World Academy of Science, Engineering and Technology
International Journal of Electrical and Information Engineering
Vol:18, No:3, 2024

Safe and Efficient Deep Reinforcement Learning Control Model: A Hydroponics Case Study

Almutasim Billa A. Alanazi, Hal S. Tharp

***Abstract*—**Safe performance and efficient energy consumption are essential factors for designing a control system. This paper presents a reinforcement learning (RL) model that can be applied to control applications to improve safety and reduce energy consumption. As hardware constraints and environmental disturbances are imprecise and unpredictable, conventional control methods may not always be effective in optimizing control designs. However, RL has demonstrated its value in several artificial intelligence (AI) applications, especially in the field of control systems. The proposed model intelligently monitors a system's success by observing the rewards from the environment, with positive rewards counting as a success when the controlled reference is within the desired operating zone. Thus, the model can determine whether the system is safe to continue operating based on the designer/user specifications, which can be adjusted as needed. Additionally, the controller keeps track of energy consumption to improve energy efficiency by enabling the idle mode when the controlled reference is within the desired operating zone, thus reducing the system energy consumption during the controlling operation. Water temperature control for a hydroponic system is taken as a case study for the RL model, adjusting the variance of disturbances to show the model's robustness and efficiency. On average, the model showed safety improvement by up to 15% and energy efficiency improvements by 35%-40% compared to a traditional RL model.

***Keywords*—**Control system, hydroponics, machine learning, reinforcement learning.

I. Introduction

HYDROPONICS is a significant system in new agricultural methods. It utilizes nutrient-rich water rather than soil for plant nourishment. In the past, hydroponics did not consider water temperature, but some studies show maintaining the water temperature in the hydroponic system in the water reservoir has many effects on the growth processes from the initial stages of development to flower formation [1]-[4]. Some experts agree that the best water solution temperature for hydroponics is between 65 °F (18 °C) and 80 °F (26 °C). Therefore, for healthy roots and the best nutrition uptake, this temperature range is recommended [5]. These innovative farming methods promise sustainable food production, especially for challenging environments like the state of Arizona, a desert environment where temperatures are high and water sources are rare. With a growing need for food due to global population increase, engineers are constantly putting forth innovative agricultural technologies.

Such innovation may be a boon, but safety is an important aspect of implementing these technologies, especially for control systems applications, which control other hardware components that might affect the surrounding environment. Some effects to the environment may occur from malfunction of components due to hardware constraints. In addition, the external environment might be extreme such that the installed hardware may not meet the desired specifications. In these cases, letting the agent/controller operate with unsuccessful accomplishments might lead to unsafe region operation. This is particularly true for hydroponic systems, which can be installed in residential homes, restaurants, hospitals, or large vertical farming for industrial purposes. Consequences of continued under performance can be serious, possibly causing damage to the property, plants, or even a loss of life if something like a fire occurred.

Energy consumption is also important with hydroponics, where the new system heating and cooling water consumes more energy than the prior approach using only water pumps and other subsystems. To reduce energy consumption, the model aims to keep the agent running in low power/idle mode just under or above the thresholds which turn the external disturbances into advantages inside the root zone temperature as the environment stochastically changes. Overall, the model seeks to promote safety and energy efficiency while maintaining optimal water temperature in the hydroponic system.

II. Literature Review

This section intends to present three topics in the literature review of improving safety and energy efficiency using RL control model and compare them with the proposed model.

A. Reference Governors

Enforcing safety to a control system via governors has been proposed in many forms in the literature. For instance, using scalar and vector reference governors, command governors, extended command governors, incremental reference governors, feedforward reference governors, parameter governors, and virtual state governors [6], [7] are all suitable methods when the controller can overcontrol the environment/ plant to the desired zone range. However, in some real-world applications, disturbances overwhelm the controller performance, meaning governing the controller via reference

Almutasim Billa A. Alanazi is a Ph.D. student with the Electrical and Computer Engineering Department, The University of Arizona, Tucson, AZ USA (e-mail: abalanazi@arizona.edu).

Hal S. Tharp is Associate Department Head of Electrical and Computer Engineering and Associate Professor of Electrical and Computer Engineering with the Electrical and Computer Engineering Department, The University of Arizona, Tucson, AZ USA (e-mail: tharp@arizona.edu).

governors may not be ineffective.

