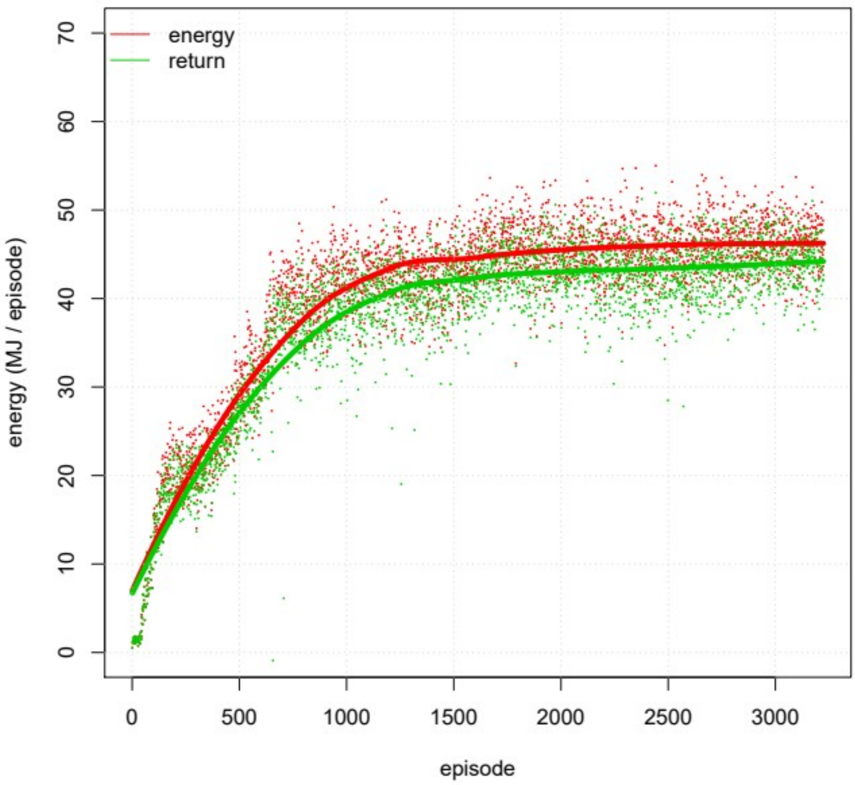

Figure 12: Convergence plot for RL training with both noise and time delay.

Further studies were conducted that trained policies on a spectrum of one particular characteristic period, and then tested these on simulations where the incident wave spectrum had a different characteristic period. The policy transfer was successful. This gives confidence that it should be possible to learn a policy that operates over a range of spectra, and in spectral shapes that are different from those used in training.

These transference studies were a first step in assessing the robustness of policies trained on simulations. They will inform future work in this area.

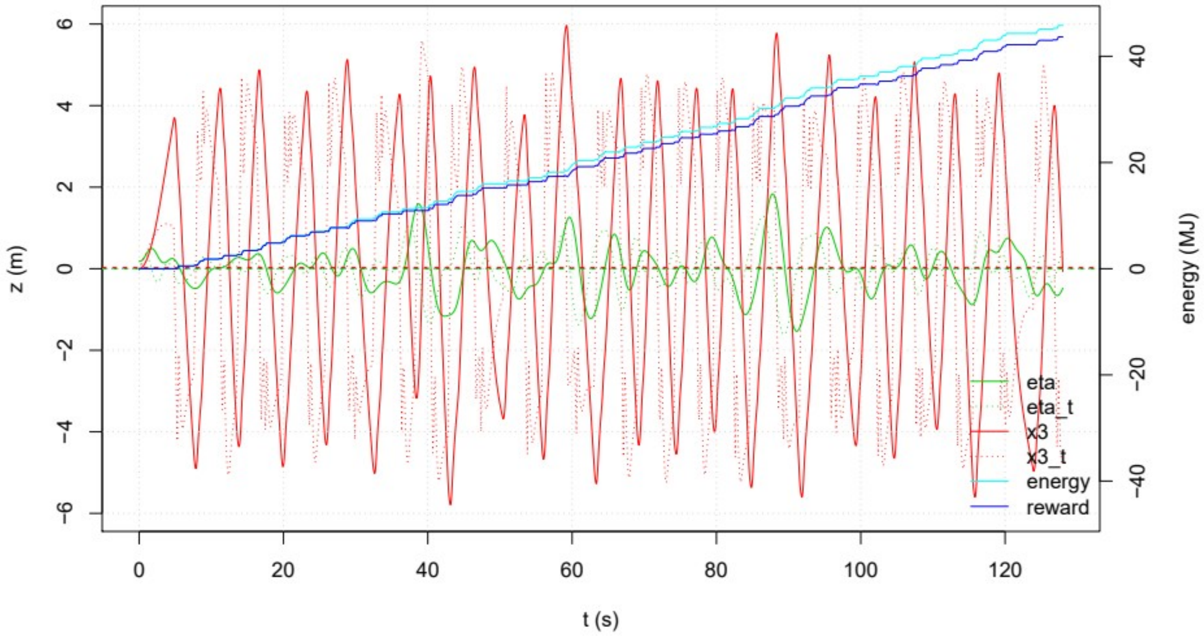

Figure 13: Successful learning with both noise and time delay. Example time-series plot showing the wave elevation (eta), the heave displacement (x3), the heave velocity (x3_t), the cumulative energy absorbed and the cumulative reward. Despite being in the legend, eta_t is not plotted.

