applied sciences

MDPI

Article

Extending Power Electronic Converter Lifetime in Marine Hydrokinetic Turbines with Reinforcement Learning

Samuel Barton, Ted K. A. Brekken * and Yue Cao

School of Electrical Engineering & Computer Science, Oregon State University, Corvallis, OR 97331, USA; bartonsa@oregonstate.edu (S.B.); yue.cao@oregonstate.edu (Y.C.)
* Correspondence: brekken@eecs.oregonstate.edu; Tel.: +1-541-737-2995

Abstract: Hydrokinetic turbines (HKTs) are a promising renewable energy source due to the consistency and high energy density in river and tidal resources. One of the primary barriers to the widespread adoption of HKT technologies is a high levelized cost of energy (LCOE). Considering the marine operating environment, the operation and maintenance costs are substantial. The power electronic converter, a key element in the electrical energy conversion system, is a common point of failure in direct-drive turbine applications—leading to increased maintenance efforts. This work presents a reinforcement learning (RL) method built within a quadratic feedback torque control framework to balance energy generation with power electronic device lifetime. The effectiveness of the RL-based control scheme is compared against a static baseline controller through two year-long tidal case studies. The results showed that the proposed method reduced cumulative damage on the device by upwards of 75% but reduced energy generation by up to 25.2%. Using a custom real-time cost estimation function that considers the sale of energy and an estimate of the costs associated with operating a device at a given temperature, it was found that the RL method can increase net income by up to 45.4% depending on the energy market conditions.

Keywords: reinforcement learning; hydrokinetic turbine; power electronics; lifetime; marine energy

Academic Editor: José A. Orosa

Received: 18 December 2024
Revised: 21 February 2025
Accepted: 22 February 2025
Published: 26 February 2025

Citation: Barton, S.; Brekken, T.K.A.; Cao, Y. Extending Power Electronic Converter Lifetime in Marine Hydrokinetic Turbines with Reinforcement Learning. *Appl. Sci.* **2025**, *15*, 2512. https://doi.org/10.3390/app15052512

1. Introduction

A 2021 study conducted by the US National Renewable Energy Laboratory found that the marine renewable energy sources in the United States have the capacity to meet up to 57% of the nation's energy needs [1]. Alaska alone accounts for approximately 40% of the nation's river and ocean wave resources and up to 90% of the national tidal resources [2]. Despite the abundant energy resources found in our oceans and streams, these resources remain largely untapped.

Of the many marine renewable energy technologies being developed, hydrokinetic turbines (HKTs) have been receiving significant attention in the literature. HKTs operate similarly to wind turbines but are located in rivers, tidal inlets, and ocean currents to harness the energy from these resources. Compared to their wind counterparts, HKTs have the added benefit of increased resource predictability, reduced variability, and higher energy density considering the relative density of water to air. However, a high levelized cost of energy (LCOE) is preventing the widespread adoption of HKT technologies.

LCOE is loosely defined as the net revenue associated with generating one kilowatt-hour (kWh) of energy, including manufacturing, operations, and other costs. LCOE reductions can be achieved in a number of ways, including turbine design optimization [3]

and improved maintenance strategies [4]. In wind turbine systems, the power electronic converter has been found to have a high failure rate, especially in direct-drive technologies [5,6], directly adding to the maintenance costs of these systems. Thermal cycling at the junction of the device has been identified as a primary cause of device failure [7]. Therefore, a key way to reduce the maintenance costs of HKTs is to reduce the severity of thermal cycling on power electronic devices.

There is a large body of research devoted to reducing thermal stress on power electronic devices, termed active thermal control (ATC) [8,9]. ATC methods have been applied in multiple areas of research, including wind turbine applications. These approaches are generally realized at either the device level, converter level, or system level (including the cooling system) [9,10]. At the converter level, these efforts to reduce thermal stresses can be built within the high-level control scheme of the system, thereby improving the lifetime of the converter while incurring little to no additional costs. The thermal response of power electronic devices is directly related to the amount of power harvested by the turbine as losses generally increase with power generation. Therefore, it is intuitive to integrate ATC within the control algorithm that is responsible for regulating turbine power, which is often some form of maximum power point tracking (MPPT) control [11,12].

When utilizing these ATC methods, it is often found that there is a direct trade-off between device lifetime and system performance [8,9,13,14]. Multiple works have been developed to solve this trade-off between device damage and efficiency using an economic-based approach to their ATC methods [10,15,16]. Although these works have presented solutions that simultaneously improve the revenue of the turbine system and minimize the damage to power electronic devices, the proposed methods are heavily reliant on models of the turbine. HKT systems have the additional challenge of being much more dependent on the natural environment. For example, biofouling (the growth of marine organisms on the HKT structure) can significantly impact the optimal operating conditions of the HKT by changing the hydrodynamics of the blades [17,18]. Therefore, the accuracy of the HKT models is not guaranteed.

Considering that the dynamics of the HKT system can vary over time, adaptive model-free control methods are increasingly attractive. Reinforcement learning (RL) is a model-free learning-based approach that develops its own understanding of the system by interacting with its environment. RL has shown promise in applications of condition monitoring, fault diagnosis, and remaining useful life estimation for power electronic converters [19] and has been used to maximize the energy generated by HKT systems in various applications [20–25]. However, few works have been published that utilize RL to extend the lifetime of power electronic converters. References [26,27] present a neural-network-based approach to predict the driving behaviors and device junction temperature for a given point-to-point trip in electric vehicle applications. The predicted temperature profile is then used alongside a rainflow counting algorithm that identifies the most harmful thermal cycles the devices may face over the drive cycle and creates a reference junction temperature at or near the peak value of the thermal cycle to reduce junction temperature fluctuation. This reference temperature is then used to control the cooling system employed in the vehicle, which requires a complex and highly capable cooling system that is not suitable for most HKT applications.

This work develops a model-free approach that reduces the thermal stresses on a device through a lightweight stochastic gradient descent (SGD) state–action–reward–state–action (SARSA) algorithm with a Gaussian radial basis function (RBF)-based approximation method, similar to the method presented in [25]. The proposed RL method in this work is built within a common k-omega-squared (or optimal torque controller [11,12]), thereby removing the need for a complex actively controlled thermal management system. Also, this work utilizes a similar economic-based reward function to that developed in [15] to

train the RL agent. The work presented in [25] was primarily focused on maximizing energy generation with the consideration of environmental impacts on HKT performance through a linear feedback torque control law. Although the proposed RL method is similar, the state space in this work is significantly more complex (considering not only the flow velocity but also the water temperature, energy cost, and current feedback gain k), leading to additional challenges in designing the RL method and thoroughly training the RL agent.

The remainder of the paper is structured as follows. Section 2 covers the modeling methods employed to represent the turbine, permanent magnet synchronous generator (PMSG), and power electronic converter. Section 3 presents an overview of the general RL framework and the specific details of the proposed SGD SARSA method. Section 4 explains how the RL framework has been adapted for this specific application (including how the monetary-based reward function has been defined), the method used to train the RL agent, and presents a case study comparing the proposed RL-based control method to a baseline controller in two different environmental settings. Section 5 presents a discussion of the performance of the proposed method, followed by the implications of the assumptions made throughout this paper and a statement regarding future work. Finally, the manuscript is concluded in Section 6.

2. System Modeling

This section covers the modeling of the three key subsystems that comprise the HKT system: the turbine, the PMSG, and the power electronic converter, as shown in Figure 1.

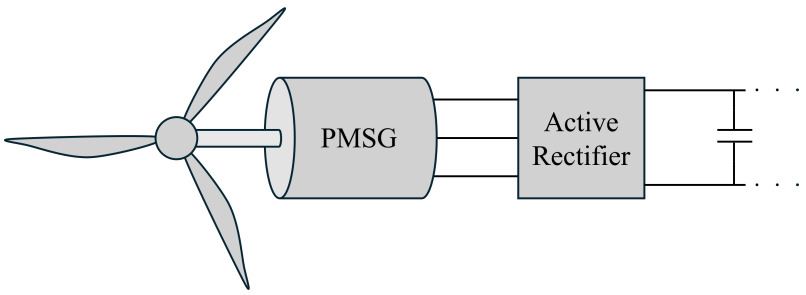

Figure 1. Hydrokinetic turbine system overview highlighting the turbine, PMSG, and generator-side converter subsystems.

2.1. Modeling of Turbine System

For a flow stream of a given fluid flowing through an area A_c with density ρ and flow velocity u_0, the amount of power available within that area P_{flow} can be calculated using (1).

$$P_{flow} = \frac{1}{2}\rho A_c u_0^3 \tag{1}$$

The amount of power absorbed by the turbine is linearly related to the total power in the incoming flow stream through the functional area of the HKT through a term named coefficient of power C_p, which is a function of the tip speed ratio λ (defined as $\lambda = \omega L/u_0$, where ω is the rotational speed of the turbine and L is the length of the turbine blades). Therefore, in terms of the HKT system, if it is assumed that the flow passes through the rotor swept area of the turbine A_c and the turbine is operating at a given tip speed ratio λ, the power harvested by the turbine $P_{turbine}$ can be estimated using (2).

$$P_{turbine} = \frac{1}{2}\rho A_c C_p(\lambda) u_0^3 \tag{2}$$

The equation of motion that describes the rate of the rotational speed is shown in (3). This formula relates the net torque T_{net} (which is defined as the difference between the

turbine torque $T_{turbine}$, generator torque T_{gen}, and friction torque T_{fric}) to the rate of change in ω through the total moment inertia of the HKT system J (the sum of the moment of inertia of the turbine $J_{turbine}$ and the generator J_{gen}: $J = J_{turbine} + J_{gen}$).

$$T_{net} = T_{turbine} - T_{gen} - T_{fric} = J\frac{d\omega}{dt} \tag{3}$$

The torques $T_{turbine}$ and T_{fric} are related to ω, as shown in (4), where C_{fric} is the coefficient of friction. T_{gen}, on the other hand, is defined as a function of the current in the stator windings, as shown in Section 2.2, and is controlled through the proposed k-omega-squared control method.

$$T_{turbine} = \frac{P_{turbine}}{\omega}, \qquad T_{fric} = C_{fric} \cdot \omega \tag{4}$$

To improve reliability and reduce system costs and complexity, it has been assumed that there is no gearing between the turbine and generator shafts, and there is no control over the pitch of the turbine blades.

Region-Based Control

The turbine modeled in this research has four regions of operation, which are defined by the flow velocity, as shown in Figure 2.