B. Safe Reinforcement Learning

The process of learning policies that optimize the expectation of the return to assure reasonable system performance with respect to safety concerns is known as "safe reinforcement learning" [8]. Nevertheless, the optimal policy may not guarantee 100% safety under uncertain hardware and environmental disturbances. The proposed model can be integrated after using the safe RL methods of exploring the safer policy to monitor the controller success and performance.

C. Reducing Energy Cost for Heating, Ventilation, and Air Conditioning

Heating, Ventilation, and Air Conditioning (HVAC) is a major energy consumer in residential sectors, and RL has had great success in controlling temperature with high energy efficiency. [9]. However, many RL models used other supplements such as electronic devices, actuators, and sensors that were interconnected with Internet of Things (IoT) technologies. In the current model, the energy efficiency approach is integrated within the agent block and operates automatically. With no external devices required, both hardware cost and energy consumption of the hardware are reduced.

III. Background

A. RL for Control System

RL, a general class of algorithms in the field of machine learning, is highly influenced by the theory of Markov Decision Processes (MDP) [10]. The general relationship framework between RL and control systems is represented by Fig. 1 [11].

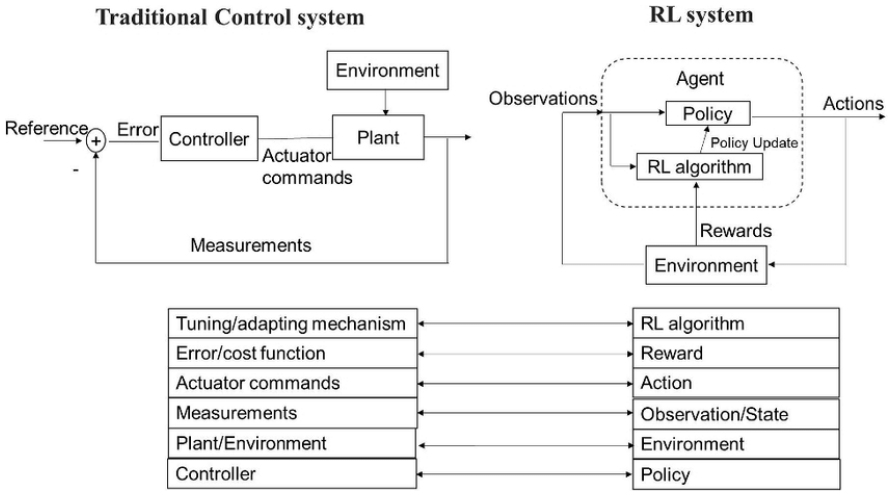

Fig. 1 Traditional RL and classical control system model [11]

The primary components of the RL system are the agent (learner), the environment (where the agent observes and interacts with changing states), the policy (controller that the agent follows to take actions), and the reward (the signal that the agent observes to identify optimality upon taking actions).

Model-based algorithms, exemplified by the SARSA algorithm [12], involve RL in initially learning the model knowledge, which represents the dynamics of the environment, and subsequently deriving the optimal strategy based on this acquired understanding. This approach allows RL to build an internal model of the environment's behavior.

On the other hand, model-free approaches, such as Q-learning algorithms [12] or policy optimization algorithms [13], take a different route. In these algorithms, RL bypasses the explicit learning of the environment's model and instead directly calculates the optimal strategy through trial and error. This model-free strategy is particularly useful in scenarios where the exact dynamics of the environment are complex or unknown, making it challenging to formulate an accurate model.

In essence, model-based RL aims to understand the underlying structure of the environment, leveraging this knowledge for strategy development, while model-free RL focuses on directly optimizing actions without explicit knowledge of the environment's dynamics.

