Comparison of Reinforcement Learning Methods to Optimize Net Profit in Lettuce Greenhouse Production

Group 2: Logan Rower, Eric Wiskandt

WAGENINGEN
UNIVERSITY & RESEARCH

Contents

Introduction

By 2050 it is expected that the global demand for food will increase by 70 %. This demand for food production stresses current conventional agricultural systems (1). Incorporating agricultural systems that utilize novel technological solutions can not only increase overall productivity, but also reduce the consumption of the heavily utilized resources in agriculture, such as land, water, and energy. Controlled Environment Agriculture (CEA) systems attempt to optimize agricultural processes through both control and automation. Greenhouses are one such system that can utilize the natural solar radiation for aiding photosynthesis in the plants, indoor temperature regulation, while also being able to incorporate additional heating, ventilation and CO_2 when needed. However, the control of these systems is complex and requires highly trained individuals to properly execute these processes to maintain optimal environmental conditions (2). Model-based approaches for predicting future disturbances to optimize the system dynamics have been used for greenhouse control but are not robust for capturing the complex dynamics of the system.

A crop suitable for such systems is lettuce, which has a shorter crop cycle from 40-60 days and can be implemented in soil or soilless systems seamlessly. Incorporating lettuce into an automated greenhouse system enables a reduction in susceptibility to disease, improved quality, and higher yields (3&4). However, achieving such results requires the control mechanism of the greenhouse adhering to the optimal thresholds of environmental parameters for lettuce. The optimal temperature for the cultivation of lettuce ranges from 15.5 C to 28 C during the day, and 3 C to 12 C during the night (3). Ambient CO_2 levels are 400 ppm, so maintaining a level between 800 ppm and 1000 ppm would be most ideal to avoid causing any form of damage to the plants while also leading to an increase in dry weight (5). Finally, it was also found in a study that lettuce plants under 85% relative humidity had the highest dry weight (6).

To capture these complex dynamics and be able to deal with the random behavior attributed to the environment, Deep Reinforcement Learning can be used to address this. This method of machine learning enables the agent to learn from its experiences, while being rewarded or penalized for its actions. These methods of control involve time intensive training in a simulated environment to ensure that a model is ready to be deployed into the real-world environment. However, it is important to note that due to the gaps in the simulation the model won't necessarily translate directly into the real-world scenario (7). In addition to this, a simulation environment for a single greenhouse would not necessarily translate to other greenhouses and because the model is trained on an individual environment it would not be able to be applied outside of its current scope. There are simpler reinforcement learning methods such as deterministic policy control. This is essentially where the agent's state results in a specific action. This policy allows for greater interpretability but less overall control of the environment.

This study examined the utilization of reinforcement learning to optimize the net profit of lettuce production in an automated greenhouse environment in the Netherlands. It is important to note that like most reinforcement learning algorithms that are to be implemented in real world scenarios the model that was created was tested and trained on a simulated greenhouse environment. The simulated greenhouse environment was based on the greenhouse model developed by Morcego et al. (8). Both a deterministic and deep reinforcement learning method were implemented. The DRL method that was examined is Proximal Policy Optimization (PPO). Two reward functions were tested for the PPO algorithm, one utilizing the net profit (Eq. 1), and a second that incorporated penalties and thresholds for the control and environmental parameters (Eq. 2). The optimal model was then trained on a growing period of 40 days and tested to determine the net profit.

Materials & Methods

Greenhouse Simulation Model

The basis of the net profit optimization problem is a greenhouse model by Boersma, Sun and van Mourik (8). This greenhouse model works with 4 state variables which are: lettuce dry weight $\left(\frac{kg}{m^2}\right)$, the indoor CO_2 concentration $\left(\frac{kg}{m^3}\right)$, indoor air temperature (°C) and the indoor humidity $\left(\frac{kg}{m^3}\right)$. Additionally, the indoor model uses weather data from a Dutch greenhouse in Bleiswijk collected in 2014 and published by Kempkes et. al in the same year (9). The weather data provides four disturbance variables that are added to the state model during calculation of the next state. These variables are the outdoor radiation $\left(\frac{W}{m^2}\right)$, CO_2 concentration of the outside air $\left(\frac{kg}{m^3}\right)$, the outside air temperature (°C) and outdoor humidity $\left(\frac{kg}{m^3}\right)$. The environment is set up as a gym environment. The Gymnasium package is a popular way of training reinforcement algorithms. It supplies some basic environments but also allows the construction of custom environments. The reinforcement algorithm is drawn from the Stable Baselines 3 library. For this project version 1.5 of Stable Baselines 3, and 0.21 of Gymnasium were used. The model can be initialized with many parameters. Some of the parameters that were utilized are the start day of the year and the length of the simulation period.

In setting up this greenhouse environment several assumptions were made with respect to data collection and simulation settings. The starting day was set to day 150, and the simulation length was set to 7 days. The late start day was needed to train the algorithm on the most challenging part of the weather data, the summer. The high radiation during summer proved to be the biggest influencing factor. The simulation length of 7 days was chosen to cut down the simulation time. On 7 days the simulation takes around 150 minutes, in comparison to that the 40-day period for a full growing cycle takes more than 180-minutes to compute. The simulation considers the time for training the policy model as well as the time for the evaluation. Training and testing are repeated 5 times to reduce the impact of the random initialization in the PPO algorithms policy. Each test runs through 100 simulation episodes and records the mean net profit of all the runs in one test.

