Optimizing Nitrogen Management with Deep Reinforcement Learning and Crop Simulations

Jing Wu*   Ran Tao*   Pan Zhao*   Nicolas F. Martin   Naira Hovakimyan

University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

{jingwu6, rant3, panzhao2, nfmartin, nhovakim}@illinois.edu

Abstract

Nitrogen (N) management is critical to sustain soil fertility and crop production while minimizing the negative environmental impact, but is challenging to optimize. This paper proposes an intelligent N management system using deep reinforcement learning (RL) and crop simulations with Decision Support System for Agrotechnology Transfer (DSSAT). We first formulate the N management problem as an RL problem. We then train management policies with deep Q-network and soft actor-critic algorithms, and the Gym-DSSAT interface that allows for daily interactions between the simulated crop environment and RL agents. According to the experiments on the maize crop in both Iowa and Florida in the US, our RL-trained policies outperform previous empirical methods by achieving higher or similar yield while using less fertilizers.

Figure 1. A framework for optimizing N management with deep RL and DSSAT-based crop simulations

1. Introduction

The agricultural industry is facing significant challenges to meet the food demand to feed more than nine billion people by 2050 [23]. The challenges are further complicated by the current context of diminishing land and water resources, degraded soil and changing climate. The world urgently needs to move towards more sustainable and resilient cropping systems [24]. Among different factors influencing crop production and the environment, *nitrogen (N) management* is a key controllable one. Nitrogen is the main nutrient affecting crop growth and yield formation, but excessive nitrogen has substantial negative environmental effects [28]. Effective nitrogen management is therefore crucial for maximizing crop yields and farmer income and minimizing negative environmental impacts. Although best-practice knowledge for N management for common scenarios exists among farmers, it is unclear whether these practices are near-optimal, or whether some specific strate-
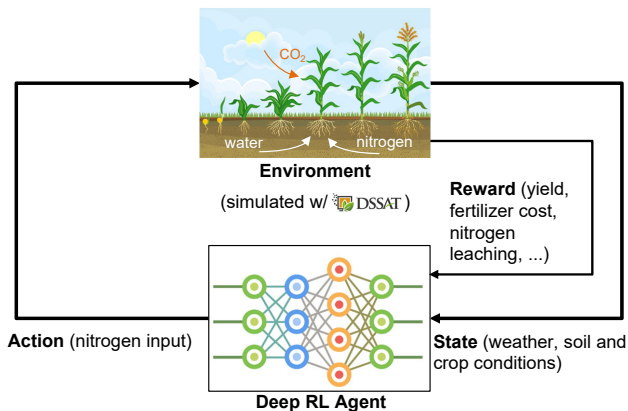
gies transfer well to adverse seasonal conditions of extreme temperature or precipitation. N management is essentially a sequential decision making (SDM) problem as a few decisions on nitrogen application time and quantities need to be made across the growth cycle of crops. Modern reinforcement learning (RL) methods, represented by deep RL, have achieved remarkable or superhuman performance on a variety of tasks involving SDM such as gaming [18, 30], data center cooling [6], and robotic control [11, 15, 26]. We expect that RL has a potential for optimizing agricultural management, improving the crop yield while minimizing the environmental impacts. Training a deep RL policy often needs numerous interactions between the RL agent and the environment, which makes it unrealistic to leverage field trial-based approaches [1]. Therefore, training the management policies in simulations [3, 4, 20], using crop models to simulate the crop and soil dynamics and interact with the RL agent, seems the only realistic solution.

In this paper, we *propose and evaluate a framework for optimizing N management using deep RL and crop simulations*, depicted in Fig. 1. In particular, we leverage Decision

---

*These authors contributed equally.

Support System for Agrotechnology Transfer (DSSAT), a widely used tool for crop modeling and simulation [10, 14], and the Gym-DSSAT interface [8] that allows users to read the simulated crop and soil conditions and apply management practices on a daily basis. As a demonstration of the use of the presented framework, we train N management policies with two deep RL algorithms, namely deep Q-network (DQN) and soft actor-critic (SAC), for the maize crop in Iowa and Florida, US. We further evaluate the performance of the trained policies in comparison with standard practices, and under different scenarios including partial observations and reduced action frequencies.

Compared to early work on RL-based crop management [7, 27], our framework, which leverages deep RL, can handle much larger state and action spaces. Compared to recent work on deep RL-based agricultural management [2, 19], the crop model adopted in our framework, i.e., DSSAT, is much more widely used globally; additionally, our experimental study is significantly more comprehensive, which involves two different deep RL algorithms, two geographic locations, and ablation study for partial observations and reduced action frequencies.