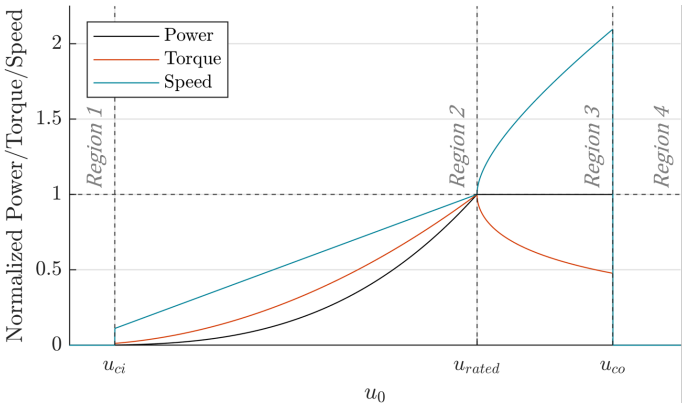

Figure 2. Plots of normalized turbine power ($P_{turbine}$), torque ($T_{turbine}$), and rotational speed (ω) compared to flow velocity (u_0), highlighting the operation of the turbine in each of the four regions of operation.

The turbine operates in "offline mode" for flow velocities below the cut-in speed u_{ci} (Region 1) and above the cut-out speed u_{co} (Region 4). The turbine is blocked from rotating in these two operating regions using an external brake, which is modeled by increasing C_{fric} in this work.

The turbine operates in the maximum power extraction region for flow velocities between u_{ci} and the rated turbine flow velocity u_{rated} (Region 2). In Region 2, the goal is generally to operate at the optimal tip speed ratio λ^* such that the C_p of the turbine is at its maximum value, often termed maximum power point tracking control.

Operation in Region 3, on the other hand, is designed to ensure the turbine operates at rated power for flow velocities between u_{rated} and u_{co}. Depending on the design of the turbine and its application environment, the harvested power can be regulated using pitch or speed control techniques [28]. As stated above, it is assumed that the turbine system in this work does not have pitch actuation. Therefore, speed control techniques must be used in Region 3 operation.

Considering the equation for turbine power (2) and the general shape of the $C_p(\lambda)$ curve, power curtailment can be achieved by operating on the left or right side of the optimal point on the $C_p(\lambda)$ curve. This entails either decreasing or increasing λ, respectively. These two methods both have their pros and cons. Operating at lower λ requires exceptionally high torques from the generator, which would often require a gearbox to reduce the torque load on the generator. On the other hand, operating at increased values of λ reduces the overall torque load. However, rapid increases in u_0 will lead the turbine to operate at values of λ closer to the optimal point, increasing the power generated by the HKT. Since the turbine system is assumed to be direct drive, the latter option will be used for Region 3 control. This will be accomplished through our existing k-omega-squared control law by decreasing the values for k as flow velocities increase, which can be determined using a lookup table approach. A significant assumption that leads to the adoption of this form of Region 3 control is that there are no limitations on the rotational speed of the turbine, and the primary limitation of the system is the PMSG torque. If this assumption fails, torque must be increased in Region 3, increasing per-phase currents and, further, device losses, and the proposed control method may not be suitable.

It should be noted that the proposed RL-based control scheme is only used in Regions 2 and 3 as operations in Regions 1 and 4 are designed for economic or safety reasons.

2.2. Modeling of Permanent Magnet Synchronous Generator

The PMSG has two primary purposes: to convert mechanical energy to electrical energy usable for the grid and to apply torques to the turbine system that oppose $T_{turbine}$ to control the speed of rotation.

For PMSGs, it is common to utilize a rotating reference frame to model and control the machine. Following the power-invariant d/q transformation presented in [29], electromagnetic torque T_{gen} can be calculated as shown in (5), where p is the number of poles, λ_r is the rotor flux linkage, L_d/L_q are the d-/q-axis stator inductances, and i_d/i_q are the d-/q-axis stator currents. In this work, it is assumed that there are no losses in the PMSG, such that T_{gen} is equal to the torque applied on the shaft of the PMSG rotor.

$$T_{gen} = \frac{p}{2}(\lambda_r - (L_d - L_q)i_d)i_q \tag{5}$$

Assuming the voltage at the DC bus of the power electronic converter is sufficiently high such that the d/q voltages do not exceed the achievable per-phase voltage at the stator terminals, no field weakening is required, and i_d can be set to 0 [29]. In this case, (5) can be simplified further, resulting in the following relationship between q-axis current and torque:

$$T_{gen} = \frac{p}{2}\lambda_r i_q \tag{6}$$

Following the power-invariant d/q transformation, the root mean square (RMS) per-phase current I_{ph} can be calculated from the q-axis current as follows:

$$I_{ph} = \frac{\sqrt{i_d^2 + i_q^2}}{\sqrt{3}} = \frac{i_q}{\sqrt{3}} \tag{7}$$

2.3. Modeling of Power Electronic Converter

For this work, it has been assumed that the active rectifier is of a two-level voltage source converter topology, as shown in Figure 3. For a metal-oxide-semiconductor field-effect transistor (MOSFET), the losses of a given device can be estimated as the sum of the switching and conduction losses of both the MOSFET and internal body diode.

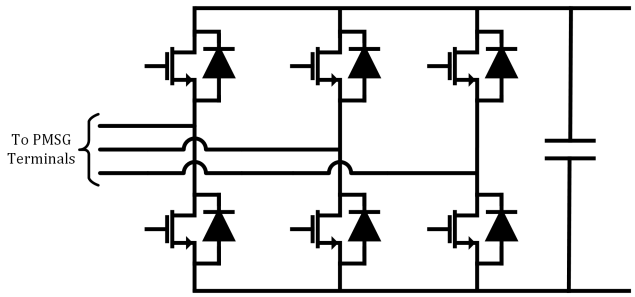

Figure 3. Schematic of a two-level voltage source converter.

In most MOSFET datasheets, the manufacturer provides an estimate of total switching energy as a function of device drain current, which can be compiled into a lookup table to reduce the complexity of the model. Using this provided information, the switching losses of the MOSFET $P_{sw,m}$ can be estimated using (8), where f_{sw} is the switching frequency, $E_{tot}(\langle I_{DS} \rangle)$ is the total switching energy at a given average drain current $\langle I_{DS} \rangle$, and k_v is a linear scaling term used to estimate how the operational drain-source voltage of the device affects the overall switching energy value provided by the manufacturer at some test voltage $V_{DS,test}$ [30]. As a worst case, it can be assumed that the drain-source voltage that the device will experience in the two-level topology is equal to the DC bus voltage V_{DC}.

$$P_{sw,m} = f_{sw} \cdot k_v \cdot E_{tot}(\langle I_{DS} \rangle), \qquad k_v = \frac{V_{DC}}{V_{DS,test}} \tag{8}$$

Assuming (i) there is no dead-time when neither of the MOSFETs that comprise a phase leg of the converter are engaged, (ii) the majority of the current flows through the channel of the MOSFET rather than its internal body diode, and (iii) a sinusoidal pulse width modulation scheme is used to generate the high-frequency switching signals that engage the MOSFETs in the power electronic converter $\langle I_{DS} \rangle$ and the RMS drain current (I_{DS}) of the MOSFET can be estimated from the RMS phase current and the current modulation index value m_a as follows, where $cos(\phi)$ relates to the power factor of the PMSG, which can be assumed to be 1 in most PMSG applications, as shown in (9).

$$\langle I_{DS} \rangle = m_a \cdot \frac{\sqrt{2} I_{ph}}{4} \cdot cos(\phi), \qquad I_{DS} = \frac{I_{ph}}{\sqrt{2}} \tag{9}$$

MOSFET conduction losses $P_{cd,m}$ are equal to the I^2R losses caused by the on-resistance of the MOSFET $R_{DS,on}$, as shown in (10).

$$P_{cd,m} = I_{DS}^2 \cdot R_{DS,on} \tag{10}$$

The switching losses of the body diode $P_{sw,d}$ are largely related to the diode's reverse-recovery time, which occurs when the MOSFET switches off. The charge accumulated during this time is denoted as Q_{rr} and is often provided in the datasheet for the device. From this, the diode switching losses can be estimated as presented in [31] and shown in (11).

$$P_{sw,d} = \frac{1}{4} \cdot f_{sw} \cdot Q_{rr} \cdot V_{DC} \tag{11}$$

Similar to the MOSFET, body diode conduction losses are related to the I^2R losses of the diode. However, under the assumption that the majority of the current flows directly through the MOSFET rather than the internal body diode, the conduction losses of the body diode can be assumed to be zero.

Overall, the total losses of a single device P_{dev} can be estimated using (12) below.

$$P_{dev} = P_{sw,m} + P_{cd,m} + P_{sw,d} \tag{12}$$

2.4. Thermal Modeling of the Power Electronic Converter and Device Lifetime Estimation

2.4.1. Thermal Modeling

Of the many methods to model the thermal response of a power electronic device under varying load conditions, equivalent circuit models such as the Foster and Cauer thermal networks are some of the simplest and most commonly found in the literature. Traditionally, thermal resistance θ and thermal capacitance C values are required for each medium between the junction of the device and ambient environment. In this work, the thermal chain is composed of the power electronic devices, thermal interface materials (TIM) that bond the device to the heat sink, and a large heat sink that interacts with the incoming flow. However, it is assumed that the time-varying aspects of the flow and environmental temperature are sufficiently slow such that the thermal capacitance present in the thermal chain of the MOSFET can be ignored and only the thermal capacitance of the heat sink is considered.

Considering that the heat sink directly interacts with the flow of the water in this system, the thermal resistance of the heat sink would have some dependence on u_0 through convective heat transfer. However, the changes in thermal resistance relative to the total equivalent thermal resistance are negligible and have been assumed to be constant in this work. It should also be noted that the thermal parameters and losses have been assumed to be equal for each device in the converter.

The simplified version of the Cauer network can be observed in Figure 4, where P_{dev} is defined as the per-device losses calculated using (12), θ_{jc} represents the device junction-case thermal resistance, θ_{TIM} represents the thermal resistance of the TIM, C_{HS} represents the thermal capacitance of the heat sink common to each device, θ_{HS} represents the thermal resistance of the heat sink, T_{amb} represents the temperature of the ambient temperature of the environment, and T_j represents the junction temperature of the device.

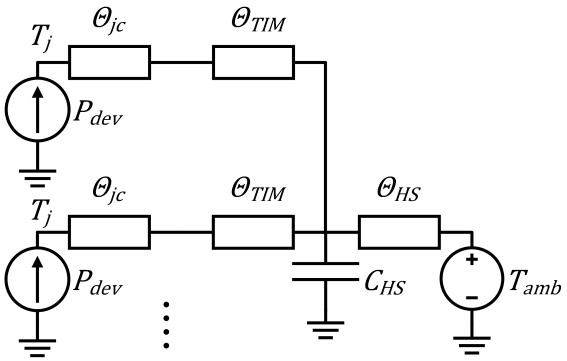

Figure 4. The simplified Cauer thermal network used to estimate junction temperature for each device. This figure only represents two of the six devices that comprise the two-level voltage source converter. The dots are representative of the other four devices feeding into the same node.