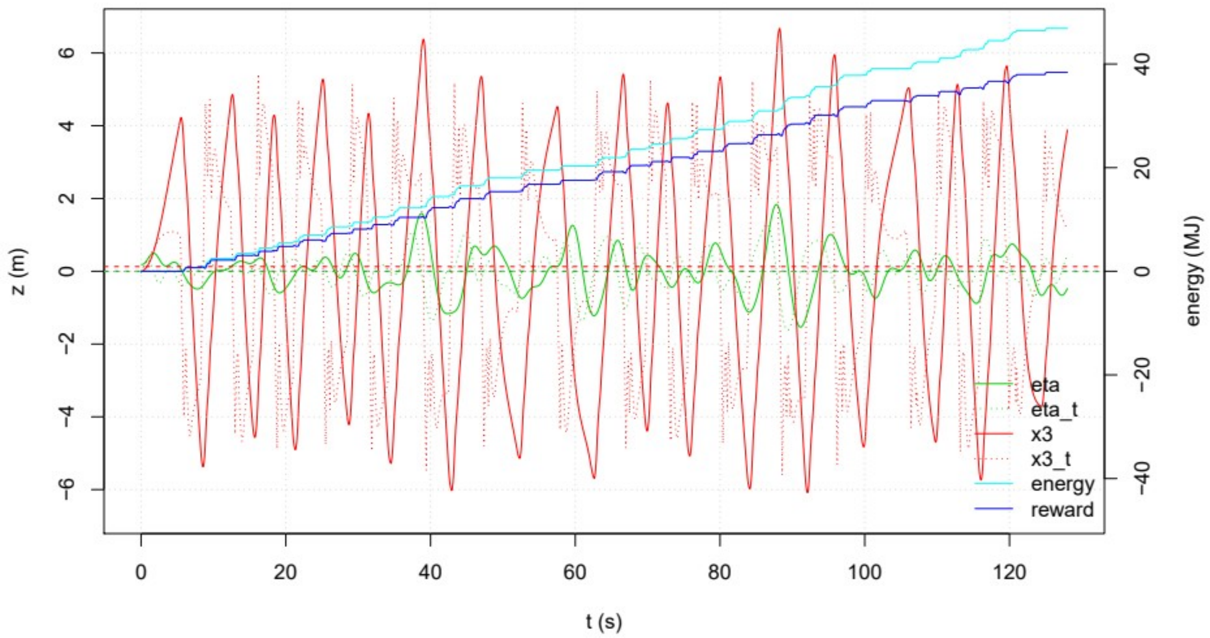

Figure 14: Example of successful policy transfer from a system trained with no noise and no delay to a system with noise and delay.

5 Recommendations for Further Work

The key steps for commercialisation consist of methodology development using simulations, hardware-in-the-loop (HIL) testing, implementation in tank tests, and implementation in sea trials. The numerical modelling (computer simulation) and HIL work is planned for Stage 3 of the WES control call. The objectives are as follows:

- Integrate RL training platforms with real-time systems in preparation for tank tests:
  - As tank tests are an expensive place to learn about software integration, there will be an intermediate HIL step.
  - A numerical model of the HIL system will be built and used as a training environment.
- Tailor numerical model to the physical model that will be tested in the tank (AquaHarmonic's 1/20th scale buoy).
  - Add sufficient complexity so that this numerical model can be used as an environment for learning a policy for use on the 1/20th scale buoy, e.g. improve models of the PTO and sensors.
  - The HIL system tests the physical systems that are actuated by the control system. Thus HIL tests will inform the level of complexity required in modelling the PTO.
- Advance policy transfer work by demonstrating that policies learnt on numerical models work on the HIL system.
- Improve policies by tailoring the training algorithms and hyper-parameters to this specific problem.

Stage 3 of the WES control programme will culminate in tank tests. The purpose of the tank tests is to:

- Provide evidence that RL can improve the control of a real WEC with a representative PTO system, in representative wave conditions.
- Advance policy transfer work by demonstrating what level of complexity is required for a training environment. This is a key step, as the states (control inputs) are influenced by body hydrodynamics, so cannot be tested with the HIL system.
- Inform planning for prototype field trials, and derisk these trials.

After the WES Control programme, a prototype will be tested in a field trial. Our preferred partners are AquaHarmonics, who have been collaborators since Stage 1. They are preparing for 1/7th scale sea trials at a test site in Hawaii.

6 Communications and Publicity Activity

The CEORL project presented a poster and a short presentation at the WES annual conference. Other than public communications via WES, there has been no other publicity activity.