B. Proximal Policy Optimization (PPO) Algorithm

The proposed model used PPO for the RL algorithm. PPO is a deep RL algorithm, one of the model free algorithms introduced in 2017 [13]. PPO algorithms have some of the benefits of trust region policy optimization [13], but they are much simpler to implement, more general, and have better sample complexity. The main principle of PPO - Clipped Surrogate Objective version is that, after an update, the new policy $\pi_\theta(a_t|s_t)$ should be not too far from the old policy $\pi_{\theta_{old}}(a_t|s_t)$.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (1)$$

$$L^{CPI}(\theta) = E_t\left[r_t(\theta)\, A_t^{\pi_{\theta_k}}\right] \quad (2)$$

$$L^{CLIP}(\theta) = E_t\left[\min(r_t(\theta)\, A_t^{\pi_{\theta_k}}, clip(\, r_t(\theta), 1-\epsilon, 1\epsilon) A_t^{\pi_{\theta_k}}\right) \quad (3)$$

The $A_t^{\pi_{\theta_k}}$ is the advantage function which is the difference between the q value for action a in state s and the average value of that state.

$$A_t^{\pi_{\theta_k}}(\mathrm{s},\mathrm{a}) = \mathrm{Q}_\mathrm{t}^{\pi_{\theta_k}}(\mathrm{s}|\mathrm{a}) - V_t^{\pi_{\theta_k}}(s) \quad (4)$$

If the probability ratio between the new policy and the old policy falls outside the range (1 — ε) and (1 + ε), the advantage function will be clipped. ε is recommended and set to 0.2 in the original PPO paper [13]. Basically, the advantage function is a measure of how much a certain action is a good or bad decision given a certain state. The power of PPO algorithms is that after an update, the new policy will not be too far from the old policy. For that, PPO uses clipping to avoid updates that are too large.

IV. Model Methodology

In this model, the agent has four components (policy, RL algorithm, safe performance, save energy). Policy and RL algorithm have been briefly covered in the background section. In Subsections IV *A* and *B*, the additional components to the general RL framework, safe performance and save energy, are explained.

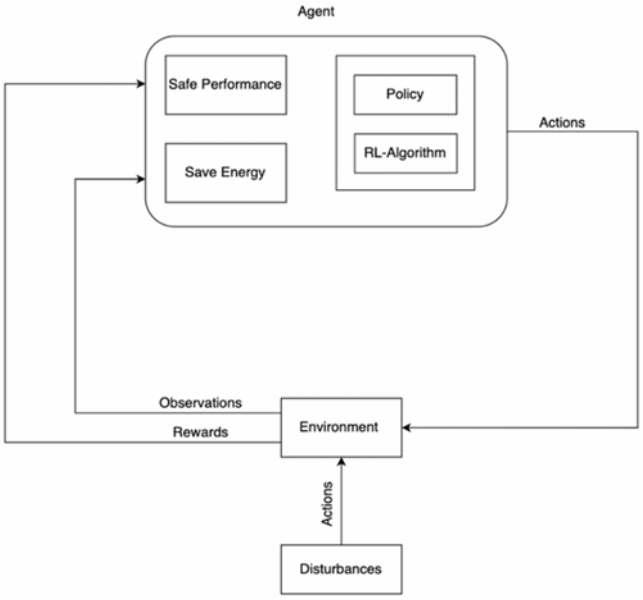

Fig. 2 Proposed Model

A. Safe Performance

Safe Performance (SP) focuses on monitoring agent success through the return rewards from the environment during the RL controller operation. For instance, if the agent is not functioning with good controlling performance for any reason (hardware constraints, rough environment disturbances), SP will turn off the agent, and it will no longer give actions to the environment (system/plant) or receive new observations. Furthermore, SP is not limited to simply turning off the agent; it can also be used to notify the user via an alarm signal, such as a red light or an IoT notification system. Depending on the application, the unsuccessful percentage can be adjusted as needed. For example, in critical applications, the control engineer could specify this metric in the design process. In other noncritical applications, the user may have the choice to modify the desired unsuccessful metric for the system. In this study, the unsuccessful metric is considered as a percentage, the percentage is the ratio between the received rewards and the total expected rewards during the operational process. For example, if the metric is set to be 25%, the SP will monitor the controller performance, and in the case of 25% failure, the SP will send a signal to turn off the agent for safety concerns.