2.4.2. Life Consumption Estimation

For silicon carbide (SiC) MOSFETs, cycles in T_j are one of the primary factors leading to device failure for failure modes such as bond wire liftoff, solder fatigue, and gate oxide breakdown [32,33]. Analysis of the impacts the temperature profile has on the lifetime of the device is typically conducted using physics-of-failure (PoF) models, data-driven models, or some combination of the two. In general, PoF models are developed for a specific failure mode, while data-driven models may consider a variety of failure modes

depending on the dataset. In this work, a common PoF method, the Coffin–Manson model, is used to estimate the number of cycles to failure N_f for a specific device operating at a given thermal cycle amplitude ΔT_j and minimum thermal cycle temperature T_{min}, as shown in (13) [34]. The Coffin–Manson model is often used to estimate device failure caused by bond wire liftoff and solder fatigue [33]. The scaling terms in (13) are defined for IGBT technologies rather than SiC MOSFETs as there is a lack of publicly available data that provides sufficient information to derive the scaling terms in (13) for SiC technologies. Unfortunately, this does lead to general inaccuracy in the estimation of N_f. The primary issue with utilizing scaling terms from other device technologies is, as the Coffin–Manson model looks at failures within and between the materials that make up the device, the scaling terms essentially represent the material properties of the specific device under test. Considering the differences between IGBTs and SiC MOSFETs, the material properties of SiC technologies are likely not encapsulated in the definition of the scaling terms in (13). Although the absolute values calculated for N_f are not determined for the specific device technology assumed in this work, these estimates provide a meaningful representation of how thermal stresses change under the proposed control method.

$$N_f = 1.017^{(20-T_{min})^{1.16}} \cdot 1.26 \cdot 10^{13} \cdot \Delta T_j^{-4.51} \tag{13}$$

Life consumption (LC) is a metric that assesses the cumulative damage a device has accumulated over a mission profile. The life consumption LC of the device during the mission profile can be estimated using the Palmgren–Miner rule [35], which is shown in (14). The Palmgren–Miner rule calculates the cumulative damage on a device considering the number of cycles $n_{c,i}$ experienced at a specific stress level defined by a thermal cycle amplitude $\Delta T_{j,i}$, minimum cycle temperature $T_{min,i}$, and a corresponding estimated number of cycles to failure $N_{f,i}$ calculated for the corresponding stress metrics. As the estimate for LC approaches 1, the device is expected to fail. The lifetime of the converter can be approximated as the reciprocal of LC, assuming the converter repetitively experiences the same profile throughout its deployed life.

$$LC = \sum_i \frac{n_{c,i}}{N_{f,i}} \tag{14}$$

Pertinent thermal cycle data must be extracted from the mission profile to be used in (14). This can be achieved using various cycle counting methods, with rainflow counting being one of the most popular. MATLAB has a built-in rainflow counting function that returns values for $n_{c,i}$, $\Delta T_{j,i}$, and mean cycle temperature (which can be used to calculate the minimum cycle temperature) that will be used to calculate the cumulative LC over the entire mission profile [36]. More information on the rainflow counting algorithm can be found in [36].

3. Reinforcement Learning Method

This section covers the basics of the RL framework and the specific RL algorithm used in this work.

3.1. Reinforcement Learning Framework Overview

The RL framework can be defined by two discrete entities: the agent and the environment. The environment can be thought of as a means of responding to the action A taken by the agent by transitioning from the current state S to the next S' and providing the agent some reward R. In general, the purpose of the agent is to choose A to maximize R. These actions are taken following some policy π, which is the core of the RL agent.

The RL agent improves the quality of π as it learns more about the dynamics of the environment. These policy improvements can be conducted in several ways, giving rise to a number of RL methods, such as temporal difference (TD)-learning, actor–critic methods, and deep-deterministic policy gradient methods, to name a few. Please refer to [37] for more information. A general block diagram portraying the RL framework is shown in Figure 5.

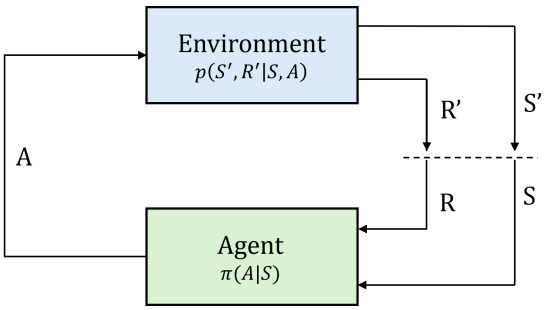

Figure 5. A block diagram representing the agent–environment interaction defined by the RL framework, derived from Sutton and Barto [37].

3.2. SGD SARSA with Linear Function Approximation

The SARSA algorithm has been chosen for this application due to its simplicity and stability [37]. The SARSA algorithm is an on-policy TD-learning method, with Q-learning being its off-policy counterpart. The terminology "on-policy" or "off-policy" refers to whether the updates occur with respect to the current policy. Q-learning is one of the most popular learning algorithms found in literature. However, it is difficult to integrate off-policy learning methods with linear function approximations as they are susceptible to diverge [37]. Hence, SARSA has been chosen for this work due to the seamless integration of the function approximation and learning methods.

The basics of the SARSA algorithm are to estimate the expected value of selecting an action A from the current state S following the policy π. This estimate is called the Q-value and is denoted $Q(S, A)$. In the tabular approach, a table (called the "Q-table") is constructed, composed of a Q-value for each state–action pair in the defined state–action space. The basic goal of the RL agent is to know the exact Q-value for each element in the Q-table to devise the best policy π. Therefore, learning is facilitated through updating. In general, updates occur in steps in the direction of the update target. The size of these steps is determined by the step-size parameter or "learning rate" α. The update target under the SARSA methodology is the difference between the new estimate of the state–action pair following policy π and the old estimate of the state–action pair. The new estimate considers the achieved reward R and the value of the future state–action pair following the current policy $Q(S', A')$. The consideration of the future Q-value is termed "bootstrapping", and the amount that this value impacts the update target is controlled through the discount factor γ. The basic SARSA update equation is shown in (15) below

$$Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma \cdot Q(S', A') - Q(S, A)\right] \tag{15}$$

The tabular approach works well for small state and action spaces. However, as the problem becomes more complex, the memory required for the Q-table increases significantly. An alternative to the tabular approach is to use function approximation to represent the Q-table such that the Q-values for each state–action pair can be approximated by some function. The function approximation approach also solves another common issue in RL: balancing exploiting known "good" actions and exploring other "better" ones. In tabular methods, for the agent to be certain it is selecting the most valuable action, all actions should

be explored for each state. With function approximation, the values of one state–action pair can be generalized to its neighbors, reducing the burden for an extensive exploration stage and decreasing training time.

There are a number of function approximation methods, such as linear approaches and neural networks (a non-linear approach), both of which have their respective advantages and disadvantages [37]. Linear methods are selected in this work to reduce the memory necessary to approximate the Q-values accurately. In linear function approximation methods, the Q-values are represented as the product of a weight vector $\mathbf{W}$ and a feature vector $\mathbf{X}(S, A)$, as shown in (16).

$$Q(S, A) = \mathbf{W}^\top \mathbf{X}(S, A) \tag{16}$$

The different classes of linear approximation methods stem from how the feature vector is defined. The fundamental types of features are polynomial functions, Fourier basis functions, coarse coding, tile coding, and radial basis functions. Of these, tile coding (a form of coarse coding) is the most computationally efficient [37]. In these methods, an area of varying shape and size is defined in the state–action space, representing each feature in the feature vector. If a given state–action pair is within this area, the corresponding element in the feature vector has the value of one and is a zero otherwise. The Gaussian RBF method is quite similar to the tile/coarse coding methods, but the features are defined as bell curves with a center point in the state space and some width σ, as shown in (17), where $[SA]$ is a vector of scalars holding the current state and action.

$$\mathbf{X}(S, A) = e^{-\frac{\|[SA]-\mathbf{C}\|^2}{2\sigma^2}} \tag{17}$$

Using the Gaussian RBF method, the feature vector defines how close a given state–action pair is to the center point of each feature rather than the feature vector describing whether or not a given state–action pair is located within the area of each feature. Therefore, graphically speaking, the approximated Q-function representing the state–action space is continuous and smooth. For a sparsely defined feature space (i.e., few features or center points), these continuous functions may improve the accuracy of the estimation of the Q-value as defined by (16). Therefore, the Gaussian RBF method has been selected in this work.

With the use of a function approximator rather than updating the Q-values directly, the SARSA update equation shown in (15) is refocused on updating $\mathbf{W}$. If combined with an SGD method, the weight update occurs in the direction that reduces the error for the current sample [37], and the weight update can be written as follows

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha\left[R + \gamma Q(S', A', \mathbf{W}) - Q(S, A, \mathbf{W})\right] \cdot \nabla Q(S, A, \mathbf{W}) \tag{18}$$

Considering the linearity of the RBF function approximation method, the gradient on the right-hand side of (18) can be simplified to the value of $\mathbf{X}(S, A)$. Hence, the final weight update equation takes the following form:

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha\left[R + \gamma Q(S', A', \mathbf{W}) - Q(S, A, \mathbf{W})\right] \cdot \mathbf{X}(S, A) \tag{19}$$

4. Case Study

This section describes how the proposed SGD SARSA algorithm and Gaussian RBF approximation method have been integrated within the high-level control strategy of the HKT, followed by a definition of the system parameters employed for the example HKT system. Then, the goals and process of training the RL agent are covered, with the analysis of the success of the training focusing on the cumulative reward achieved per

training epoch. Finally, the trained RL agent is compared against a baseline controller in two different year-long mission profiles, with the results highlighting the learned policy, the impact on the thermal response of the power electronics, and the difference in energy generation in each case.

4.1. SGD SARSA in the HKT System

A common method of controlling the turbine system is through a quadratic feedback control law called k-omega-squared torque control [38]. If k is tuned properly, the HKT will track the optimal tip speed ratio λ^* for all flow velocities below u_{rated}. The goal of the proposed RL method in the HKT system is to find a value or profile of values for k to ensure ample energy generation while reducing LC over the deployment. A high-level block diagram showing how the proposed RL-based control scheme is integrated with the k-omega-squared framework is shown in Figure 6, and a brief overview of the function of the proposed control system is as follows. At each time step, the RL agent samples the current state $[k, u_0, T_{amb}, c_e]$ and the reward for taking the previous action R. From these sampled values, the RL agent updates the weight vector $\mathbf{W}$ following (19) and selects a new action following the current policy π. These actions result in a change in the feedback gain k. The new value of k changes the generator torque command T_{gen} following the k-omega-squared framework and is realized by changing the q-axis current i_q following (6). From the new values of T_{gen} and i_q, the operating conditions of the turbine and the power electronics are changed following the methods outlined in Section 2. The reward function block then samples the generator power P_{gen}, the junction temperature of the power electronic device T_j, and the cost of energy c_e to provide the RL agent with a meaningful representation of the quality of the last action selected following the current policy π.