2. Related Work

2.1. Reinforcement Learning in Agricultural Management

Reinforcement learning, as a sub-field of machine learning, aims to solve SDM problems by letting an agent directly interact with the environment and learn from trial and error [29]. As a pioneering work, [7] proposed to use a simple RL method (namely, R-learning) and crop simulations to optimize management of wheat crops in France. [27] studied the use of SARSA(λ), an on-policy RL method, and crop simulations to optimize the irrigation for the maize crop in Texas, US. However, the state and action spaces in [7, 27] were quite small due to the curse of dimensionality from which early RL methods suffered. For instance, the state space in [27] has only one state, i.e., total soil water (TSW) level. In contrast, modern RL methods, represented by deep RL, are able to handle extremely large state and action spaces due to the use of deep neural networks (DNNs) (to approximate the value functions or policies), and have achieved remarkable or superhuman performance on a variety of high-dimensional problems such as gaming [18, 30], data center cooling [6], and robotic control [11, 15, 26]. Deep RL based on the proximal policy optimization (PPO) algorithm was used in [19] to optimize the fertilizer management for the wheat crop. Additionally, [2] studied the use of PPO to optimize the irrigation management for russet potatoes. However, the study is quite coarse and the results are not promising. For instance, in terms of results, [2] included only a simple learning curve showing the normal-

ized reward, while the variables farmers mostly care such as yield and management cost were not included. Additionally, the trained policy performed much worse than a simple policy which applies a constant amount of water.

2.2. Crop Models

Crop models can simulate crop growth in response to soil, water, nutrient, and weather dynamics. They are playing increasingly important roles in the development of sustainable agricultural management, because field and farm experiments require large amounts of resources and may still not provide sufficient information in space and time to identify appropriate and effective management practices [13]. The development of crop models dated back to 1950s. In the past seven decades, many crop models of varying complexities have been developed by different groups, which include Agricultural Production Systems Simulator (APSIM), CERES (now contained in the DSSAT Suite of crop suite), CROPSYST, EPIC, WOFOST, and COUP. See the survey paper [13] and a comparison of different crop models for yield response prediction [22]. Among the existing crop models, the ones that are extensively used globally are APSIM and DSSAT, which are still constantly evolving and currently open-source to facilitate community-based development.

Most of the existing crop models need the management practices to be pre-specified before the start of a simulation, while RL-based training of management policies requires the management practices to be determined according to the soil, plan and weather conditions on a daily or weekly basis during the simulation. In light of this, the authors of [19] developed the CropGym environment for training of N management policies, which provides an interface to Open AI Gym [5], a widely used toolkit for RL research, and enables an RL agent to interact with the crop environment weekly. However, CropGym is based on the LINTUL-3 model [25] for the wheat crop, which has limited use. In a similar spirit, [2] presented another crop environment with the Open AI Gym interface for the russet potato based on the SIMPLE crop model [31], which, again, has limited use, potentially because the model is over-simplified. Recently, a Gym-DSSAT environment for the maize crop, which is based on the widely used DSSAT suite of crop models and provides a Gym interface, was developed [8] and enables an RL agent to interact with the environment on a daily basis. However, there have been no results on the use of Gym-DSSAT for training crop management policies up to now.

3. Methods

We now present technical details for the N management framework based on deep RL and crop simulations depicted in Fig. 1.

3.1. MDP Problem Formulation

The N management problem can be formulated as a finite Markov decision process (MDP) problem. In this formulation, a decision-making agent continuously interacts with the environment. At each day t, the agent selects an action (i.e., management practice), a_t, from the action space $\mathcal{A}$, based on the current state s_t, which is an array of elements from the state space $\mathcal{S}$. The selected action is applied to the environment and a new state (s_{t+1}) is generated based on this action; meanwhile, a reward signal $r_t = r(s_t, a_t)$ is produced to evaluate the immediate consequence of the selected action. This interaction repeats until the termination of the interaction, e.g., when the crop is harvested. The goal of the agent is to select optimal actions to maximize the future discounted return. The future discounted return at time t is defined as $R_t = \sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau$, where T is the time step at termination.

For N management, the action space $\mathcal{A}$ contains all possible amounts of nitrogen applied at a day. All the states that compose the state space are listed in Table 7. The reward function $r(s_t, a_t)$ at day t is set as:

$$r(s_t, a_t) = \begin{cases} w_1 Y - w_2 a_t - w_3 N_{l,t} - w_4 P_t & \text{if harvest at } t, \\ -w_2 a_t - w_3 N_{l,t} - w_4 P_t & \text{otherwise,} \end{cases} \tag{1}$$

where a_t is the action (i.e. amount of nitrogen applied at day t), $N_{l,t}$ is the nitrate leaching at day t, Y is the crop yield at the harvest date represented by the top weight at maturity, and P_t is the additional penalty on large *total amount* of nitrogen applied. In particular, $P_t = \sum_{k=1}^{t} a_k - threshold$ if $a_t \neq 0$ and $P_t = 0$ if $a_t = 0$, where $threshold$ represents the allowable total amount of nitrogen inputs. It may be worth mentioning that nitrate leaching occurs when nitrate is washed out of the root zone by heavy rainfall. Leaching is undesirable because it leads to the waste of the fertilizers, and more importantly, causes environmental problems such as eutrophication of watercourses and soil degradation. Thus, we include a penalty on nitrate leaching in the reward function. Finally, the positive constants $w_1 \sim w_4$ are selected to balance the different aspects mentioned above.