B. Save Energy

Save Energy (SE) enables the agent to be in an idle mode when the environment is within the desired region, but it switches the agent to active mode when the environment (system/plant) needs to be controlled. It determines its state is by constantly monitoring the environment's observations. The policy has the highest computation operation due to its layer analysis starting with the inputs (observations) and continuing to output (actions). Therefore, when the agent is in an idle mode, the policy will not receive observations or give actions, which will reduce the controller power consumption. In the meantime, the disturbances' actions are the ones who change the environmental states. Research on idle/sleeping modes is ongoing in the embedded system field. Some experimental academic research showed a power consumption reduction of 48%-80% when an embedded system was in idle/sleeping mode [14].

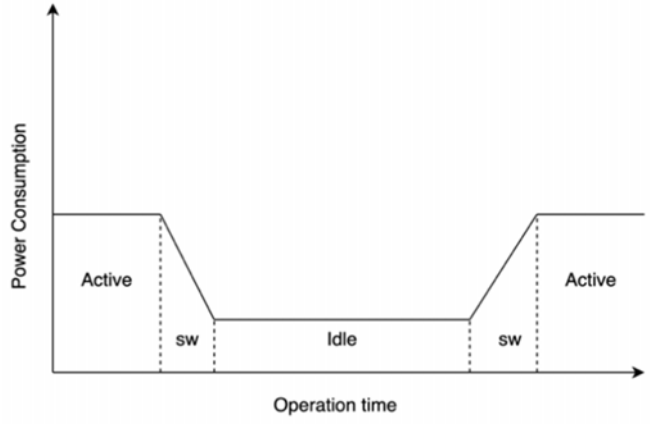

Fig. 3 Active and idle illustration

V. Experimental Setup

The experiment was designed to control temperature of water in a hydroponic system, with the desired range being between 18 °C and 26 °C. The experiment was divided into three stages: setting up the RL environment; training the agent; and testing and evaluating the model. All these steps were carried out with the help of OpenAI Gym toolkit [15] and Stable-Baselines3 [16] using Python.

A. Creating the RL - Environment

The RL environment can be real or simulated. In this study, the simulated environment was chosen because it gives the

ability to trial and train in a safe and cost-effective manner. The main goal of the environment is to be responsive to the agent's actions and to external disturbances. Based on those conditions, it gives new observations and rewards. *Class hydroponic (Env)* is the created environment, and it was tested for one episode for 120-time steps (minutes) with random actions to observe the environmental response.

B. Training the Agent

Training was performed via Stable-Baselines3 [16] using the PPO algorithm [13]. The agent was trained for a total of 100,000 timesteps with random initial conditions. If the temperature falls within the desired temperature range, the agent will be rewarded positively (+1); otherwise, it will be penalized negatively (-1). The training mechanism allows the agent to learn how to make the right decisions while performing actions in the environment to maximize the cumulative positive rewards.

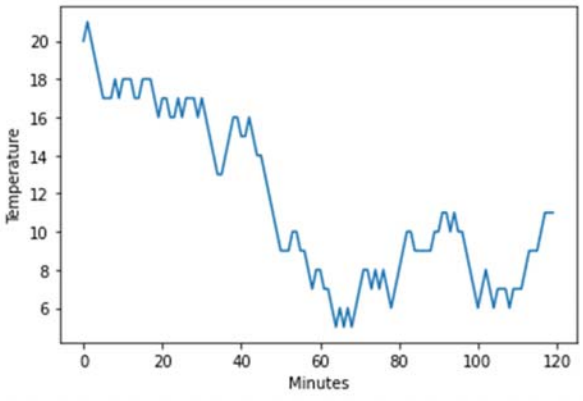

Fig. 4 Environment Response for Random Actions

rollout/	
ep_len_mean	120
ep_rew_mean	-34.7
time/	
fps	449
iterations	2
time_elapsed	9
total_timesteps	4096
train/	
approx_kl	0.008753238
clip_fraction	0.0327
clip_range	0.2
entropy_loss	-1.1
explained_variance	0.00142
learning_rate	0.0003
loss	53.8
n_updates	10
policy_gradient_loss	-0.00135
value_loss	131

(a)

rollout/	
ep_len_mean	120
ep_rew_mean	10.2
time/	
fps	786
iterations	5
time_elapsed	13
total_timesteps	10240
train/	
approx_kl	0.007588242
clip_fraction	0.0397
clip_range	0.2
entropy_loss	-1.04
explained_variance	0.182
learning_rate	0.0003
loss	37.1
n_updates	40
policy_gradient_loss	-0.00982
value_loss	80.4