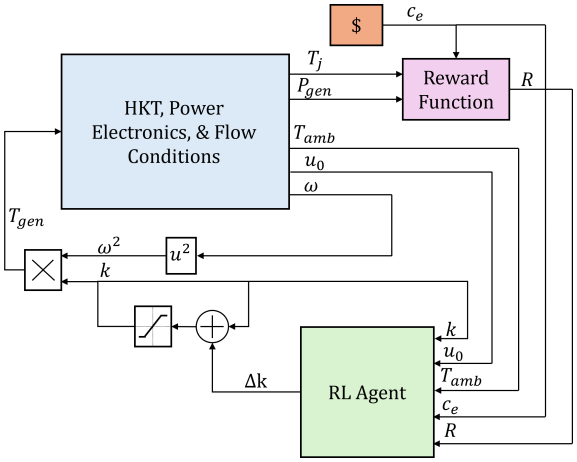

Figure 6. A block diagram highlighting the interaction between the RL agent and the HKT system. Descriptions of the environment block (blue), reward function block (pink), and RL agent block (green) can be found in Sections 2 and 3.2.

In terms of the RL framework, the RL agent considers four states: the current value of k, u_0, T_{amb}, and the current cost of energy c_e. Each state is then normalized by some base value to represent the values of each state between 0 and 1. These base values are represented as k_b, u_b, T_b, and c_b, respectively. The states have been normalized such that the value of σ in (17) was suitable for all state/action dimensions, simplifying the design space.

The actions the RL agent can take are incremental changes to the current value of k, Δk. Although the state space is continuous, the action space is discretized and normalized between -0.5 and 0.5 using a normalization variable Δk_b. The agent takes actions following an ε-greedy policy, where random actions are taken with some probability ε, and actions that correspond to the highest action value (called "greedy" actions) are taken otherwise.

After each training epoch, the value of ε is reduced using a linear decay parameter ε_d as shown in (20). The ε-greedy policy was chosen over other policies due to its popularity, simplicity, and ability to effectively manage the exploration–exploitation problem that is common in RL.

$$\varepsilon \leftarrow \varepsilon \cdot \varepsilon_d \tag{20}$$

A common downfall of applying RL methods to continuous problems with multiple tasks is forgetfulness [39]. Forgetfulness relates to the phenomenon in which the learned parameters from one task are forgotten while the agent is being trained on a new task. Agents are especially prone to this issue when there are significant differences between rewards for each task. With the variability in environmental conditions between training years, the proposed RL method is highly susceptible to this unfortunate side effect.

To avoid this problem, this work considers a dynamic definition of α that is incrementally reduced as the agent is trained. The value of α begins at some initial value α_0 and is reduced by the number of training epochs n_e, as follows:

$$\alpha = \frac{\alpha_0}{n_e} \tag{21}$$

This ensures that, the more the RL agent learns, the amount in which the weights vary from task to task is reduced to prevent the agent from overwriting what has previously been learned.

As described above, the RBF function approximation method is defined by $\mathbf{C}$ and σ. For this work, n_c RBF center points are designated for each dimension of the state and action space. The center points in the state space are equally spaced values from 0 to 1. The possible actions the RL agent can take are defined directly by the center points of the RBF in the action dimension such that the possible actions the agent can take are defined as the set $k_b \cdot \{-0.5, -0.25, 0, 0.25, 0.5\}$.

One of the novelties of this work is the definition of the reward function R. Considering that the RL agent in this system should be maximizing energy generation while simultaneously minimizing the aging effects on the power electronic converter, this problem presents itself as a multi-objective reinforcement learning problem. It is typical for an RL agent to only handle problems with one reward. In these scenarios, it is common to use a linear combination of multiple reward functions, termed scalarization [40]. Scalarization often includes weighting of n reward functions depending on the application, where the values of weights and reward functions depend on the application, as shown in (22). However, it is quite challenging to determine meaningful weights for each reward signal.

$$R = w_1 \cdot r_1 + w_2 \cdot r_2 + ...w_n \cdot r_n \tag{22}$$

In this work, an intuitive approach to determining the weighting of the respective rewards is developed using a monetary-based reward function. This methodology is similar to that presented in [15]. However, this work considers an estimate of the costs associated with operating the power electronic converter at a given value of T_j, while other system-level costs are excluded (such as regular maintenance costs, environmental protection costs, and depreciation costs).

First, a term related to the revenue from energy generation r_g is defined in (23), where T_s is the sample time for the control algorithm and c_e is the cost of energy in USD/kWh. Many areas in the United States have energy prices that are dictated by a local energy market. Therefore, c_e becomes a time-varying function that depends on the amount of load and generation at a given time.

$$r_g = \frac{P_{gen} \cdot T_s}{3.6 \cdot 10^6} \cdot c_e \tag{23}$$

Next, the costs r_c associated with operating at some T_j are approximated using (24). This cost-based reward considers the cost of the power electronic converter c_c and the expected lifetime of the converter L_c, which are dependent on the components selected for the converter and the given application. This cost model utilizes the Arrhenius lifetime acceleration model $A_f(T_j)$ to estimate how the lifetime estimate is impacted when operating at temperatures other than the reference temperature T_{ref}, where T_j and T_{ref} are provided in Kelvin [41]. The definition of $A_f(T_j)$ can be found in (25), where E_a is the activation energy for a given failure mechanism and k_B is Boltzmann's Constant.

$$r_c = A_f(T_j) \cdot \frac{T_s}{L_c} \cdot c_c \tag{24}$$

$$A_f(T_j) = e^{\frac{E_a}{k_B} \cdot \left(\frac{1}{T_{ref}} - \frac{1}{T_j}\right)} \tag{25}$$

Lastly, a penalty term r_p is developed to deter the agent from applying Δk, which would result in negative values of T_{gen}, or from exceeding some pre-defined maximum feedback coefficient k_{max}. In (26), k is the feedback gain that is realized after the agent takes the action Δk, and the value of k_{max} is defined to be the base value k_b for the normalization of the state space. It should be noted that, although the penalty r_p acts to deter the RL agent from searching in an area where there is no solution, an additional saturation limit has also been added that prevents any damage to the turbine. In Region 3, the upper end of the saturation limit is dynamic depending on the calculated value of k, which ensures that the turbine power is equal to the rated power. However, no penalty is associated with the RL agent attempting to exceed this limit.

$$r_p = \begin{cases} |k/10|, & k < 0 \\ 0, & 0 \leq k \leq k_{max} \\ |k_{max} - k|/10, & k > k_{max} \end{cases} \tag{26}$$

Therefore, the cumulative reward function is defined as the sum of r_g, r_c, and r_p, as shown in (27).

$$R = r_g - r_c - r_p \tag{27}$$

4.2. System Parameters

The HKT system in this work is derived from the work presented by Bahaj et al., which is a three-bladed horizontal axis turbine with a rotor diameter of 0.88 m [42]. For this example HKT system, the rotor dimensions have been linearly scaled from the dimensions presented in [42] to reach 6 kW at a flow velocity of 2.25 m/s, resulting in a blade radius of 0.8538 m. The C_p versus λ curve derived from the results presented by Bahaj et al. is presented in Figure 7. A summary of all HKT parameters is shown in Table 1.

The PMSG for this case study was modeled off the custom-made PMSG used in [34]. This PMSG was designed specifically for direct-drive HKT applications operating at similar rated flow velocity conditions. A list of all the utilized PMSG parameters can be found in Table 2.

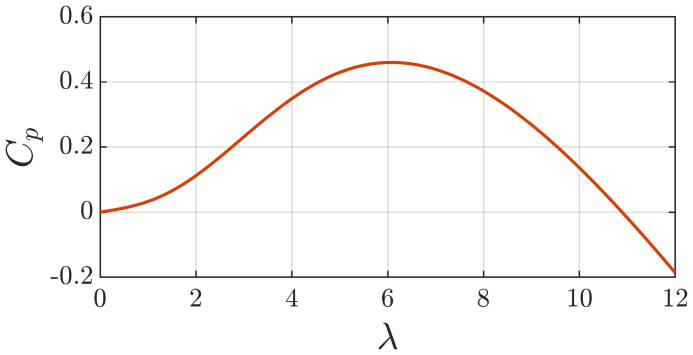

Figure 7. C_p versus λ curve derived from the results presented in [42].

Table 1. Parameters for example HKT.

Parameter	Value
Rated flow velocity $u_{0,rated}$	2.25 m/s
Rotor Length R	0.8538 m
Maximum C_p	0.46
Optimal Tip Speed Ratio λ^*	6.1
Coefficient of Friction C_{fric}	1.667 Nm·s/rad
Moment of Inertia $J_{turbine}$	2.1816 kg·m^2

Table 2. PMSG parameters derived from [34].

Parameter	Value
Rated Power P_{rated}	6 kW
Rated Speed ω_{rated}	15.7 rad/s
Number of Poles p	24
Moment of Inertia J_{gen}	0.26 kg·m^2
Stator Resistance R_s	3.7110 Ω
d-/q-Axis Inductance L_d/L_q	45.25 mH/64.88 mH
Rotor Flux Linkage λ_r	1.9059 Wb

All pertinent information for the power electronic converter is shown in Table 3. For this case study, a Wolfspeed E3M0160120J2 1.2 kV rated SiC MOSFET (Wolfspeed, Durham, NC, USA) has been selected as the example device. It should be noted that details on the internal control loops have been omitted as it is assumed that the dynamics of the power electronic converter are significantly faster than the dynamics of the HKT response. Also, all thermal parameters required for the simplified Cauer thermal model shown in Figure 4 are included in Table 3.

Table 3. Power electronic converter parameters.

Parameter	Value
DC Bus Voltage V_{DC}	800 V
Modulation Scheme	Sinusoidal Pulse Width Modulation
Switching Frequency f_{sw}	20 kHz
MOSFET ON-Resistance $R_{DS,on}$	159 mΩ
Diode Reverse Recovery Charge Q_{rr}	111 nC
Switching Loss Voltage Scaling k_v	1.33
Package Dimensions	10 mm × 9.2 mm
MOSFET Thermal Resistance θ_{JC}	1.3 K/W
TIM Thermal Resistance θ_{TIM}	1.2422 K/W
Heat Sink Thermal Capacitance C_{HS}	$5.43 \cdot 10^3$ J/K
Heat Sink Thermal Resistance θ_{HS}	$50 \cdot 10^{-3}$ K/W

The RL parameters used to define the SARSA agent, the RBF approximation method, and the monetary-based reward function are shown in Table 4. The values of the SARSA parameters (α_0, γ, ε, and ε_d) shown in Table 4 were chosen through a heuristic search. However, some insight can be drawn from [37] on how setting these parameters' values may affect the RL agent's performance. For example, reducing the value of α_0 may make learning more stable but will also extend the required amount of time to converge to the final policy.

Table 4. Reinforcement learning parameters.