3.2. Training Management Policies using Deep RL

For solving the formulated MDP problem, we leverage the recently proposed deep RL algorithms, which have achieved remarkable performance on a variety of tasks [6,11,12,15,18,26,30]. We choose deep Q-network (DQN) [18] and soft actor-critic (SAC) [9] for the experimental study, but other deep RL algorithms capable of handling continuous state spaces can also be applied.

3.2.1 Policy Training with DQN

DQN is a model-free value function based deep RL algorithm, which uses a deep neural network (DNN) to approximate the action-value function in Q-network [18]. The essential idea of DQN is to learn an optimal action-value function $Q^\star(s, a) = \max_\pi \mathbb{E}[R_t | s_t = s, a_t = a, \pi]$, where π is a policy mapping a state s_t to an action a_t at a given time t. With the $Q^\star$ function, given an action s_t, an optimal action $a_t^\star$ can be determined, e.g., by following a greedy policy defined by $a_t^\star = \max_{a \in \mathcal{A}} Q^\star(s_t, a)$. From the interaction between the agent and environment, tuples of (s, a, r, s') are generated and stored in a replay buffer, where s, a, r and s' denote current state, current action, immediate reward obtained by applying the action a at the state s, and next state, respectively. Due to the nature of continuity of the action space, we discretized the action space. At iteration i, the Q network can be trained by minimizing the loss function:

$$L_i(\theta_i) \triangleq \mathbb{E}_{(s,a,r,s')} \left[r + \gamma \max_{a' \in \mathcal{A}} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i) \right], \tag{2}$$

where the tuples (s, a, r, s') are sampled from the replay buffer, θ_i are the parameters of the Q-network at iteration i, and θ_i^- are the network parameters used to compute the target at iteration i. The optimization problem can be solved using stochastic gradient descent algorithms [18].

3.2.2 Policy Training with SAC

SAC is a policy-gradient deep RL algorithm that represents the state of the art among model-free RL algorithms in terms of sample efficiency and stability with respect to the hyperparameters [9]. Besides the expected sum of rewards, SAC introduces the expected entropy to favor stochastic policies, which leads to a cost function L defined by

$$L \triangleq -\sum_{t=0}^{T} \mathbb{E}_{(s_t, a_t) \sim p_\pi} \left[r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | s_t)) \right], \tag{3}$$

where p denotes the state-action marginals of the trajectory distribution, $\mathcal{H}$ determines the entropy for the evaluation of randomness given the state s_t, and $r(s_t, a_t)$ is the immediate reward at time t. The temperature parameter α decides the trade-off between the entropy term and rewards.

3.3. Simulating the Crop Response using Gym-DSSAT

DSSAT has been used for various crop simulations worldwide in the last 30 years [14]. However, limited interactions can be reached during the running period of simulation, leading to a possible delay of adjustment for management decisions. Recently, Gym-DSSAT [8] has been developed to bridge the communication gap between the simulation environment and daily management decisions. This communication pipeline enables RL researchers to manipulate DSSAT like Open AI Gym in machine learning and

robotics [18, 30]. In Gym-DSSAT, the environment is defined at a field scale with a time step corresponding to one day. An episode typically covers about 160 days from planting to harvest, and its state is automatically set as "done" at crop maturity. Weather is randomly generated via WGEN's [21] built-in stochastic weather generator and can be fixed depending on simulation purposes.

With Gym-DSSAT, millions of daily interactions between an RL agent and the simulated crop environment can be achieved in a few minutes, and used for training the management policies.

4. Experiments and Results

We conducted experiments on training N management policies for the maize crop in both Florida and Iowa. These two locations are selected since they have different weather and soil conditions, which can be leveraged to test the general applicability of the proposed framework. Also, DSSAT includes templates for simulating the maize crop in these two locations, which facilitates the implementation of our proposed framework. We evaluated the performance of trained policies in comparison with the standard practice proposed in [17].