Proximal Policy Optimization

For this project Proximal Policy Optimization (PPO) was chosen as reinforcement algorithm. PPO is a model free policy gradient method developed by OpenAI in 2017. This means that the algorithm doesn't need a model of the environment. It can, however, still operate with models, as has been done in this project. Policy gradient methods like PPO map the finite dimensional parameter space to the policy space. After doing that the algorithm works with gradient ascend to tune the parameters that lead to an optimal policy. During this convergence PPO applies a novel objective function called *clipped surrogate objective function*. This clipping function constrains the possible policy update. As a result, the policy updates are smaller than in other comparable policy gradient methods. The clipped updates reduce the chance of so-called catastrophic forgetting which is induced by large policy updates that wipe out critical parts of the policy. Smaller policy updates lead to the new policy being in close proximity to the old one and this is where the method has its name from (10).

Reward Functions

When training the PPO policy model, the reward function of the model is called which needed to be implemented in the scope of this project. The first version of the reward function focused on the net revenue gained through the harvest of lettuce, while considering the cost for heating, CO_2 supply and ventilation. This reward function delivered unsatisfactory results when trying to tune it. After optimizing this reward function the optimal solution was to take no action on the control inputs and in this way minimize cost. The dry weight still increased for most of the year, except for a few days in summer and in winter when

temperatures were too extreme, and the plant died. In fact, the first reward function did optimize more on the costs than on the profit, as the dry weight influencing factors were not included in the reward structure.

Eq. 1: Net Profit Reward Equation using parameters from Table 1

$$Cost\ CO_2\left[\frac{€}{m^2}\right] = CostCO_2\left[\frac{€}{kg}\right] \cdot Supply\ Rate\ CO_2\left[\frac{mg}{m^2 \cdot s}\right] \cdot timestep$$

$$Cost\ Ventilation\left[\frac{€}{m^2}\right] = Cost\ Energy\left[\frac{€}{J}\right] \cdot vent\_cap\left[\frac{J}{m^3 \cdot °C}\right] \cdot vent\_rate\left[\frac{mm}{s}\right] \cdot IndoorAirTemp[°C] \cdot timestep \cdot \frac{m}{mm}$$

$$Cost\ of\ Heating\left[\frac{€}{m^2}\right] = Cost\ Energy\left[\frac{€}{J}\right] \cdot Energy\ Supplied\ by\ Heating\left[\frac{J}{s \cdot m^2}\right] \cdot timestep\ [s]$$

$$Revenue\left[\frac{€}{m^2}\right] = \left(dryweightNEW\left[\frac{kg}{m^2}\right] - dryweightOLD\left[\frac{kg}{m^2}\right]\right) \cdot LettucePrice\left[\frac{€}{kg}\right]$$

$$Net\ Profit\left[\frac{€}{m^2}\right] = Revenue - (Cost\ of\ CO_2 + Cost\ Ventilation + Cost\ Heating)$$

For that reason, it was decided to change the reward function to focus more on the growing parameters like lettuce dry weight, and CO_2 supply, and put cost factors on the control inputs themselves. The new reward function was based on a previous reward function from a paper by Morcego et. al. (8). This reward function includes the change in dry weight per timestep, the value of CO_2 in the indoor air in comparison with a lower and upper threshold, a similar setup for the temperature and a punishment for the use of the control input, as these result in costs. The threshold values for the CO_2 were provided in ppt and ranged from 0.4 to 0.9 ppt (400ppm to 900ppm). The optimal value range for temperature was set to be from 5 to 30 °C. Both thresholded regions are relatively equivalent to the literature ranges previously discussed (3&5). The original reward function from the paper used constant weight factors for all these different parts. These weight factors were changed during tuning to optimize the profit.

Eq. 2: Penalized Reward Function as in Morcego et al. using parameters from Table 2 (8).

$$r_{CO_2}(k) = \begin{cases} -c_{r,CO_2,1} \cdot (y_2(k) - CO_{2_{min}}(k))^2 & \text{if } y_2(k) < CO_{2_{min}}(k) \\ -c_{r,CO_2,1} \cdot (y_2(k) - CO_{2_{max}}(k))^2 & \text{if } y_2(k) > CO_{2_{max}}(k) \\ c_{r,CO_2,2} & \text{otherwise} \end{cases}$$

$$r_T(k) = \begin{cases} -c_{r,T,1} \cdot (y_3(k) - T_{min}(k))^2 & \text{if } y_3(k) < T_{min}(k) \\ -c_{r,T,1} \cdot (y_3(k) - T_{max}(k))^2 & \text{if } y_3(k) > T_{max}(k) \\ c_{r,T,2} & \text{otherwise,} \end{cases}$$

$$r(k) = c_{r,1}\Delta_{y_1}(k) + r_{CO_2}(k) + r_T(k) - \left(\sum_{j=1}^{3} c_{r,u_j} \cdot u_j(k-1)\right)$$