	Parameter	Value
SARSA Parameters	Initial Step-Size α_0	0.05
	Discount Factor γ	0.995
	Initial Exploration Rate ε	0.7
	Exploration Rate Decay ε_d	0.6
RBF Parameters	RBF Centers per Dimension n_c	5
	RBF Width σ	0.125
	Base Feedback Coefficient k_b	1.445 Nm/s
	Base Flow Velocity u_b	3.5 m/s
	Base Flow Temperature T_b	26 °C
	Base Energy Cost c_b	USD 0.55/kWh
Reward Parameters	Sample time T_s	350 s
	Estimated Converter Cost c_c	USD 5000
	Reference Temperature T_{ref}	353.15 K
	Reference Converter Lifetime L_c	12,000 h
	Activation Energy E_a	0.9 eV
	Boltzmann's Constant k_B	$8.617 \cdot 10^{-5}$ eV/K

The values used for the RBF approximation method were determined as follows. The value of n_c was chosen heuristically, converging on a value that enables sufficient mapping of Q-values in each dimension without causing a significant strain on memory. With the final value of n_c (5) and the four states and one action the agent must consider, the feature vector is constructed of 3125 different RBF center points. The value of σ was chosen to be one-half the distance between the evenly spaced RBF center points in each dimension. The base value k_b was designated to be the value that would ensure the turbine system would track the optimal value of λ and maximize C_p for all flow conditions. The base values u_b, T_b, and c_b were designated to be the maximum values the agent could expect for flow velocity, water temperature, and cost of energy, respectively.

Finally, the values used for the variables that comprise the reward function were determined as follows. The value of T_s was selected to be approximately equal to seven multiples of the time constant created by the thermal capacitance and resistance of the heat sink, ensuring the device reaches thermal steady state. The value for c_c was chosen through a survey of similarly rated power electronic converters on the market. The values for T_{ref}, L_c, and E_a were derived from examples from the AEC-Q101 Rev. E documentation [43] as these values would have to be defined from a variety of experimental lifetime tests for the specific power electronic device in each system.

Mean water velocity and water temperature data from a river near the central California coast with tidal influence were collected from the United States Geological Survey (USGS) online database [44]. The raw flow velocity data from the USGS database were normalized and scaled such that they had a mean flow velocity of 1.5 m/s, a middle ground in terms of the regions of control.

Lastly, as the reward function requires energy price data, historical data on the locational marginal price (LMP) from the day-ahead market for Pacific Gas and Electric were collected from [45]. Data could only be found from March 2019 to October 2024, which limited the available training data for the agent. The LMP data were also normalized and scaled so the LMP profiles for each year reached a maximum of USD 0.52/kWh, the absolute maximum price over the defined period from 2019 to 2024. This was conduct3ed to ensure ample exploration of the LMP state space, considering this region's lack of LMP data. These profiles of LMP data are then used for the values of c_e in (23).

4.3. Training Method

Training enables the RL agent to explore the state and action space to find the best policy over all years in the defined dataset. This section will analyze the cumulative reward for each training year to verify that the RL agent has converged on a suitable policy.

The training dataset is composed of LMP and flow data from 2020 to 2023, with the RL agent interacting with each year 50 times (i.e., 50 training epochs). After each training epoch, the exploration rate ε is reduced following (20) to promote the agent to exploit known "good" actions as the agent experiences more of the state space, while the learning rate α is also reduced to mitigate forgetfulness following (21). The specific training year is randomly selected from the available dataset so that the agent does not learn the relative pattern of flow and LMP data from year to year. Figure 8 shows the LMP, T_{amb}, and u_0 data for the training years 2020 to 2023, respectively. As shown, the data for the natural environment (T_{amb} and u_0) are similar between years, but the LMP profiles vary significantly from year to year.

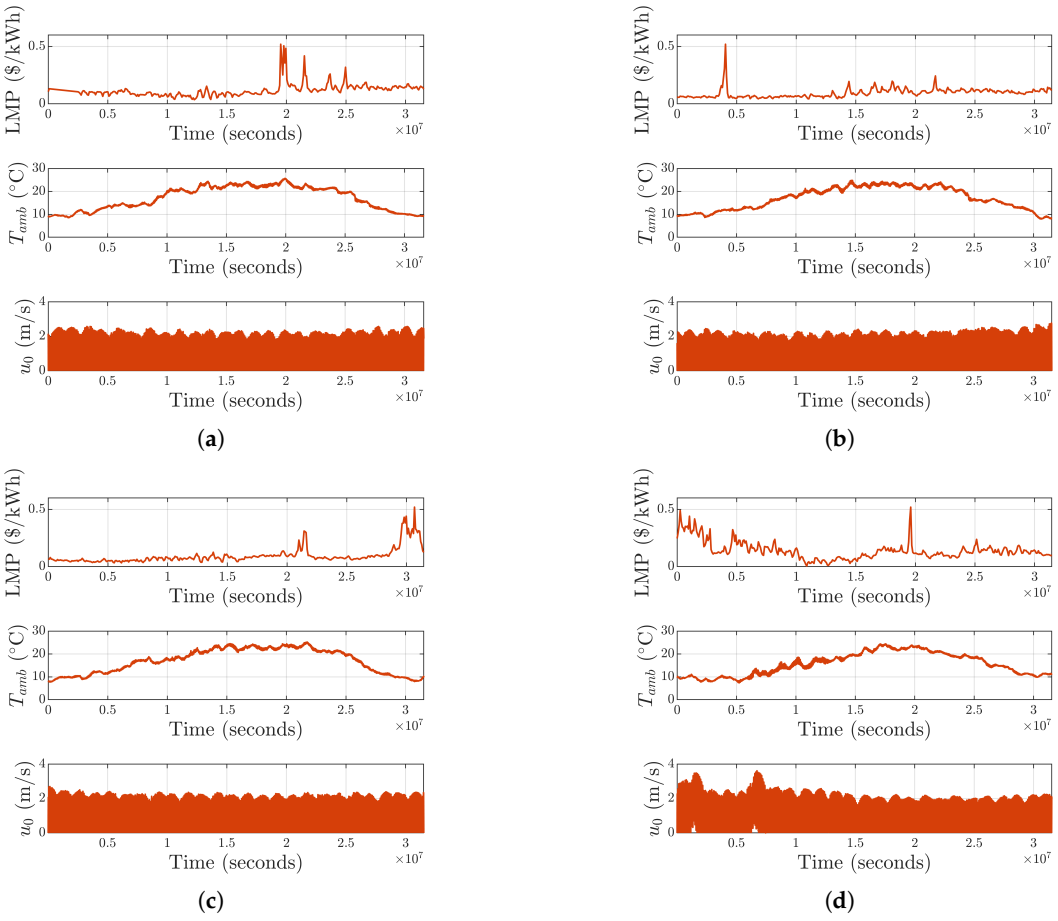

Figure 8. Data representative of the training environment in terms of locational marginal price (LMP), ambient temperature (T_{amb}), and flow velocity (u_0) for (**a**) 2020, (**b**) 2021, (**c**) 2022, and (**d**) 2023.

The total reward collected over a complete training iteration (one year) was used to analyze the success of the training protocol. As the LMP data shown in Figure 8 vary significantly between training years, there were large differences in the reward accumulated over each year in the same epoch. Therefore, to analyze how much of an improvement the RL method was making in each year, the cumulative reward for each year at each training epoch was normalized by the cumulative reward collected in the first training epoch in each respective year. These results are shown in Figure 9. As indicated, the normalized reward for 2023 is significantly higher than the other three test years, which can largely be attributed to the higher mean LMP values for 2023. This also shows that years with higher mean LMP values often underwent consistent improvements in accumulated reward throughout the training. For example, the reward continually increases until convergence for the 2023 dataset. In summary, the reward begins to plateau around the 10th epoch as the policy begins to converge and the learning rate is reduced, ultimately achieving increases in accumulated reward of 1.4%, 1.9%, 3.8%, and 6.9% in 2020, 2021, 2022, and 2023, respectively.

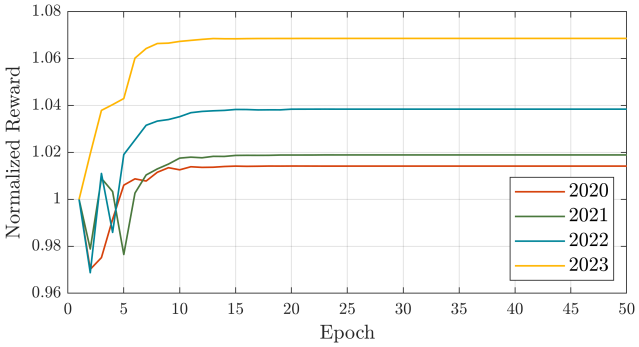

Figure 9. Normalized reward profiles for each training year in the training set. The reward signals for each year are normalized by the initial reward received under the highly exploratory policy in the first training epoch.

4.4. Case Study Comparison

Two case studies have been created using the partial datasets from 2019 and 2024. The goal of these case studies is to verify the effectiveness of the proposed RL method against a common baseline, specifically looking at the annual energy harvested $E_{turbine}$ and annual LC. In this case, a static k-omega-squared controller acts as the baseline, with the feedback gain k optimally tuned such that the HKT continuously tracks the optimal λ throughout Region 2.

Figure 10 shows the environmental conditions for the 2019 and 2024 case study years, respectively. Similar to the training environments, the profiles of T_{amb} and u_0 are quite similar between the two case study years. However, the LMP data in 2019 are significantly higher than in 2024.

Figure 11 shows the applied values of k for 2019 and 2024, respectively. A key takeaway from this figure is that, in relation to Figure 10, as T_{amb} increases and peak values of u_0 decrease toward the midpoint of the mission profile, the applied values of k also begin to decrease in both years. This phenomenon is somewhat intuitive as T_{amb} acts as a lower bound for T_j in reference to the thermal network in Figure 4. Therefore, as T_{amb} increases, the cost associated with the operation of the power electronics estimated using (24) also increases, and torque should be reduced to minimize these costs. Also, during instances of high environmental temperatures, rapid increases in the applied values of k can be found around local peaks in LMP data. This phenomenon is much more apparent in the 2024 case study, most notably between 1.75×10^7 and 2.25×10^7 s. This shows that, despite the increased negative reward that accompanies the high environmental temperatures, the

reward associated with maximizing energy outweighs the negative rewards associated with device aging.