4.1. Datasets for Florida and Iowa

Two experiments were studied. The first one is for the maize crop in Ames, Iowa, in 1999. The simulation starts on April 25th, the planting happens on May 27th, and the crop is harvested no later than Oct 24th. The soil has a depth of 151 cm, and the plant density is 7.6 plant/m^2. The second experiment is for the maize crop in Gainesville, Florida, in 1982. In the Florida setup, the simulation starts on Jan 30th, while the crop is planted on Feb 26th and harvested when reaching maturity. The soil in this case has a depth of 180 cm, and the plant density is 7.2 plant/m^2. For both simulations, the irrigation is set to 0. This is consistent with the current practice in Iowa, where the maize crop is not irrigated. On the other hand, irrigation is crucial to improve the maize yield in Florida in reality. Setting the irrigation to 0 for the Florida case can be considered as an emulation of the extreme case of severe drought and limited water supply, which allows us to compare the RL-based management strategy with the standard practice under this extreme case.

4.2. Implementation Details

For all the experiments, weight parameters w_1, w_2, and w_3 in the reward function (1) were set to be 0.1, and w_4 is set to be 1. For both DQN and SAC, we implemented the training using Pytorch, and used the Adam [16] optimizer with an initial learning rate of 0.00005 and a batch size of 64 to train the neural network. We trained the policies for 1200 episodes with the exploration rate ϵ decreasing from 1 to 0, following a decay factor of 0.994 for the Florida case and of 0.992 for the Iowa case.

For DQN, the discrete action space was defined to be $\mathcal{A} = \{40k\frac{\text{kg}}{\text{ha}}|k = 0, 1, 2, 3, 4\}$. The discount factor was set to be 0.99.

For SAC, the agent action a_{sac} varies from 0 to 200 and is discretized into the same action space as the one used for DQN through the mapping: $\arg\min_{a \in \mathcal{A}} \|a_{sac} - a\|$, for both training and testing. The discretization is for being consistent with farmers' fertilization patterns, i.e., fertilize only a few times in the whole growth cycle. The discount factor and smoothing constant for updating the target network were set to be 0.98 and 0.001, respectively.

For comparison with the trained policies, we also implemented the standard management practice in [17], which suggests to add nitrogen at vegetative growth stage (vstage) 5, the stage when crop reaches five expanded leaves.

4.3. Results for the Iowa Maize

The training curve using DQN, averaged over five trials, is shown in Fig. 2. During the first 200 episodes of exploration, due to the large exploration rate, the DQN agent over-fertilizes, causing significant penalties. After 800 episodes of training, the learning converges, constantly giving a cumulative reward of over 2000. Concretely, Table 1 compares the performance of the DQN-trained policy and three baseline strategies, corresponding to 160, 240, 280 kg/ha of nitrogen applied at stage v5, as suggested in [17]. The trained DQN agent decides to apply a total of 240 kg/ha nitrogen input during the growing season, and achieves 21711.8 kg/ha top weight of maize at maturity and a cumulative reward of 2126.3. Among three baseline policies, the one with 280 kg/ha achieves the largest cumulative reward of 2142.9 and largest top weight of 21709.5 kg/ha. Compared with the best baseline, DQN achieves slight improvement on top weight at maturity while using 14% less nitrogen input, being more cost-efficient. Compared to the baseline using same amount of nitrogen input, DQN achieves a 1% increment on the top weight at harvest. In general, the trained DQN agent achieves better results than the baseline methods.

The performance of SAC is shown in Table 2. Although using less nitrogen that causes a smaller top weight at maturity, the SAC policy still achieves a cumulative reward similar to that achieved by the DQN policy. Thus, both RL algorithms succeeded in finding a better management policy than the baselines. However, the learning process converged much faster with SAC. Specifically, the cumulative reward with SAC reached 2100 within 700 episodes, while additional 300 episodes were needed to achieve similar results with DQN.

Table 1. Performance comparison between DQN and baseline policies for Iowa. Baseline (X) indicates that X kg/ha of nitrogen is applied at stage v5.

Methods	Nitrogen input (kg/ha)	Nitrate leaching (kg/ha)	Nitrogen uptake (kg/ha)	Top weight at maturity (kg/ha)	Cumulative reward
Baseline (160)	160	0.11	219.1	21133.3	2097.3
Baseline (240)	240	0.11	264.0	21502.9	2126.3
Baseline (280)	280	0.11	272.3	21709.5	2142.9
DQN	240	0.12	290.4	**21711.8**	**2147.1**

Table 2. Performance comparison between SAC and DQN policies for Iowa.