(b)

rollout/	
ep_len_mean	120
ep_rew_mean	109
time/	
fps	1422
iterations	49
time_elapsed	70
total_timesteps	100352
train/	
approx_kl	0.0041626994
clip_fraction	0.0783
clip_range	0.2
entropy_loss	-0.92
explained_variance	0.00363
learning_rate	0.0003
loss	84
n_updates	480
policy_gradient_loss	0.001
value_loss	170

(c)

Fig. 5 Training Process

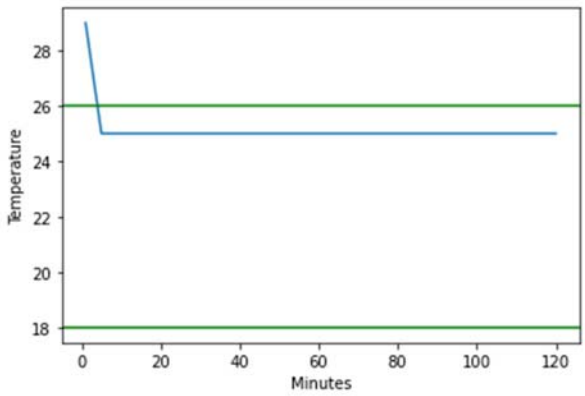

Fig. 6 Performance Test without Disturbance

In Fig. 5 (a), at the early training process, the rewards mean is still negative -34.7. In Fig. 5 (b), after 10240 timesteps, the agent now is getting positive rewards, +10.2. Fig. 5 (c) is the last time steps where the mean reward is +109.

C. Testing and Evaluating

First, the model was tested with no disturbances to make sure that the agent is following the desired performance (Fig. 6).

To ensure that the model is robust and resilient, it was evaluated with external disturbances (change between +1, -1 °C randomly at each time step) starting with low initializations (cold water) (Fig. 7 (a)) and progressing to high initializations (hot water) (Fig. 7 (b)).

VI. Evaluating Safe Performance

SP will track the controller performance by monitoring the rewards received from the environment. For this hydroponic case study, the reward is between -120 and +120. For example, if the temperature falls within the root zone temperature for the whole operation time, the cumulative rewards would be +120. Experimentally, 50% were selected to verify SP functionality with high level of disturbances. So, if the agent is unsuccessful in controlling the environment for 50% of the operation (i.e., it allows the water temperature to fall outside of the root temperature region for any reason out of the controller's ability, such as rough external disturbances or hardware constraints), the controller and the hardware will be turned off for safety concerns.