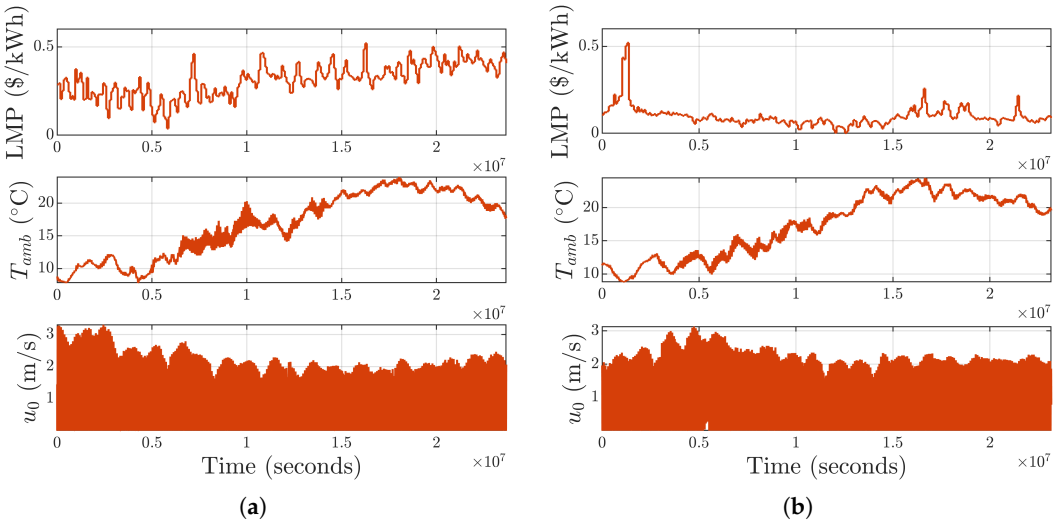

Figure 10. Plots of ambient temperature T_{amb}, LMP data, and flow velocity u_0 for the (**a**) 2019 and (**b**) 2024 case studies, respectively.

The most significant results that can be drawn from the plots in Figure 11 are the differences in the applied values of k between 2019 and 2024. Referring to Figure 10, the LMP data in 2019 are significantly higher than in 2024. This difference is highlighted by the k profiles between the two years. In 2019, the high cost of energy leads the RL agent to apply values of k much closer to k_{opt} to maximize energy generation at the expense of damaging the power electronic devices. In 2024, the low cost of energy drives the RL agent to favor the preservation of the power electronics by applying values of k that are significantly lower than k_{opt}, except in times where the cost of energy spikes. Therefore, significant reductions in energy generation can be expected in the 2024 case study.

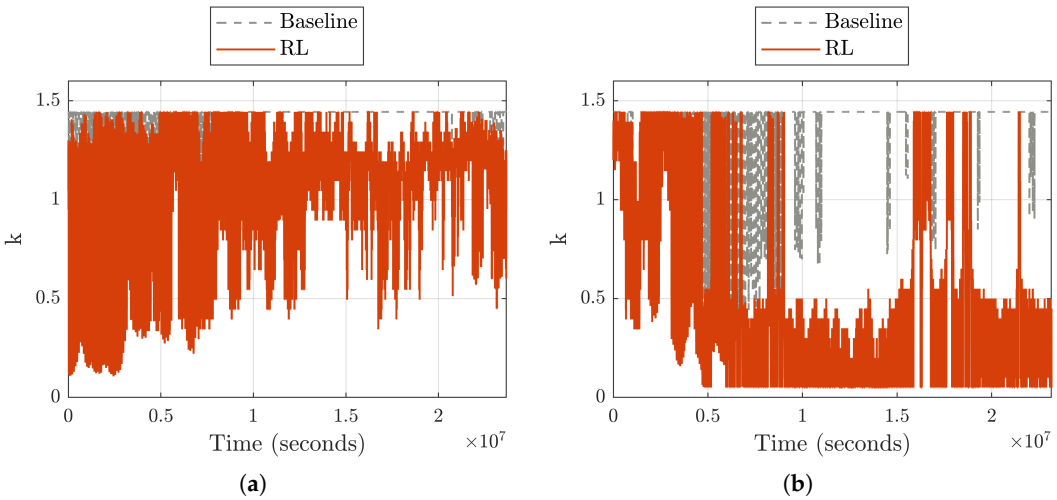

Figure 11. Profiles of feedback gain values k applied by the baseline and RL-based controllers for (**a**) 2019 and (**b**) 2024.

The thermal profiles of T_j for the baseline and RL-based control for both case study years are shown in Figure 12. The general relationship between the two cases is that both the thermal cycle amplitudes ΔT_j and mean junction temperatures are significantly lower under the RL-based control method compared to the baseline. The results in Figure 13 show that, under the proposed control scheme, the number of thermal cycles with amplitude

greater than 20 °C has been greatly reduced, but the number of low-amplitude thermal cycles less than 5 °C has increased. As ΔT_j is the primary player in estimating the amount of damage accumulated by the power electronic converter, even slight reductions can lead to significant savings in *LC*. It should also be noted that ΔT_j has a non-linear relationship with N_f in reference to (13), implying that larger-amplitude thermal cycles will cause more damage to the device than low-amplitude thermal cycles. Following the method outlined in the Lifetime Estimation section, it was found that this assumption holds, and that *LC* using the RL-based control scheme was lower compared to the baseline by 76.7% and 84.5% in 2019 and 2024, respectively.

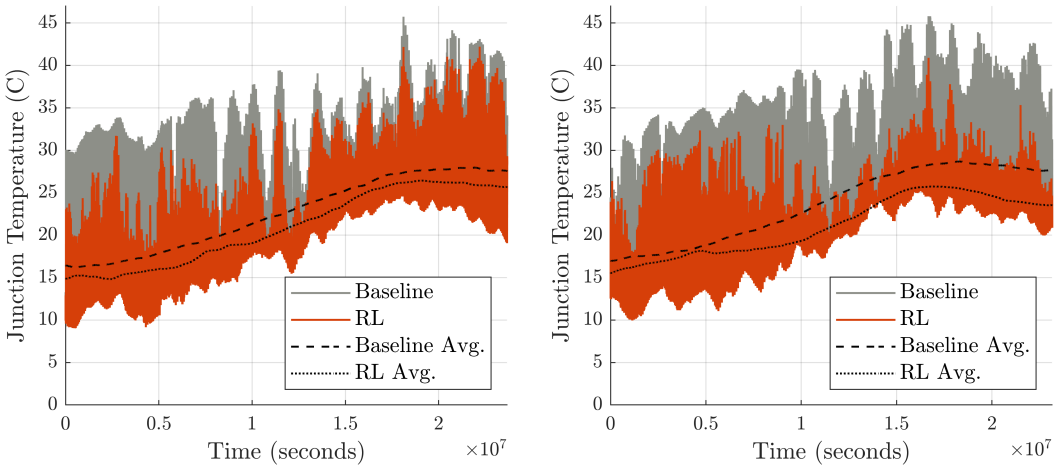

Figure 12. Junction temperature (T_j) profiles for both the baseline and RL-based control case studies completed for (**a**) 2019 and (**b**) 2024. The dashed and dotted lines are rolling mean profiles of T_j with an averaging window of three months for the baseline and RL-based control cases, respectively.

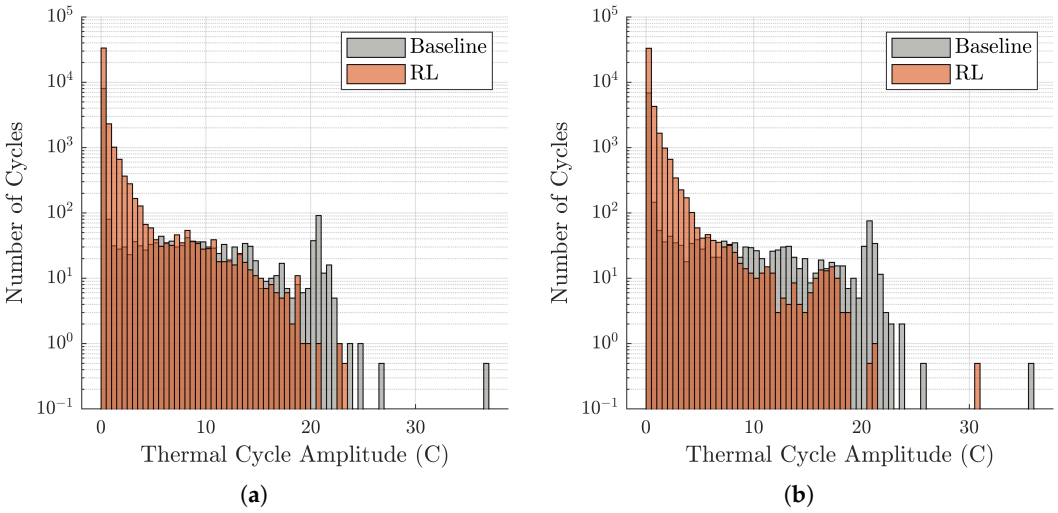

Figure 13. Extracted thermal cycle amplitude data from the output of the rainflow counting algorithm highlighting the number of thermal cycles experienced at each stress level during the case studies completed for (**a**) 2019 and (**b**) 2024.

Figure 14 shows the accumulation of energy harvested by the turbine ($E_{turbine}$) while using both control schemes. As shown, the RL-based controller under-performs compared to the baseline in this category. Considering the nature of the proposed RL method, this outcome is expected as any deviation from the optimal value of k will lead to lower

values of C_p, which, in turn, reduces $P_{turbine}$. However, the impacts on $E_{turbine}$ varied quite significantly between the two years. In a side-by-side comparison, the RL method reduced the annual energy harvested by 5.8% and 25.2% in 2019 and 2024, respectively, when compared to the baseline. Referring to the k profiles in Figure 11, this difference in energy generation can be explained by the large difference in applied values of k between the two years. In 2024, the applied values of k are significantly lower, yielding both higher rotational speeds and values of λ that stray further from the optimal point, resulting in values of C_p that are much lower than the maximum.

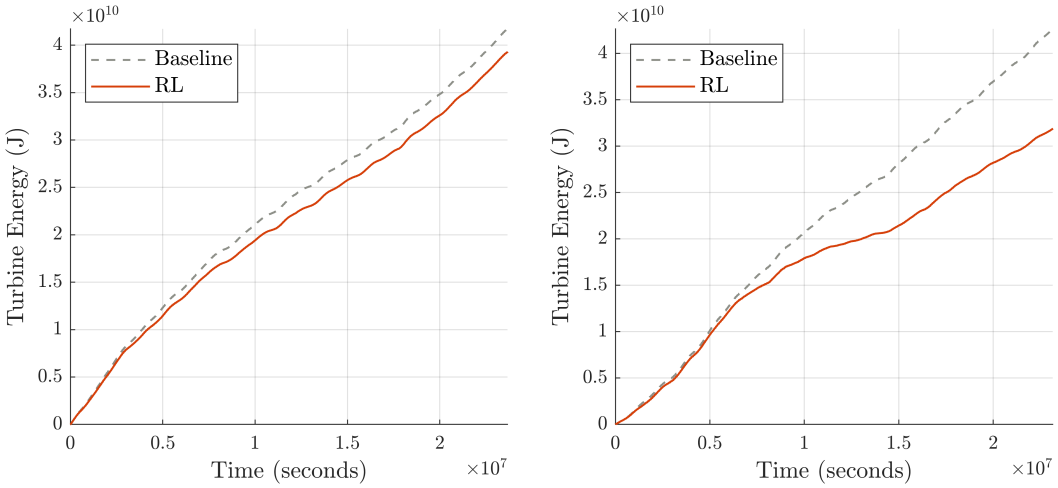

Figure 14. Energy harvested by the turbine ($E_{turbine}$) using both the baseline and RL-based control schemes for the (**a**) 2019 case study and (**b**) 2024 case study.