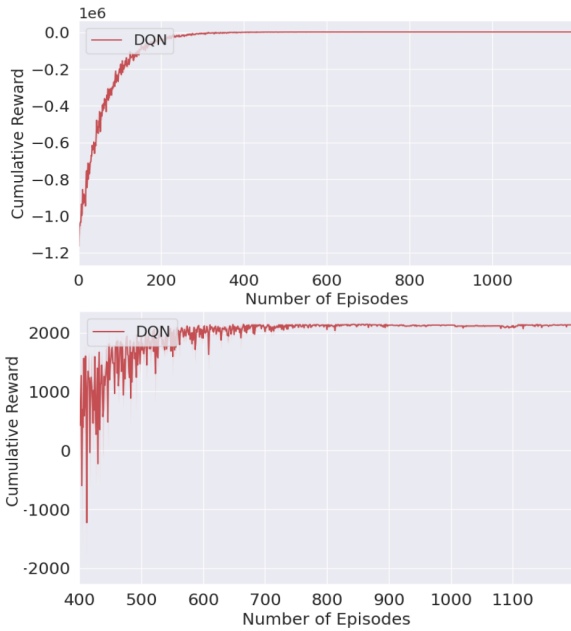| Methods	Nitrogen input (kg/ha)	Top weight at maturity (kg/ha)	Cumulative reward	Episodes of convergence
DQN	240	21711.8	2147.1	1000
SAC	200	21503.3	2144.2	700

Figure 2. Cumulative reward versus episodes with DQN for Iowa. Results are averaged over five trials, with the light-red shaded area denotes the variance. Top: full view. Bottom: zoomed-in view for 400–1200 episodes

4.4. Results for Florida Case

As we mentioned in 4.1, the Florida case is not realistic due to the 0 irrigation setup, and can be considered as an extreme weather case under severe drought with water shortage. Accordingly, the yields obtained under this setup are much smaller compared to those under the Iowa case. The training curve under DQN averaged over five trials is shown in Fig. 3. The performance comparison between our trained DQN policies and baselines is summarized in Table 3. As one can see, the DQN policy shows a stable improvement in terms of the top weight and rewards, which is consistent with the results for Iowa. The performance comparison between SAC and DQN is shown in Table 4. Similar to the

Iowa case, the SAC policy achieved a similar cumulative reward as the DQN policy but the training with SAC converged much faster.
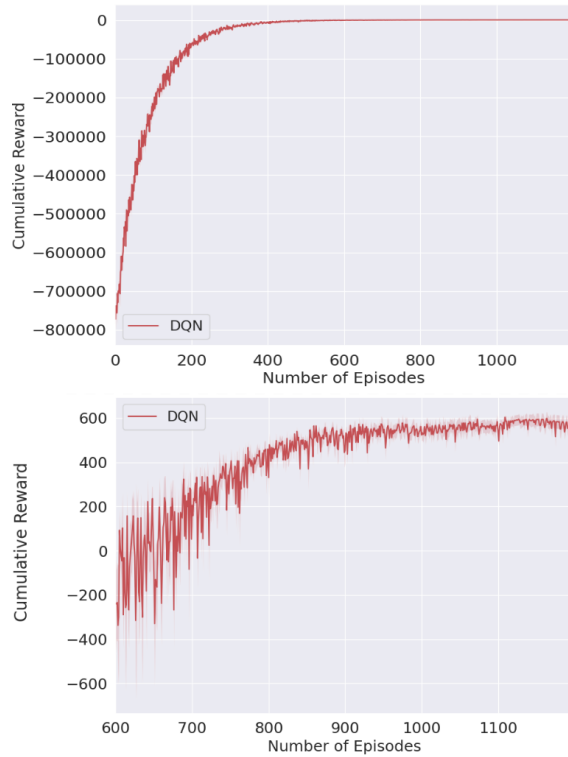

Figure 3. Cumulative reward versus episodes with DQN for Florida. Results are averaged over five trials, with the light-red shaded area denotes the variance. Top: full view. Bottom: zoomed-in view for 400–1200 episodes

4.5. Ablation Study

In practice, not all the states used in the training and testing of the management policies in the previous sections are accessible. Additionally, from an economic perspective, farmers prefer to make decisions less frequently,

Table 3. Performance comparison between DQN and baseline policies for Florida. Baseline (X) indicates that X kg/ha of nitrogen is applied at stage v5.

Methods	Nitrogen input (kg/ha)	Nitrate leaching (kg/ha)	Nitrogen uptake (kg/ha)	Top weight at maturity (kg/ha)	Cumulative reward
Baseline (40)	40	46	55	4393.3	430.7
Baseline (80)	80	65	66	4673.1	452.8
Baseline (160)	160	97	86	5190.4	493.3
DQN	80	33	105	**6310.8**	**619.7**

Table 4. Performance comparison between SAC and DQN policies for Florida.

Methods	Nitrogen input (kg/ha)	Top weight at maturity (kg/ha)	Cumulative reward	Episodes of convergence
DQN	80	6310.8	619.7	900
SAC	80	6308.0	610.2	700

e.g., weekly and biweekly, instead of daily. Therefore, in this section, we study the effect of full/partial observation and action frequencies on the performance of the proposed framework.