Deterministic Model

As a point of reference, a deterministic policy was used to calculate the net profit. This deterministic policy was designed as a stepwise threshold controller, based on a lower and upper bound for indoor temperature. This threshold for temperature was set to 5 to 30 °C. If the indoor temperature was not within these bounds, then the ventilation and heating are stepwise increased or decreased accordingly. The CO_2 supply was set to a constant predetermined value of 0 $\frac{mg}{m^2 s}$ that kept the CO_2 level in a range that did not influence the plant growth negatively due to the high ventilation ate during the specific simulation frame. This deterministic

controller was not optimized and just keeps the indoor values in bounds that hinder plant death by overheating or CO_2 undersupply. The deterministic policy was initialized for starting the simulation on day 150, for a range of 7 days.

Results

The deterministic policy performed as expected during the warm weather period with no energy supplied due to high solar radiation. In addition to this the ventilation was turned on for the entirety of this period, indicating again this controller was behaving as expected for this warm period. The CO_2 concentrations are displayed in parts per thousand (ppt), but even if converted to ppm turned out to be 1ppm, the CO_2 concentrations for the deterministic policy are extremely low even with the relatively continuous ventilation (Figure 1). There is a clear diurnal cycle that can be seen with the temperature, and CO_2 concentrations indoors, so although the controller is not optimized for the model for these points in time, it is still able to capture the natural cycle. An intriguing observation was that even with venting constantly that the CO_2 concentrations were not higher especially during the night (Figure 1). If the vents are open and the plants are not taking up any CO_2, which is what occurs during the night, the CO_2 concentration would have been expected to have reached at least 4.0 ppt (400 ppm). The overall net profit over the 7-day period for this policy was found to be 24.8 $\frac{\text{cents}}{m^2}$. This indicated that for a 1-acre greenhouse the net profit would be €52,188/year if this 7-day period was extrapolated over the 52 weeks of the year. Although similar profit might not be expected for the winter month.

The PPO algorithm was then implemented first with the net profit reward function (Eq. 1). This reward function proved to be extremely faulty and led to errors within the environment when running over many episodes. In addition to this it ended up with a mean net profit of -inf over the 100 episodes using the trained PPO algorithm starting at 150 and over the duration of a week. Due to this the plots were not further examined as it was assumed there were errors in their production, so the next reward function (Eq. 2) was tested to see if it performed better than the deterministic policy.

The second reward function based on Morcego et al. was then tested (8). The initial constant values detailed in this publication were used to check the initial performance for the PPO algorithm (Table 2). Figure 2 showcases the average state and action values over 100 episodes for a duration of 7 days, with each episode resetting the start day of the simulation to another random day. This random sampling of the data was done to ensure that the trained reinforcement learning model would behave appropriately. In Figure 2, the actions performed by the agent by adjusting the control variables of energy supplied to heat, the ventilation rate, and the CO_2 supplied are very twitchy. These three plots showcase a limited form of constancy that is related to the day and night cycle. Besides the turbulent nature of the control parameters, the indoor CO_2 levels were within the optimal threshold for plant production from 0.4 ppt (400ppm) to 0.80 ppt (800ppm). However, the indoor temperature exceeds the threshold set of 30 °C, multiple times going above 35 ℃. It is important to note that despite the indoor CO_2 concentrations being so low and the possibly dangerously high temperatures, the growth was relatively constant over the period producing roughly 80 $\frac{\text{g}}{m^2}$. The production and control efficiency led to a net profit of 17.65 $\frac{\text{cents}}{m^2}$, which when extrapolated to a 52-week period and for a 1-acre greenhouse the net profit would be €37,134.19/year.

After performing multiple tests of the manually tuned constant values, the values from run 4 were selected as the most optimal (Table 3). The performance of this test shows greater constancy among the control variables, with heating being applied minimally, and a diurnal cycle that can be seen for the ventilation with it being on during the day, but then it reduces during the night (Figure 3). The CO_2 supplied also exhibited a similar pattern with more CO_2 being supplied during the day, to aid in the

plants' photosynthetic growth. During the night the CO_2 supplied was reduced due to the low ventilation rate and a high indoor concentration of CO_2. The indoor CO_2 concentration reached peaks of up to 1.4 ppt (1400ppm), far exceeded the set upper bound of .9 ppt but still within reasonable range for continued plant growth (5, Figure 3). This additional CO_2 allowed for a greater production of biomass as over 100 $\frac{g}{m^2}$ of dry weight was produced. Finally, the temperature within the greenhouse remained within the set threshold from 5 C to 30 C. The overall net profit of the greenhouse with the reward function and system constraints was 38.89 $\frac{\text{cents}}{m^2}$. This indicates the largest profit of the tested methods with a net profit for a 1-acre greenhouse of €81,821.45/year.