5. Discussion

5.1. Performance

During the training protocol, the RL agent converged to a policy within 15 epochs, which led to an improvement in terms of reward for all four training years. The results for 2023 found the largest benefit during the proposed training, with significantly higher increases in rewards and a much more consistent reward profile. The RL agent's strong performance during the 2023 training sessions is just one example of how the cost of energy or the LMP profile influences its learning ability. As shown, the reward impacts are directly related to the average cost of energy each year: the lower the energy costs, the lower the total reward accumulated by the RL agent.

For the case study, the focal point was comparing the life consumption and the total annual energy generation between the RL-based method and the baseline controller for the two mission profiles. A summary of the case study results is shown in Table 5, where LC Impact, Energy Impact, and Revenue Impact are defined as the relative differences between the values calculated for the RL-based control scheme and the baseline method normalized by the values achieved by the baseline method. In Table 5, a negative sign represents a relative decrease and a positive sign represents a relative increase in the respective category. It should be noted that the estimates for revenue have been calculated as the difference in reward signals r_g and r_c, which are defined in (23) and (24), respectively.

While it has been shown that the RL method leads to significant reductions in *LC* in both case studies, the impacts on energy generation are non-negligible, especially in 2024. The Revenue Impact category provides more insight as to how the system costs will be impacted under the proposed RL method. These results show that, despite the 25% reduction in energy generation in 2024, the estimated revenue increases by over 45%

when the RL method is in use. On the other hand, it was found that, in 2019, there was a reduction in estimated income of 1.3% under the proposed control method. These impacts show how crucial the LMP profile is in determining how useful the proposed RL method is. Further, 2019 consistently has higher energy costs than in 2024, as shown in Figure 10. With these higher energy costs, the potential benefits of extending the lifetime of power electronics do not outweigh the relative revenue coming in from the generation of energy. On the other hand, in 2024, when the energy costs are significantly lower with short-term spikes, the RL method excels because the relative cost of the power electronic converter is on the same order of magnitude as the price of energy.

Table 5. A summary of the impacts on life consumption, energy generation, and estimated revenue comparing the proposed RL method to the baseline static k-omega-squared controller.

Year	LC Impact	Energy Impact	Revenue Impact
2019	−76.7%	−5.8%	−1.3%
2024	−84.5%	−25.2%	+45.4%

In summary, as shown by the results of these two case studies, the potential improvements in the RL-based control system in tidal HKT applications are highly dependent on the current energy market. In areas where the price of energy is high, the benefits in extending the lifetime of the power electronic converter do not outweigh the loss of revenue from reduced HKT performance. On the other hand, in areas where the energy market covers a wider range of energy costs, the proposed RL method has shown to lead to significant increases in estimated net income.

5.2. Implications and Future Work

First, to use the RBF-based approximation method, the designer must designate a large enough state space that includes all the conditions the RL agent may experience. If the RL agent experiences states outside of what has been defined, the learned Q-surface may be heavily skewed to approximate the Q-value of these outlier states, resulting in poor performance within the defined state space. Considering the high variability in LMP data, it may be significantly challenging to decipher the total range that the LMP data may cover while the RL agent is online. More extensive datasets with various LMP profiles would provide the designer more insight into what the landscape for the energy market may look like for a given year, but even possessing extensive historical data may not be sufficient to estimate what might happen in the future. Therefore, it may be beneficial to explore other function approximation methods that do not require explicit bounds on the allowable state space.

Second, the deployed environment for the HKT can significantly impact the lifetime of the power electronic converter. For example, in riverine environments, the fluctuations in flow velocity are of much lower frequency, typically on the order of months. In contrast, the flow velocity undergoes multiple cycles per day for tidal applications, as presented in this work. As flow conditions play a major role in the thermal cycling of the device, the proposed RL method would likely have less of a positive impact on the lifetime of the converter in riverine applications. A side-by-side comparison of the relative impacts on LC and $E_{turbine}$ should be carried out for both tidal and riverine applications to more broadly generalize how well the proposed RL-based control scheme performs in these two cases.

Third, several electrical characteristics were simplified in this work, such as generator efficiency, field weakening effects, and switching/fundamental frequency effects on the thermal response of the power electronics. These characteristics may have a significant impact on the cumulative energy generation of the HKT, as well as the device losses. For example,

the implementation of the RL method may lead to an increased number of fundamental frequency cycles as the speed of the turbine is higher under this control scheme, which may reduce the overall *LC* impact realized through the case study. In future works, these effects could be realized by utilizing a multi-physics modeling scheme such as the HKT-specific method proposed in [34]. After the RL agent has been trained on the high-level HKT model, the switching-level effects and generator impacts can be analyzed to present a more realistic picture of the device's thermal response and overall energy generation.

Fourth, this work only considered the impacts of ΔT_j and T_{min} on the life consumption of the device. As noted before, there are several other more complex lifetime consumption estimation models, but, without accurate scaling terms, the estimates for *LC* are significantly devalued. Although there are a number of resources for IGBT devices, little to no data have been published for SiC MOSFETs. Future work should aim to either complete the required lifetime tests and derive the pertinent scaling terms or partner with a SiC device manufacturer willing to share these values.

Lastly, as mentioned earlier, deploying RL in continuous tasks makes the agent susceptible to forgetfulness. In serious cases, forgetfulness can lead to the agent becoming untrained over time. Forgetfulness can be solved in several ways, broadly classified as memory-based replay, regularization, and parameter isolation approaches [39]. However, in this work, learning is purposely stopped using the learning parameter α to avoid this problem. Although this method proves convergence, this approach does not provide a guarantee that the policy learned by the agent corresponds to optimal performance as learning is hard-coded to slow down over time. Future works should focus on methods to prevent or minimize the effects of forgetfulness. This should be conducted either through one of the advanced approaches listed above or by restructuring the training protocol.

6. Conclusions

This paper has proposed a lightweight SGD SARSA algorithm utilizing a Gaussian RBF approximation method to balance the trade-off between energy generation and power electronic converter lifetime in a marine HKT. The proposed control scheme was developed within a common k-omega-squared torque control framework, in which the RL agent varies the applied value of k depending on the state of the environment. The RL agent was trained using a reward function that estimated the net revenue of the turbine, considering both the income from the sale of energy in a dynamic energy market and the cost of operating the power electronic converter at a specific junction temperature. The training environment was derived from real-world flow velocity, flow temperature, and LMP data, creating a realistically difficult challenge for the RL agent to overcome. The results of the training protocol demonstrated that the RL agent achieved the greatest improvement with a higher cost of energy. Also, the training revealed the method's susceptibility to forgetfulness, which was solved by controlling the learning rate. The trained RL agent was then compared against an optimally tuned static k-omega-squared torque controller in two separate case studies. The case studies showed that the applied value of k (and, hence, the amount of power generated by the HKT) was directly related to the cost of energy at a given moment. Higher energy costs promoted the RL agent to maximize energy generation, while lower energy costs led to a reduction in energy generation to preserve the lifetime of the power electronics. In both case studies, it was also found that utilizing the proposed RL method reduced the mean junction temperature and the number of high-amplitude thermal cycles but increased the number of low-amplitude thermal cycles. As high-amplitude thermal cycles have more of an impact on life consumption, the RL-based scheme had less impact on the power electronic converter lifetime than the baseline controller. However, the reduction in thermal stress on the device also resulted in a non-negligible reduction in the cumulative

energy harvested by the HKT. To quantify the relative improvements achieved by the proposed method, a custom estimate for net revenue considering the sale of energy and the cost of operating the power electronic converter was used. At the end of the case study, the improvements achieved by the proposed control method were determined by the state of the energy market. In the best case, the proposed adaptive control scheme increased the estimated net revenue by 45.4%. In the worst case, the RL-based method led to a net loss in revenue of 1.3%. These outcomes show that the RL method is more valuable in scenarios where the cost of energy is comparable to the cost of the converter such that decreasing the energy generated by the HKT to reduce the damage to the power electronic devices does not result in significant losses in revenue.

Author Contributions: Conceptualization, S.B., T.K.A.B. and Y.C.; methodology, S.B. and T.K.A.B.; writing—original draft preparation, S.B.; writing—review and editing, T.K.A.B. and Y.C.; supervision, T.K.A.B. and Y.C.; project administration, T.K.A.B. and Y.C.; funding acquisition, T.K.A.B. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the US Department of Energy Advanced Research Projects Agency-Energy (ARPA-E) under Award DE-AR0001438 and in part by the US Department of Energy Water Power Technologies Office under Award DE-EE0011381.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ATC	Active Thermal Control
BTB	Back-to-Back
HKT	Hydrokinetic Turbine
kWh	Kilowatt hour
LC	Life Consumption
LCOE	Levelized Cost of Energy
LMP	Locational Marginal Price
MOSFET	Metal–Oxide–Semiconductor Field-Effect Transistor
MPPT	Maximum Power Point Tracking
PMSG	Permanent Magnet Synchronous Generator
PoF	Physics of Failure
RBF	Radial Basis Function
RL	Reinforcement Learning
RMS	Root Mean Square
SARSA	State–Action–Reward–State–Action
SGD	Stochastic Gradient Descent
SiC	Silicon Carbide
TD	Temporal Difference
TIM	Thermal Interface Material
USGS	United States Geological Survey