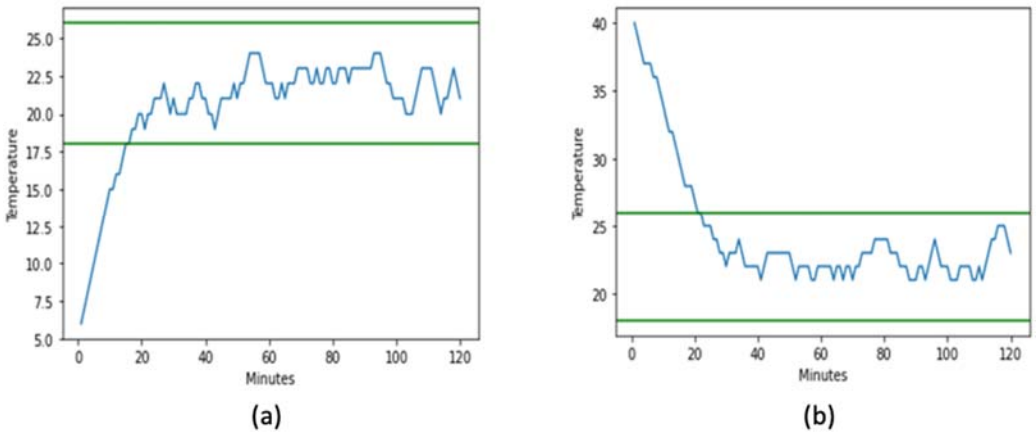

Fig. 7 Performance Test with Disturbances

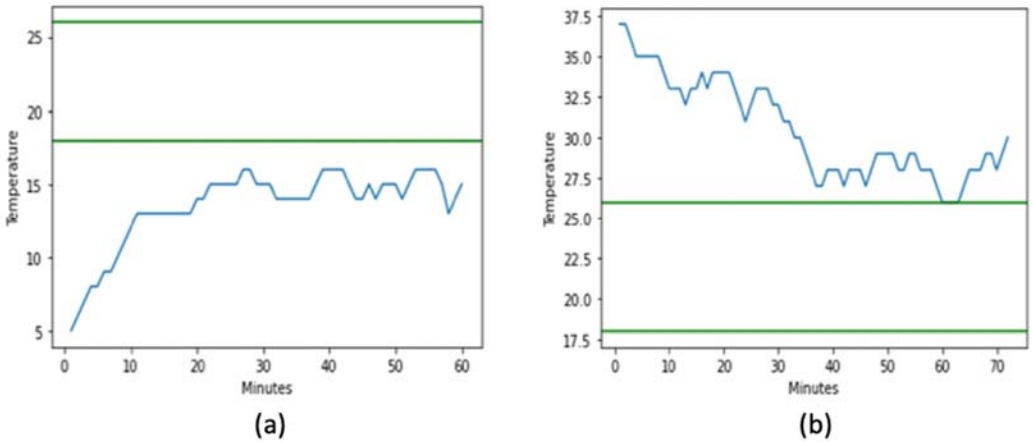

Fig. 8 SP Performance Test

Fig. 8 (a) indicates that the controller could not achieve the desired temperature zone. Thus, after 60 minutes out of 120 minutes with unsuccessful achievements, it turned off. Fig. 8 (b) indicates that the shutdown occurred after 73 minutes because it counted some successful performances between 58 and 65, after which the temperature rose far beyond the desired region. When the controller begins unsuccessfully controlling the temperature due to rough disturbance and/or weak equipment capabilities, the SP plays its role by either turning off the agent (such as in this case study) or sending a notification to the user, designer, or manufacturing company. That is all automatically performed as part of the control system and does not need external devices such as fuses, sensors, or actuators.

VII. Evaluating Safe Energy

SE will let the agent stay in an idle mode automatically when there is no need to let the policy do high computations. For the hydroponic study, when the temperature is inside the root temperature zone, the agent will be in an idle mode. SE was tested without disturbances (Fig. 9) and with disturbances (Fig. 10) beginning with random initializations.

VIII. Average Result

In this study, 1000 episodes (cases) were performed on the proposed RL model. Each episode represents a different case, and each has an initial temperature that was chosen at random between 5 °C and 50 °C with different disturbances between (+1 and -1) and (+2 and -2) °C in each time step (minute). The goal of the model is to stay in the desired root zone region (between 18 °C and 26 °C) while monitoring the agent success/ unsuccess and the energy consumption via the SP and SE. The disturbances manifest themselves as changes in external temperature, hardware constraints, or any type of noise that interferes with controller operation.

Tables I and II represent the average results considering two types of disturbances: (+1, -1) and (+2, -2) °C in each time step (minute), with different unsuccessful metrics (25%, 50%, 75%). The turn off cases are out of the total cases (1000), and the active/idle mode percentage is out of the total operation time for these 1000 cases. Given that the idle mode could consume 50% less energy than normal operation [14], [17], [18], the energy saving is estimated as 50% of the idle mode operation.