A final model on the best algorithm or policy was performed over a 40-day period. The constants from Run 4 were utilized for this model to run over this 40-day period to simulate an actual cultivation period using PPO (Table 3). Upon testing this model, the average cumulative reward increased over time but then in 2000 timesteps it's reward started to rapidly decrease. Although the other reward plots were not included within the contents of this study, for the initial three tests the reward plots were steadily increasing overtime with positive reward contrary to this model's reward (Figure 4). In Figure 5 there is a plot of average action and state values using the optimal reward constants. There is a clear diurnal cycle for the supply of CO_2 and the ventilation rate, but no energy supplied for heating over the examined periods (Figure 5). The indoor CO_2 concentration also appears to maintain a level that is within the range of 0.2 ppt (200 ppm) and 1 ppt (1000ppm). The indoor concentration goes below the threshold that has been set at 0.4 ppt (400ppm) (Figure 5). This is likely attributed to the level of biomass production and uptake of CO_2 by the lettuce plants, resulting in upwards of 250 $\frac{g}{m^2}$. of dry weight of lettuce. An big issue with this model is that the temperature exceeds the acceptable threshold of 30 C reaching temperatures close to 50 C (Figure 5). Having this high of temperatures over this period should have led to a decrease in production as above 40 C is far outside of the optimal threshold of lettuce (3). Despite having this extremely hot environment the environment was relatively productive and resulted in a mean net profit of 71.76 $\frac{\text{cents}}{m^2}$., and over a period of 1 year in a 1-acre greenhouse €26,493.61/year.

Discussion and Conclusion

It was assumed that the PPO model would perform the best with a model based strictly on net profit and utilizing the change in profit to aid in the learning of the algorithm. A net profit reward function was developed without any additional penalties for this purpose (Eq. 1). Due to the lack of additional penalties on the control a second reward function was formulated focusing on the growing process rather than the net profit (Eq. 2). In selecting models an important feature to consider is not only the model performance but also its efficiency. In the end the deterministic model was efficient in taking roughly 5 seconds to achieve results for the 7-day period. While running the PPO algorithm for the 7-day period it took roughly 150 minutes in comparison to the 40-day period which took 180 minutes. So, the efficiency of the deterministic model is a factor that would need to be weighed with respect to reinforcement learning algorithms. It is important to note however, that once the PPO algorithm had been trained its testing was quite fast, taking only a few seconds to complete for both the 7 and 40-day periods.

It was clear that of the 3 original models over 7-day periods, the best model was the PPO algorithm using the second reward function with the tuned constraints (Eq 2, Table 3). This model was found to have a 56.83% increase in profit compared to the deterministic model. However, upon training and testing a PPO algorithm over a 40-day simulated cultivation period it was found that the overall net profit in a year was significantly less than the extrapolated amount achieved from the same model but over the 7-day period, with a difference of roughly €55,000. The biggest difference in these two models is that the resulting

temperature plots of the period showcased more extreme temperatures for the 40-day than the 7-day that would have indicated greater loss in biomass due to damage to the lettuce plants (Figure 3 & 5). In addition to this the reward for the 40-day model, although it was the best selected model for this period, seemed to suffer from catastrophic forgetting, indicated by the steep decline in the reward. Catastrophic forgetting is something typical of neural networks where upon learning new information while training it forgets the previously learned information. The basis of the PPO algorithm policy that was used in this study is a multi-layered perceptron, a type of neural network. Some solutions to deal with this potential issue is to stop training as early as possible, and normalize the state and action space, which avoids any issues with units and scaling (11). Upon noticing this issue initially, the training timesteps were significantly reduced, but no normalization was applied to the state space, though the action space was normalized.

It is important to note that there is likely bias in the results achieved. This is due to the random start date, at each new episode during the testing of the PPO model. A set random seed was applied to allow for reproducibility of results, which led to the same series of random start days being used for each model. This testing procedure was only run on a single random seed, and it is possible that because of this there is still little generalization. This is because the sampled data and many of the test dates are within a similar warm period. However, when looking at the results for the 40-day simulation, the results are worse, this is likely because there is a greater number of colder days included within the extended simulation range. To improve future results and to validate the results that were achieved more test runs will need to be conducted on the trained model using different random seeds to determine the magnitude of variability in results.

Another problem was the limited ability to cool the greenhouse down. The only available option is to increase the ventilation which leads to a problem when the simulated outside air is close to or above the desired inside temperature. This temperature dilemma was the reason why day 150 was chosen as the start day for the outlined experiments. The period following day 150 is one of the warmest periods in the weather data set. As shown in the results, the deterministic policy and PPO policy reacted with increased ventilation but even with full ventilation it was not possible to keep the indoor temperature below the desired values. The addition of a dedicated cooling system, like water dispersion, into the model could help reduce the temperature more effectively.

Some additional limitations to the PPO reinforcement learning algorithm were that it was an on-policy algorithm. Although this is an algorithm that allows for greater stability, it is quite resistant to hyperparameter tuning, and has reduced change from its previous policy. Testing a different off-policy method that would have exhibited greater exploration to compare with the PPO results could have also provided greater insight not only into the performance of PPO but also whether for applications such as this if on-policy algorithm would be preferred.