4.5.1 Full vs. Partial Observation

To understand the contribution of observed states in the fertilizer optimization process, we carry out an ablation study on (i) full observation case, in which all the states listed in Table 7 are used for policy training and testing and (ii) partial observation case, in which only 10 states (indicated in Table 7) are used. The study on partial observation is motivated by the fact that not all the states output by DSSAT can be accessed by farmers without professional agricultural tools for detection and inspection. Experiments of full observation have been conducted in Sec. 4. The results under partial observation, which are based on DQN, are shown in Fig. 4. For both Florida and Iowa, the policy training and testing under partial observation were conducted three times, and Fig. 4 shows the results averaged over the three trials.

For both Florida and Iowa setups, the policies under partial observation always underperformed those under full observation. In particular, we observed 30.15% and 3.58% decreases in reward and 27.94% and 3.68 % drops in the final yield for Florida and Iowa, respectively. The decrease for Florida is relatively large compared to that for Iowa. This could be attributed to the extreme weather condition associated with Florida which requires the RL agent to have more comprehensive information to make a good decision.

4.5.2 Action Frequency

We ablate the action frequency in the life cycle of maize to further understand the applied actions during the simulation process. We continue to conduct all the experiments with DQN in a discrete space to ensure the consistency with farmers' fertilization patterns. Concretely, we experiment
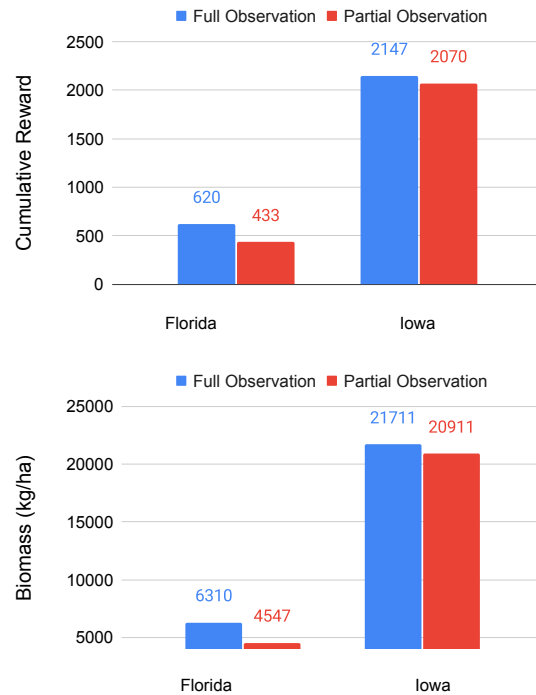

Figure 4. Comparison of cumulative rewards (top) and final above the ground population biomass (bottom) obtained under partial observation and full observation. The results are averaged over three trials.

with the trained DQN policy using two action frequencies: (i) RL agents are allowed to fertilize every day and (ii) RL agents are only permitted to fertilize every ten days.

For Iowa, Fig. 5 shows the applied actions and Table 5 lists achieved cumulative reward and top weight at maturity under different actions. The DQN agents fertilized 5 to 6 times. Both the baseline method and the DQN policy with a 10-day action frequency used 280 kg/ha nitrogen input. However, their results are relatively poor compared with the DQN policy with a 1-day action frequency.
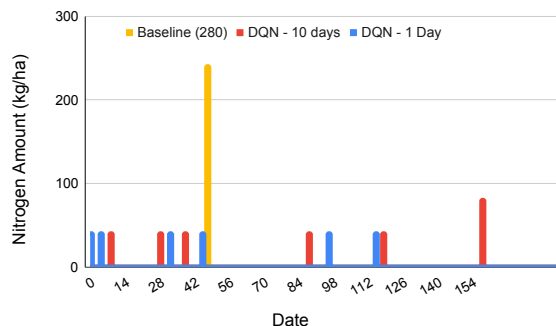
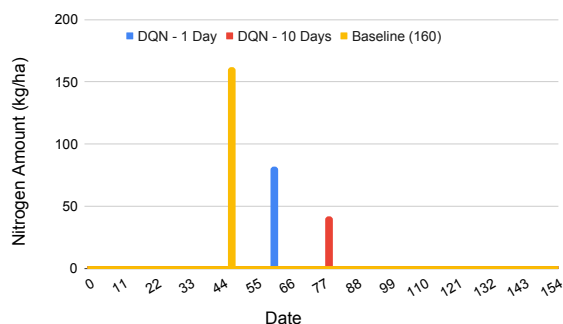Figure 5. Applied actions under different action frequencies for Iowa.

Figure 6. Applied actions under different action frequencies for Florida.

The results for Florida are shown in Fig. 6 and Table 6. One can see that both baseline methods and our RL agent chose to fertilize once across the crop's life cycle. However, the RL agent tended to apply less nitrogen, as shown in Fig. 6. According to Table 6, the performance of the DQN agent degrades a little bit under the reduced action frequency of 10 days, but is still better than that of the best baseline.