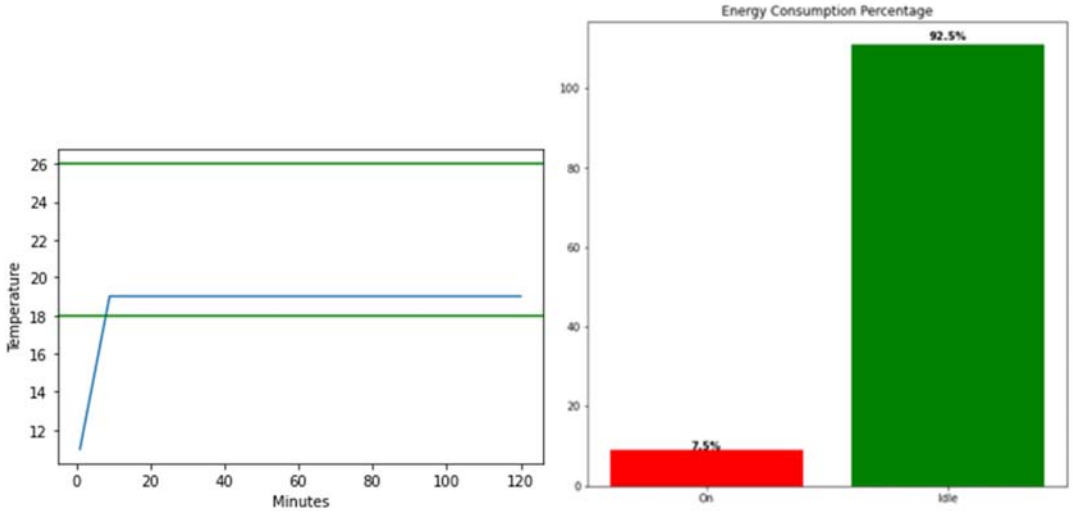

Fig. 9 SE Performance Test without Disturbances

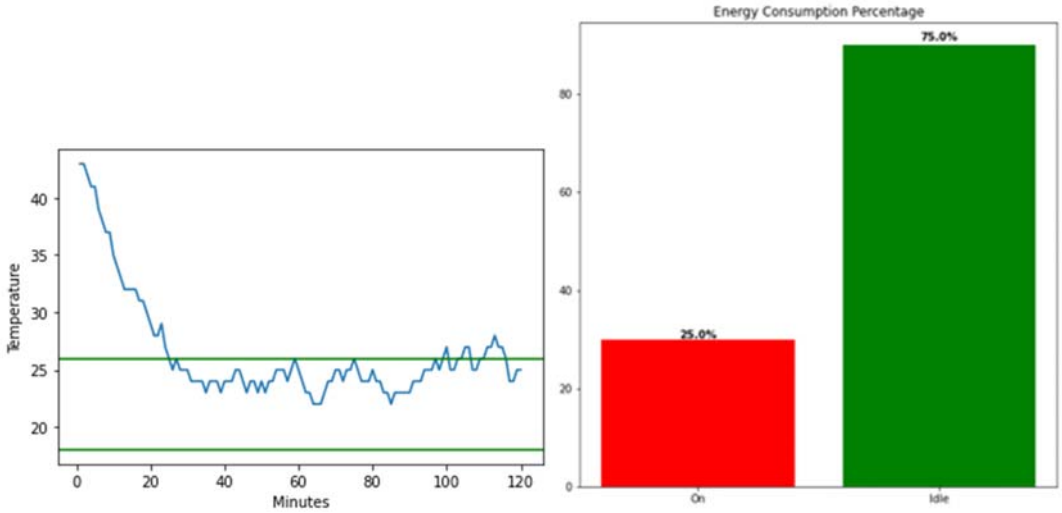

Fig. 10 SE Performance Test with Disturbances

TABLE I
MODEL PERFORMANCE FOR 1,000 CASES WITH DISTURBANCE CHANGE IN TEMPERATURE BETWEEN (+1, -1) EACH TIME STEP

	25%	50%	75%
Turn off cases	148	1	0
Safety improvement	14.8%	0.1%	0%
Active mode	18.9%	17.9%	18.3%
Idle mode	81.1%	82.1%	81.7%
Energy saving	40.5%	41.05%	40.85%

TABLE II
MODEL PERFORMANCE FOR 1,000 CASES WITH DISTURBANCE CHANGE IN TEMPERATURE BETWEEN (+2, -2) EACH TIME STEP

	25%	50%	75%
Turn off cases	156	5	2
Safety improvement	15.6%	0.5%	0.2%
Active mode	28.4%	28.65%	28.2%
Idle mode	71.6%	71.35%	71.8%
Energy saving	35.8%	35.6%	35.9%

IX. CONCLUSION AND FUTURE RESEARCH

This paper represents a RL model for control applications that can be used to improve safety and reduce energy consumption. Temperature control for a hydroponic system was provided to illustrate the model's concept and techniques.

Among average results of 1000 episodes, the model showed up to 15% improved safety and 35%-40% energy saving, depending on the initialization parameters such as the initial temperature, unsuccessful metric percentage, and the roughness level of the disturbance. This model can be adjusted as needed based on the criticality of the application.