Addendum:

Upon completion of this work, additional testing was performed using the net profit reward function, determined that an initial error in the equation that was subsequently fixed was not evaluated for further testing of the reward function before moving to the next one. The results of the net profit reward function showcase the second highest mean net profit for the 7-day period of 35.34 cents/m^2 evaluating to 74,352.53 euro/year for a 1-acre greenhouse. Additional comparisons between both reward functions utilized in this report can be examined further using different reinforcement learning algorithms.

References:

1. Food and Agriculture Organization. (2009). In How to Feed the World in 2050. Rome.
2. Graamans, L., Baeza, E., van den Dobbelsteen, A., Tsafaras, I., & Stanghellini, C. (2018). Plant factories versus greenhouses: Comparison of resource use efficiency. Agricultural Systems, 160, 31–43. https://doi.org/10.1016/j.agsy.2017.11.003
3. Petropoulou, A. S., van Marrewijk, B., de Zwart, F., Elings, A., Bijlaard, M., van Daalen, T., Jansen, G., & Hemming, S. (2023). Lettuce production in intelligent greenhouses—3d imaging and computer vision for plant spacing decisions. Sensors, 23(6), 2929. https://doi.org/10.3390/s23062929
4. Van Henten, E. J., & Bontsema, J. (2009). Time-scale decomposition of an optimal control problem in Greenhouse Climate Management. Control Engineering Practice, 17(1), 88–96. https://doi.org/10.1016/j.conengprac.2008.05.008
5. Greenhouse Carbon Dioxide Supplementation - Oklahoma State University. Greenhouse Carbon Dioxide Supplementation | Oklahoma State University. (2017, March 1). https://extension.okstate.edu/fact-sheets/greenhouse-carbon-dioxide-supplementation.html
6. Chia, S. Y., & Lim, M. W. (2022). A critical review on the influence of humidity for plant growth forecasting. IOP Conference Series: Materials Science and Engineering, 1257(1), 012001. https://doi.org/10.1088/1757-899x/1257/1/012001
7. van den Bemd, W. J. G. M. (2022). Robust Deep Reinforcement Learning for Greenhouse Control and Crop Yield Optimization (thesis).
8. Morcego, B., Yin, W., Boersma, S., van Henten, E., Puig, V., & Sun, C. (2023). Reinforcement Learning Versus Model Predictive Control on Greenhouse Climate Control. arXiv preprint arXiv:2303.06110.
9. Kempkes, F. L. K., Janse, J., & Hemming, S. (2013, October). Greenhouse concept with high insulating double glass with coatings and new climate control strategies; from design to results from tomato experiments. In International Symposium on New Technologies for Environment Control, Energy-Saving and Crop Production in Greenhouse and Plant 1037 (pp. 83-92).
10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
11. Cahill, A. C. (2010). Catastrophic forgetting in reinforcement-learning environments (thesis).

Appendix:

Table 1: Equation 1 Parameters:

Parameter	Value
$CostCO_2\left[\frac{euro}{kg}\right]$	$\frac{42e^{-2}}{2.20371}$
$Cost\ Energy\left[\frac{euro}{J}\right]$	$\frac{6.35e^{-9}}{2.20371}$
$vent\_cap\left[\frac{J}{m^3 \cdot °C}\right]$	1290
$LettucePrice\left[\frac{€}{kg}\right]$	$\frac{16}{2.20371}$
$timestep\ (s)$	900

Table 2: Equation 2 Variables [8]

Parameter	Value	Parameter Name	Value
$c_{r,1}$	16	$c_{r,CO_2,1}$	1
c_{r,u_1}	-4.536×10^{-4}	$c_{r,CO_2,2}$	.0005
c_{r,u_1}	$-.0075$	$c_{r,T,1}$	.001
c_{r,u_3}	-8.5725×10^{-4}	$c_{r,T,1}$	.0005

Table 3: Results of five iterations of tuning parameters for the PPO model using the second reward function.

Run	$c_{r,1}$	$c_{r,T,1}$	c_{r,u_3}	Net Profit
1 (Original Parameters)	16.0000	0.0010	-0.0009	17.6518
2	8.0000	0.0150	-0.0943	34.9299
3	4.0000	0.0150	-0.0943	38.6719
4 (Optimal Parameters)	1.0000	0.0150	-0.0943	38.8941
5	1.0000	0.0200	-0.0943	29.6473

Figure 1: 7-day simulation starting on day 150 using deterministic policy.