Table 5. Performance of trained policies under different action frequencies for Iowa. DQN (1 Day): RL agents are allowed to fertilize every day. DQN (10 Days): RL agents are only permitted to fertilize every ten days.

Method	Top weight at maturity (kg/ha)	Cumulative reward
DQN (1 Day)	21710.7	2147.1
DQN (10 Days)	21700.6	2142.0
Baseline	21709.5	2142.9

5. Conclusion

Effective nitrogen (N) management is crucial for maximizing crop yields and minimizing negative environmental

Table 6. Performance of trained policies under different action frequencies for Florida.

Method	Top weight at maturity (kg/ha)	Cumulative reward
DQN (1 Day)	6310.8	619.8
DQN (10 Days)	5728.9	565.7
Baseline	5190.4	493.3

impacts. We present a framework for optimizing N management with deep reinforcement learning (RL) and crop simulations based on DSSAT. With the proposed framework, we train management policies with deep Q-network (DQN) and soft actor-critic (SAC) for the maize crop in Iowa and Florida, which are shown to outperform standard management practices. We also evaluate the effect of partial observation and reduced action frequencies. We believe our work demonstrates the potential of deep RL in optimizing crop management for more sustainable and resilient agriculture.

Current study is focused on testing the proposed framework on a single type of crop in the simulator under a fixed weather condition (once the location is fixed). In the future, we will randomize weather during training and test the robustness of trained policies in the presence of uncertain weather conditions. Moreover, we plan to leverage real-world data to calibrate the crop simulation models and/or fine-tune the policy, which will help bridge the sim-to-real gap. Additionally, we plan to explore more advanced methods to improve the learning performance under partial observability. Besides, we plan to include other management practices such as irrigation and tillage in the action space. Finally, modifications can be made to Gym-DSSAT so that more types of crops can be tested with our proposed framework.

6. Acknowledgment

Table 7. State space description

State	Description	Included in Partial Observation Study?
cumsumfert	cumulative nitrogen fertilizer applications (kg/ha)	✓
dap	days after simulation started	✓
dtt	growing degree days for current day (C/d)	✓
istage	DSSAT maize growing stage	✓
vstage	vegetative growth stage (number of leaves)	✓
pltpop	plant population density (plant/m2)	✓
rain	rainfalls for the current day (mm/d)	✓
srad	solar radiations during the current day (MJ/m2/d)	✓
tmax	maximum temperature for current day (C)	✓
tmin	minimum temparature for current day (C)	✓
nstres	index of plant nitrogen stress (unitless)	
pcngrn	massic fraction of nitrogen in grains (unitless)	
swfac	index of plant water stress (unitless)	
tleachd	daily nitrate leaching (kg/ha)	
grnwt	grain weight dry matter (kg/ha)	
cleach	cumulative nitrate leaching (kg/ha)	
cnox	cumulative nitrogen denitrification (kg/ha)	
tnoxd	daily nitrogen denitrification (kg/ha)	
trnu	daily nitrogen plant population uptake (kg/ha)	
wtnup	cumulative plant population nitrogen uptake (kg/ha)	
xlai	plant population leaf area index (m2_leaf/m2_soil)	
topwt	top weight (kg/ha)	
es	actual soil evaporation rate (mm/d)	
runoff	calculated runoff (mm/d)	
wtdep	depth to water table (cm)	
rtdep	root depth (cm)	
totaml	cumulative ammonia volatilization (kgN/ha)	
sw	volumetric soil water content in soil layers (cm3 [water] / cm3 [soil])	

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019. 1

[2] Chace Ashcraft and Kiran Karra. Machine learning aided crop yield optimization. *arXiv preprint arXiv:2111.00963*, 2021. 2

[3] J-M Attonaty, M-H Chatelin, F Garcia, and S Ndiaye. Using extended machine learning and simulation technics to design crop management strategies. In *EFITA First European Conference for Information Technology in Agriculture, Copenhagen (Denmark)*, 1997. 1

[4] Jacques-Eric Bergez, Nathalie Colbach, Olivier Crespo, Frederick Garcia, Marie-Helene Jeuffroy, Eric Justes, Chantal Loyce, Nicolas Munier-Jolain, and Walid Sadok. Designing crop management systems by simulation. *European Journal of Agronomy*, 32(1):3–9, 2010. 1

[5] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016. 2

[6] Chris Gamble and Jim Gao. Safety-first AI for autonomous data centre cooling and industrial control. 1, 2, 3

[7] Frédérick Garcia. Use of reinforcement learning and simulation to optimize wheat crop technical management. In *Proceedings of the International Congress on Modelling and Simulation*, pages 801–806, 1999. 2

[8] Romain Gautron and Emilio J Padrón González. gym-DSSAT - A crop model turned into a Reinforcement Learning environment, 3 2022. 2, 3