References

1. Kilcher, L.; Fogarty, M.; Lawson, M. *Marine Energy in the United States: An Overview of Opportunities*; Technical Report; National Renewable Energy Laboratory: Golden, CO, USA, 2021.
2. Johnson, J.B.; Pride, D.J. *River, Tidal, and Ocean Current Hydrokinetic Energy Technologies: Status and Future Opportunities in Alaska*; Technical Report; Alaska Center for Energy and Power: Fairbanks, AK, USA, 2010.
3. Tehrani, K.; Beikbabaei, M.; Mehrizi-Sani, A.; Jamshidi, M. A smart multiphysics approach for wind turbines design in industry 5.0. *J. Ind. Inf. Integr.* **2024**, *42*, 100704. [CrossRef]
4. Kirke, B. Towards more cost-effective river hydrokinetic turbines. *Energy Sustain. Dev.* **2024**, *78*, 101370. [CrossRef]
5. Carroll, J.; McDonald, A.; McMillan, D. Reliability Comparison of Wind Turbines With DFIG and PMG Drive Trains. *IEEE Trans. Energy Convers.* **2015**, *30*, 663–670. [CrossRef]
6. El-Naggar, M.F.; Abdelhamid, A.S.; Elshahed, M.A.; El-Shimy Mahmoud Bekhet, M. Ranking Subassemblies of Wind Energy Conversion Systems Concerning Their Impact on the Overall Reliability. *IEEE Access* **2021**, *9*, 53754–53768. [CrossRef]
7. Pu, S.; Yang, F.; Vankayalapati, B.T.; Akin, B. Aging Mechanisms and Accelerated Lifetime Tests for SiC MOSFETs: An Overview. *IEEE J. Emerg. Sel. Top. Power Electron.* **2022**, *10*, 1232–1254. [CrossRef]
8. Kuprat, J.; Van Der Broeck, C.H.; Andresen, M.; Kalker, S.; Liserre, M.; De Doncker, R.W. Research on Active Thermal Control: Actual Status and Future Trends. *IEEE J. Emerg. Sel. Top. Power Electron.* **2021**, *9*, 6494–6506. [CrossRef]
9. Ibrahim, A.; Salem, M.; Kamarol, M.; Delgado, M.T.; Desa, M.K.M. Review of Active Thermal Control for Power Electronics: Potentials, Limitations, and Future Trends. *IEEE Open J. Power Electron.* **2024**, *5*, 414–435. [CrossRef]
10. Zhang, J.; Du, X.; Qian, C.; Du, R.; Hu, X.; Tai, H.M. Thermal Management of IGBT Module in the Wind Power Converter Based on the ROI. *IEEE Trans. Ind. Electron.* **2022**, *69*, 8513–8523. [CrossRef]
11. Chihaia, R.A.; Vasile, I.; Cîrciumaru, G.; Nicolaie, S.; Tudor, E.; Dumitru, C. Improving the Energy Conversion Efficiency for Hydrokinetic Turbines Using MPPT Controller. *Appl. Sci.* **2020**, *10*, 7560. [CrossRef]
12. Zhang, X.; Jia, J.; Zheng, L.; Yi, W.; Zhang, Z. Maximum power point tracking algorithms for wind power generation system: Review, comparison and analysis. *Energy Sci. Eng.* **2023**, *11*, 430–444. [CrossRef]
13. Alhmoud, L. Reliability Improvement for a High-Power IGBT in Wind Energy Applications. *IEEE Trans. Ind. Electron.* **2018**, *65*, 7129–7137. [CrossRef]
14. Andresen, M.; Ma, K.; Buticchi, G.; Falck, J.; Blaabjerg, F.; Liserre, M. Junction Temperature Control for More Reliable Power Electronics. *IEEE Trans. Power Electron.* **2018**, *33*, 765–776. [CrossRef]
15. Chen, Y.; Wang, L.; Liu, S.; Wang, G. A Health-Oriented Power Control Strategy of Direct Drive Wind Turbine. *IEEE Trans. Power Deliv.* **2022**, *37*, 1324–1335. [CrossRef]
16. Cui, H.; Guo, T.; Yang, C.; Dai, Y.; Wang, C.; Du, H.; Qin, L.; Yu, H. A New Thermal Management Strategy of IGBT in DFIG for Economic Benefit Maximization. *IEEE Trans. Ind. Inform.* **2024**, *20*, 1335–1347. [CrossRef]
17. Stringer, C.C.; Polagye, B.L. Implications of biofouling on cross-flow turbine performance. *SN Appl. Sci.* **2020**, *2*, 1–13. [CrossRef]
18. Walker, J.M.; Flack, K.A.; Lust, E.E.; Schultz, M.P.; Luznik, L. Experimental and numerical studies of blade roughness and fouling on marine current turbine performance. *Renew. Energy* **2014**, *66*, 257–267. [CrossRef]
19. Zhao, S.; Blaabjerg, F.; Wang, H. An Overview of Artificial Intelligence Applications for Power Electronics. *IEEE Trans. Power Electron.* **2021**, *36*, 4633–4658. [CrossRef]
20. Nambiar, A.; Anderlini, E.; Payne, G.; Forehand, D.; Kiprakis, A.; Wallace, A. Reinforcement Learning Based Maximum Power Point Tracking Control of Tidal Turbines. In Proceedings of the European Wave and Tidal Energy Conference, Cork, Ireland, 27 August–1 September 2017.
21. Hasankhani, A.; Tang, Y.; VanZwieten, J.; Sultan, C. Comparison of Deep Reinforcement Learning and Model Predictive Control for Real-Time Depth Optimization of a Lifting Surface Controlled Ocean Current Turbine. In Proceedings of the 2021 IEEE Conference on Control Technology and Applications (CCTA), San Diego, CA, USA, 9–11 August 2021; pp. 301–308. [CrossRef]
22. Hasankhani, A.; Ondes, E.B.; Tang, Y.; Sultan, C.; Van Zwieten, J. Integrated Path Planning and Tracking Control of Marine Current Turbine in Uncertain Ocean Environments. In Proceedings of the 2022 American Control Conference (ACC), Atlanta, GA, USA, 8–10 June 2022; pp. 3106–3113. [CrossRef]
23. Hasankhani, A.; Tang, Y.; VanZwieten, J. Reinforcement Learning for Underwater Spatiotemporal Path Planning, with Application to an Autonomous Marine Current Turbine. In Proceedings of the 2023 American Control Conference (ACC), San Diego, CA, USA, 31 May–2 June 2023; pp. 3715–3721. [CrossRef]
24. Ouyang, Y.; Zhao, W.; Wang, H. Simulation and study of maximum power point tracking for rim-driven tidal current energy power generation systems. *Energy Rep.* **2023**, *9*, 792–801. [CrossRef]
25. Barton, S.; Brekken, T.K.; Cao, Y. Reinforcement Learning Control for Enhancing Marine Hydrokinetic Turbine Energy Generation. In Proceedings of the 2024 American Control Conference (ACC), Toronto, ON, Canada, 10–12 July 2024; pp. 1044–1050. ISSN: 2378-5861. [CrossRef]

26. Chu, A.; Xie, X.; Hermann, C.M.; Stork, W.; Roth-Stielow, J. Towards Predictive Lifetime-Oriented Temperature Control of Power Electronics in E-vehicles via Reinforcement Learning. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 1667–1676. [CrossRef]
27. Chu, A.; Hermann, C.M.; Silz, J.; Pfau, J.; Barón, K.M.; Anantharajaiah, N.; Schmidt, P.; Hotfilter, T.; Xie, X.; Becker, J.; et al. LETSCOPE: Lifecycle Extensions Through Software-Defined Predictive Control of Power Electronics. In Proceedings of the IEEE EUROCON 2023—20th International Conference on Smart Technologies, Torino, Italy, 6–8 July 2023; pp. 665–670. [CrossRef]
28. Sale, D.; Jonkman, J.; Musial, W. Development of a Hydrodynamic Optimization Tool for Stall-Regulated Hydrokinetic Turbine Rotors. In Proceedings of the Volume 4: Ocean Engineering; Ocean Renewable Energy; Ocean Space Utilization, Parts A and B, Honolulu, HI, USA, 31 May–5 June 2009; pp. 901–906. [CrossRef]
29. Mohan, N.; Raju, S. *Analysis and Control of Electric Drives: Simulations and Laboratory Implementation*; Wiley: Hoboken, NJ, USA, 2020.
30. Lei, T.; Barnes, M.; Smith, S.; Hur, S.h.; Stock, A.; Leithead, W.E. Using Improved Power Electronics Modeling and Turbine Control to Improve Wind Turbine Reliability. *IEEE Trans. Energy Convers.* **2015**, *30*, 1043–1051. [CrossRef]
31. Graovac, D.D.; Pürschel, M.; Kiep, A. *MOSFET Power Losses Calculation Using the Data- Sheet Parameters*; Infineon Application Note: Neubiberg, Germany, 2006; pp. 1–23.
32. Ma, K.; Liserre, M.; Blaabjerg, F.; Kerekes, T. Thermal Loading and Lifetime Estimation for Power Device Considering Mission Profiles in Wind Power Converter. *IEEE Trans. Power Electron.* **2015**, *30*, 590–602. [CrossRef]
33. Ni, Z.; Lyu, X.; Yadav, O.P.; Singh, B.N.; Zheng, S.; Cao, D. Overview of Real-Time Lifetime Prediction and Extension for SiC Power Converters. *IEEE Trans. Power Electron.* **2020**, *35*, 7765–7794. [CrossRef]
34. Tariquzzaman, M.; Li, P.; Barton, S.J.; Thurlbeck, A.P.; Kilgore, T.; Brekken, T.K.A.; Cao, Y. Multi-physics and Multi-timescale Modeling of Hydrokinetic Turbine Energy Conversion System. *IEEE J. Emerg. Sel. Top. Power Electron.* **2024**, *12*, 6028–6041. [CrossRef]
35. Miner, M.A. Cumulative Damage in Fatigue. *J. Appl. Mech.* **1945**, *12*, A159–A164. [CrossRef]
36. The MathWorks Inc. Rainflow. Available online: https://www.mathworks.com/help/signal/ref/rainflow.html (accessed on 15 September 2024).
37. Sutton, R.S.; Barto, A. *Reinforcement Learning: An Introduction*, 2nd ed.; Adaptive Computation and Machine Learning; The MIT Press: Cambridge, MA, USA, 2020.
38. Reza Amini, M.; Jiang, B.; Liao, Y.; Naik, K.; Martins, J.R.; Sun, J. Control Co-design of a Hydrokinetic Turbine: A Comparative Study of Open-loop Optimal Control and Feedback Control. In Proceedings of the 2023 American Control Conference (ACC), San Diego, CA, USA, 31 May–2 June 2023; pp. 3728–3734. [CrossRef]
39. Evron, I.; Moroshko, E.; Ward, R.; Srebro, N.; Soudry, D. How catastrophic can catastrophic forgetting be in linear regression? In Proceedings of the Thirty Fifth Conference on Learning Theory, London, UK, 2–5 July 2022; Loh, P.L., Raginsky, M., Eds.; PMLR: New York, NY, USA, 2022; Volume 178, pp. 4028–4079.
40. Hayes, C.F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L.M.; Dazeley, R.; Heintz, F.; et al. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Auton. Agents Multi-Agent Syst.* **2022**, *36*, 26.
41. Ellerman, P. Calculating Reliability Using FIT & MTTF: Arrhenius HTOL Model. *MicroNote* **2012**, *1002*, 1–6.
42. Bahaj, A.; Molland, A.; Chaplin, J.; Batten, W. Power and thrust measurements of marine current turbines under various hydrodynamic flow conditions in a cavitation tunnel and a towing tank. *Renew. Energy* **2007**, *32*, 407–426. [CrossRef]
43. Automotive Electronics Council. Failure Mechanism Based Stress Test Qualification for Discrete Semiconductors in Automotive Applications, 2021. Available online: http://www.aecouncil.com/AECDocuments.html (accessed on 3 September 2024).
44. U.S. Geological Survey. Sacramento River at Rio Vista CA—11455420. Available online: https://waterdata.usgs.gov/monitoring-location/11455420/#parameterCode=72255&period=P7D&showMedian=false (accessed on 31 October 2024).
45. GridStatus.io. Available online: https://www.gridstatus.io/live (accessed on 31 October 2024).