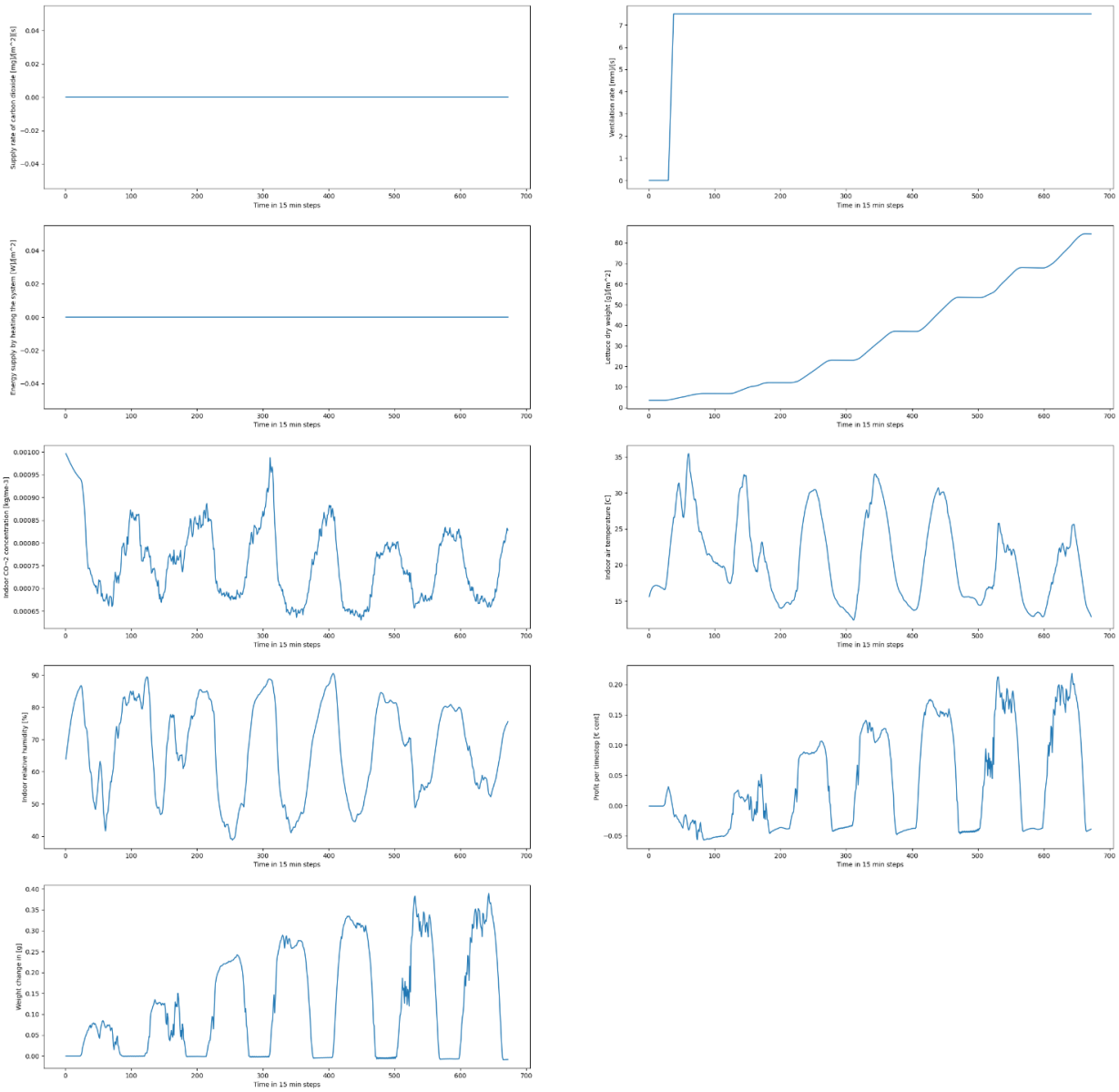

Figure 2: Plot of average action and state values from literature-based reward constants with PPO model