Other methods in the literature [6]-[9] to improve safety and energy efficiency have used external devices to the control system like fuses, sensors, or IoT technologies. However, the present approach integrates Safe Performance (SP) and Save Energy (SE) inside the control system, which all operate intelligently in synchronous processes. Therefore, this model reduces the total costs of hardware. Also, since it uses fewer electronic devices, it is more compatible and simpler to

manufacture and integrated into complex systems.

As future research, other approaches for ensuring safety, such as safe RL learning or reference governors, can be added to the model in the future to make it more robust for RL control critical applications. This RL model control can also be applied to any control system application with a desired zone region, such as continuous glucose monitoring for diabetic patients. Furthermore, the model can be applied on real-world physical systems in addition to simulations.

REFERENCES

[1] Bridgewood, L. (2003). Hydroponics: Soilless gardening explained. Ramsbury, Marlborough, Wiltshire: The Crowood Press Limited.

[2] Carotti, Laura, et al. "Plant factories are heating up: Hunting for the best combination of light intensity, air temperature and root-zone temperature in lettuce production." Frontiers in plant science 11 (2021): 592171.

[3] Nguyen, Duyen TP, et al. "Short-term root-zone temperature treatment enhanced the accumulation of secondary metabolites of hydroponic coriander (Coriandrum sativum L.) grown in a plant factory." Agronomy 10.3 (2020): 413.

[4] Zhang, Yong-Ping, et al. "Temperature effects on the reactive oxygen species formation and antioxidant defence in roots of two cucurbit species with contrasting root zone temperature optima." Acta Physiologiae Plantarum 34 (2012): 713-720.

[5] "Best Temperature for Hydroponics." North Slope Chillers, 20 Aug. 2018, northslopechillers.com/blog/best-temperature-for-hydroponics. Accessed 13 Apr. 2023.

[6] Garone, E., Di Cairano, S., & Kolmanovsky, I. (2017). Reference and command governors for systems with constraints: A survey on theory and applications. Automatica, 75, 306-328. https://doi.org/10.1016/j.automatica.2016.08.013.

[7] I. Kolmanovsky, E. Garone and S. Di Cairano, "Reference and command governors: A tutorial on their theory and automotive applications," 2014 American Control Conference, Portland, OR, USA, 2014, pp. 226-241, doi: 10.1109/ACC.2014.6859176.

[8] A Comprehensive Survey on Safe Reinforcement Learning Javier García Fernando Fernández Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganes, Madrid, Spain.

[9] Kotevska, Olivera, et al. Rl-hems: Reinforcement learning based home energy management system for HVAC energy optimization. Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States), 2020.

[10] Wiering, Marco A., and Martijn Van Otterlo. "Reinforcement learning." Adaptation, learning, and optimization 12.3 (2012): 729.

[11] Huang, Jing-Wen & Gao, Jia-Wen. (2020). How could data integrate with control? A review on data-based control strategy. International Journal of Dynamics and Control. 8. 1-11. 10.1007/s40435-020-00688-x.

[12] Wang, Qiang, and Zhongli Zhan. "Reinforcement Learning Model, Algorithms and Its Application." 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC) (2011): 1143-146. Web.

[13] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. ArXiv. https://doi.org/10.48550/arXiv.1707.06347.

[14] J. Segawa, Y. Shirota, K. Fujisaki, T. Kimura and T. Kanai, "Aggressive use of Deep Sleep mode in low power embedded systems," 2014 IEEE COOL Chips XVII, Yokohama, Japan, 2014, pp. 1-3, doi: 10.1109/CoolChips.2014.6842956.

[15] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. ArXiv. /abs/1606.01540.

[16] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. The Journal of Machine Learning Research, 22(1), 12348-12355.

[17] Pedram, Massoud. "Power optimization and management in embedded systems." Proceedings of the 2001 Asia and South Pacific Design Automation Conference. 2001.

[18] Homayoun, Houman, Mohammad Makhzan, and Alex Veidenbaum. "Multiple sleep mode leakage control for cache peripheral circuits in embedded processors." Proceedings of the 2008 international conference on Compilers, architectures and synthesis for embedded systems. 2008.