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 3

[10] G. Hoogenboom, K.J. Boote C.H. Porter, V. Shelia, P.W. Wilkens, U. Singh, J.W. White, S. Asseng, J.I. Lizaso, L.P. Moreno, W. Pavan, R. Ogoshi, L.A. Hunt, G.Y. Tsuji, and J.W. Jones. The DSSAT crop modeling ecosystem. In K.J. Boote, editor, *Advances in Crop Modeling for a Sustainable Agriculture*, pages 173–216. Burleigh Dodds Science Publishing, Cambridge, United Kingdom, 2019. 2

[11] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019. 1, 2, 3

[12] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2193–2201, 2018. 3

[13] James W Jones, John M Antle, Bruno Basso, Kenneth J Boote, Richard T Conant, Ian Foster, H Charles J Godfray,

Mario Herrero, Richard E Howitt, Sander Janssen, et al. Brief history of agricultural systems modeling. *Agricultural systems*, 155:240–254, 2017. 2

[14] James W Jones, Gerrit Hoogenboom, Cheryl H Porter, Ken J Boote, William D Batchelor, LA Hunt, Paul W Wilkens, Upendra Singh, Arjan J Gijsman, and Joe T Ritchie. The DSSAT cropping system model. *European Journal of Agronomy*, 18(3-4):235–265, 2003. 2, 3

[15] Elia Kaufmann, Antonio Loquercio, Rene Ranftl, Alexey Dosovitskiy, Vladlen Koltun, and Davide Scaramuzza. Deep drone racing: Learning agile flight in dynamic environments. In *Conference on Robot Learning*, pages 133–145, 2018. 1, 2, 3

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[17] German Mandrini, Cameron M Pittelkow, Sotirios V Archontoulis, Taro Mieno, and Nicolas F Martin. Understanding differences between static and dynamic nitrogen fertilizer tools using simulation modeling. *Agricultural Systems*, 194:103275, 2021. 4

[18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 1, 2, 3, 4

[19] Hiske Overweg, Herman NC Berghuijs, and Ioannis N Athanasiadis. CropGym: a reinforcement learning environment for crop management. *arXiv preprint arXiv:2104.04326*, 2021. 2

[20] Mike W Palmer, Julia Cooper, Catherine Tétard-Jones, et al. The influence of organic and conventional fertilisation and crop protection practices, preceding crop, harvest year and weather conditions on yield and quality of potato (Solanum tuberosum) in a long-term management trial. *European Journal of Agronomy*, 49:83–92, 2013. 1

[21] CW Richardson. Weather simulation for crop management models. *Transactions of the ASAE*, 28(5):1602–1606, 1985. 4

[22] Tapio J Salo, Taru Palosuo, Kurt Christian Kersebaum, Claas Nendel, Carlos Angulo, Frank Ewert, Marco Bindi, Pierluigi Calanca, Tommy Klein, Marco Moriondo, et al. *The Journal of Agricultural Science*, 154(7):1218–1240, 2016. 2

[23] Tim Searchinger, Richard Waite, Craig Hanson, Janet Ranganathan, Patrice Dumas, Emily Matthews, and Carni Klirs. Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050. Final report, 2019. 1

[24] Tim Searchinger, Richard Waite, Craig Hanson, Janet Ranganathan, Patrice Dumas, Emily Matthews, and Carni Klirs. *Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050. Final report*. World Resources Institute, 2019. 1

[25] ME Shibu, PA Leffelaar, H Van Keulen, and PK Aggarwal. LINTUL3, a simulation model for nitrogen-limited situations: Application to rice. *European Journal of Agronomy*, 32(4):255–271, 2010. 2

[26] Yunlong Song, Mats Steinweg, Elia Kaufmann, and Davide Scaramuzza. Autonomous drone racing with deep reinforcement learning. *arXiv preprint arXiv:2103.08624*, 2021. 1, 2, 3

[27] Lijia Sun, Yanxiang Yang, Jiang Hu, Dana Porter, Thomas Marek, and Charles Hillyer. Reinforcement learning control for water-efficient agricultural irrigation. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pages 1334–1341, 2017. 2

[28] Mark A Sutton, Oene Oenema, Jan Willem Erisman, Adrian Leip, Hans van Grinsven, and Wilfried Winiwarter. Too much of a good thing. *Nature*, 472(7342):159–161, 2011. 1

[29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 2

[30] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. 1, 2, 3, 4

[31] Chuang Zhao, Bing Liu, Liujun Xiao, Gerrit Hoogenboom, Kenneth J Boote, Belay T Kassie, Willingthon Pavan, Vakhtang Shelia, Kwang Soo Kim, Ixchel M Hernandez-Ochoa, et al. A SIMPLE crop model. *European Journal of Agronomy*, 104:97–106, 2019. 2