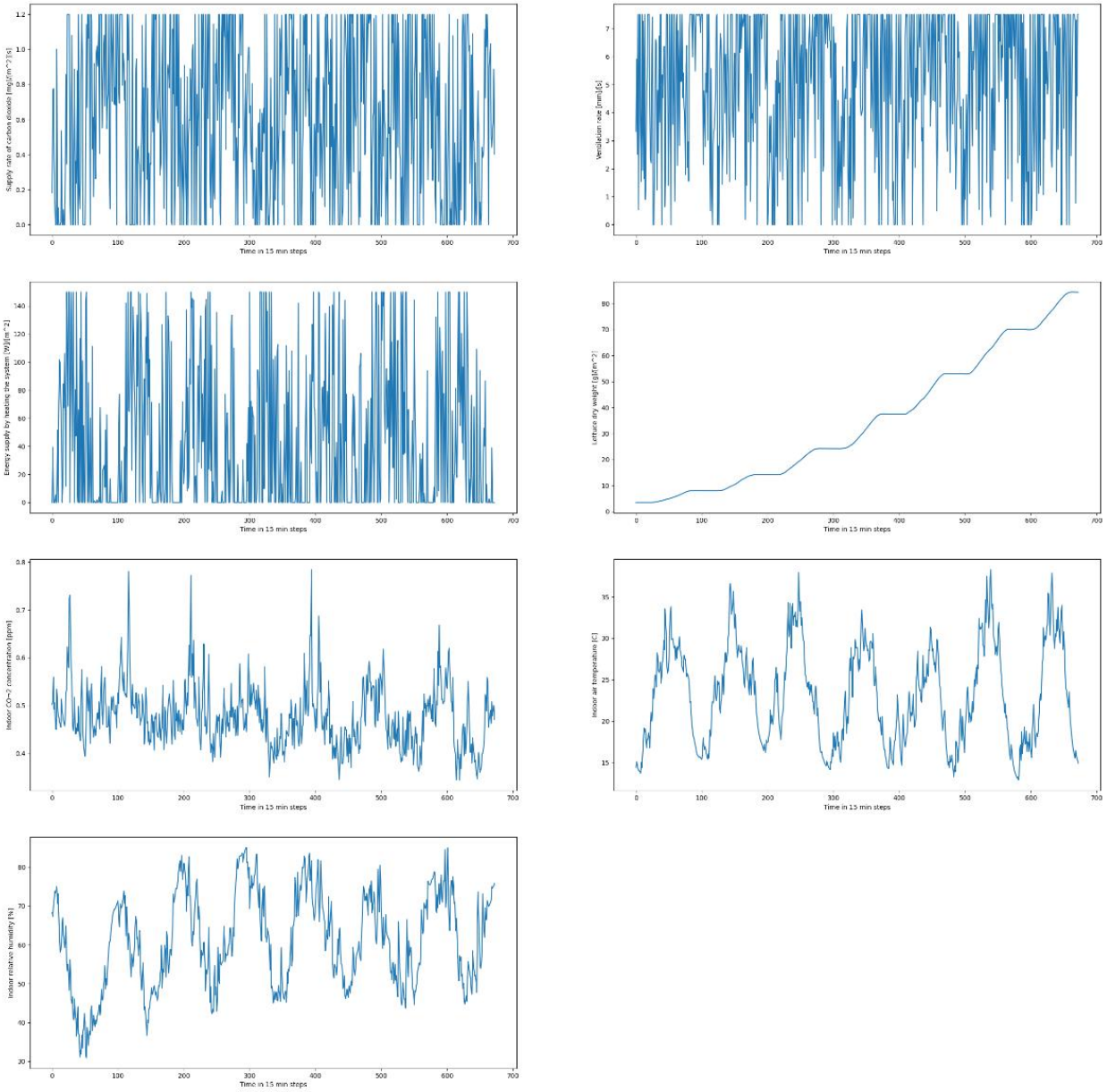

Figure 3: Plot of average action and state values from Optimal reward constants with PPO model (7 day)

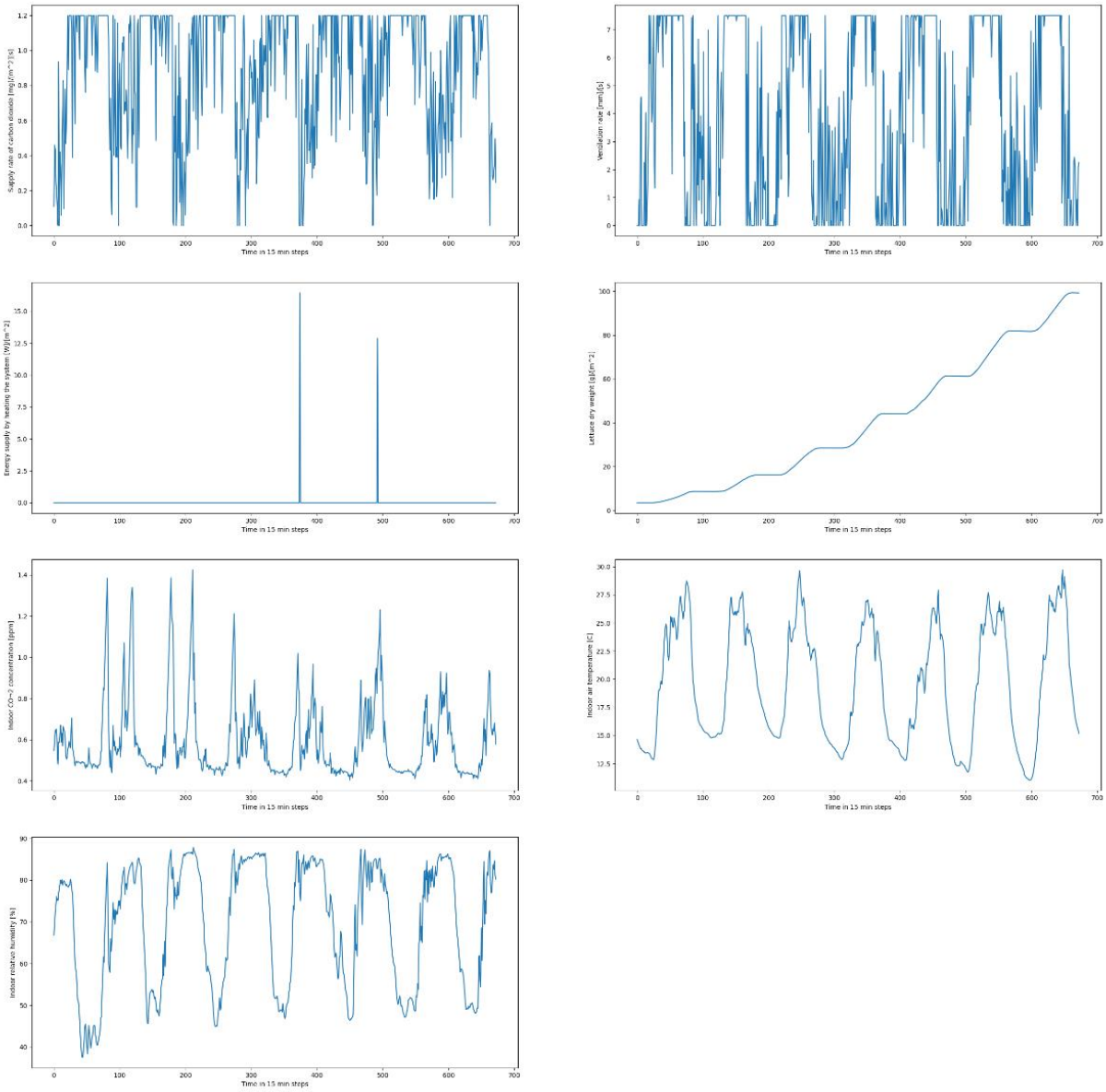

Figure 4: Plot of the average cumulative reward over 100 episodes optimal constants (40 days).

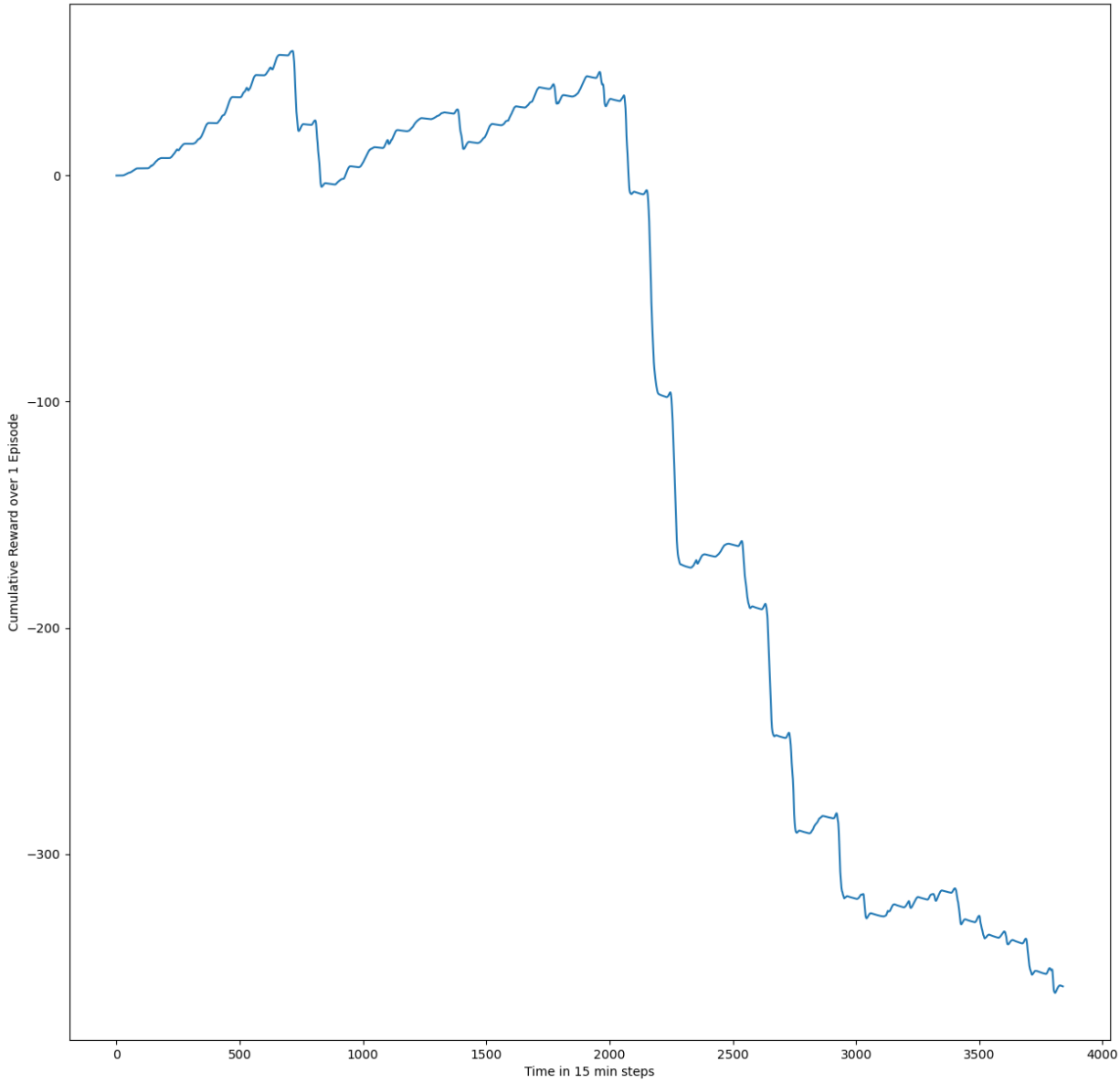

Figure 5: Plot of average action and state values from Optimal reward constants with PPO model (40 day)