Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Proximal Policy Optimization Framework for Farm-Scale Crop Allocation under Rotation, Subsidy, and Multi-Crop Expansion Constraints

Niantang Liu

liuniantang@caas.cn

Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences

Guijun Yang

Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences

Xiaodong Yang

Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences

Si-Bo Duan

Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences

Liang Sun

Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences

Zixuan Liu

Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences

Research Article

Keywords: crop planning, reinforcement learning, Proximal Policy Optimization (PPO), crop rotation, agricultural decision support, multi-objective optimization, soybean-corn allocation

Posted Date: June 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9917542/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Proximal Policy Optimization Framework for Farm-Scale Crop Allocation under Rotation, Subsidy, and Multi-Crop Expansion Constraints

Niantang Liu[1], Guijun Yang[2], Xiaodong Yang[2], Si-Bo Duan[1], Liang Sun[1], Zixuan Liu[1]

[1] *State Key Laboratory of Efficient Utilization of Arid and Semi-arid Arable Land in Northern China, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China*

[2] *Key Laboratory of Quantitative Remote Sensing in Agriculture of Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*

Abstract

Farm-level crop allocation is a sequential, constraint-rich decision problem in which farmers must reconcile biological rotation requirements, subsidy-driven economic incentives, and externally imposed crop-expansion targets. Traditional mathematical-programming and population-based evolutionary approaches require explicit reformulation for each new constraint and lack the ability to learn from sequential parcel-by-parcel decisions; heuristic rules in turn provide no formal handle on the trade-offs between competing objectives. We address this gap by formulating the annual crop allocation problem as a Markov Decision Process (MDP) over parcels and training a Proximal Policy Optimization (PPO) agent whose reward jointly encodes target-crop expansion, rotation compliance, economic surplus, and fertilizer-budget guards. The framework is evaluated on Hongxing Farm (Bei'an, Heilongjiang Province, Northeast China), a 308,443-mu farm with 553 parcels, across two policy-driven scenarios. In the soybean-expansion scenario the learned policy achieves a 4,599 mu soybean increase (92% of the 5,000 mu target), maintains rotation compliance at 85.71% (above the 85% policy threshold), and improves total economic return by 4.91% with only a 6.31% increase in fertilizer expenditure. The same trained policy supports built-in subsidy counterfactual analysis through a single-parameter retraining sweep, positioning the framework as a forward-looking complement to ex-post supply-response analyses of recent policy-reform studies. Benchmarked against the farm's own realized 2023 and 2024 allocations, the learned plan is the only one of the three that satisfies every policy constraint, restoring rotation compliance to 85.71% (above the 0.85 floor) where the unguided 2024 outcome falls to 78.84%.

Keywords: crop planning; reinforcement learning; Proximal Policy Optimization (PPO); crop rotation; agricultural decision support; multi-objective optimization; soybean-corn allocation

1. Introduction

Modern farm management sits at the intersection of agronomic, economic, and environmental imperatives. A farmer's annual choice of which crop to grow on each parcel propagates through soil nitrogen pools, pest and disease cycles, working-capital flows, and household labor schedules; over multiple seasons it determines the long-run sustainability of the farm operation [1]. The problem is structurally difficult because crop yields and revenues respond non-linearly to prior-crop identity through rotation effects [2–4],

policy instruments shift the return surface every season [5, 6], and the parcel space is heterogeneous in size, soil, and last-year crop, creating a combinatorially large action space at farm scale.

Chinese cropping policy in the northeast region provides a particularly demanding test bed for any farm-scale decision-support tool. Following the 2016 reform that replaced the temporary purchase-and-storage scheme for corn with a producer-subsidy mechanism, and the parallel target-price pilot for soybean, farms in Heilongjiang, Jilin, Liaoning, and Inner Mongolia have faced sharp shifts in the relative profitability of corn, soybean, wheat, and rice. The realized supply response was strongly heterogeneous across farm sizes: large-scale operations adjusted acreage decisively, whereas smaller farms largely retained continuous-corn rotation despite the price signal, indicating that price-only instruments are insufficient when rotation benefits, sunk fixed assets, and household labor jointly constrain the choice set [7]. At the national scale the strategic emphasis has shifted from absolute grain self-sufficiency to structural balance between food grain and feed grain, with provincial-level spatial reallocation between soybean, corn, wheat, and rice now an explicit policy lever [8]. Recent quantitative tools for this layer, for example the AgentLA agent-based model applied to three Heilongjiang counties [9], operate at the multi-county scale on raster grids and assume identical agents within a cell; they do not represent the parcel-by-parcel sequential decision that an individual farm manager actually makes when reconciling rotation history, subsidy structure, and area bounds across hundreds of contiguous fields.

A long line of work in agricultural operations research has approached crop allocation through linear and integer programming since Heady's foundational work [10], surveyed by Glen [11]; Hazell and Norton [12] compiled the canonical formulations and Kennedy [13] extended the framework to dynamic programming. Two families of population-based metaheuristics subsequently came to dominate the multi-objective crop-planning literature: particle swarm optimization, introduced by Kennedy and Eberhart [14] and extended to multi-objective settings by Coello et al. [15], with crop-planning applications by Marquez et al. [16] and the recent multi-objective PSO of Begam et al. [17]; and the elitist non-dominated sorting genetic algorithm NSGA-II [18, 19], refined for crop planning by Sarker and Ray [20]. Bio-economic farm models with embedded Pareto search (Groot et al.'s FarmDESIGN [21] built on the differential evolution of Storn and Price [22] and the Pareto-ranking ideas of Goldberg [19], and the spatio-temporal strip-cropping extension of Juventia et al. [23]) operate at the farm scale through frontier exploration. Stochastic dynamic programming under explicit rotation benefits, characterized in closed form by Boyabatlı et al. [6] building on Hennessy [3] and Livingston et al. [4], provides a theoretical baseline that Cai et al. [2] complement with numerical evidence on yield and price volatility. Supply-chain analyses extending to upstream and downstream actors [24, 25] and the comprehensive review of Jain et al. [26] complete the picture. These methods deliver optimality guarantees under specific functional forms but require costly reformulation whenever a new constraint type is introduced; heuristic decision rules are flexible but tied to a single policy scenario [27]. Crucially, none of these frameworks addresses the parcel-by-parcel sequential structure that arises when a single farm manager allocates hundreds of parcels under coupled rotation, subsidy, and area-bound constraints, the setting that motivates the present work.

Reinforcement learning (RL) offers a complementary approach. By framing crop allocation as a Markov Decision Process (MDP) [28, 29, 30], the agent learns from rewards that can be arbitrarily shaped to reflect biological, economic, and policy considerations. Early agricultural RL focused on within-season operational decisions: Garcia [31] used R-Learning for wheat technical management, Bergez et al. [32] applied Q-learning [33] to maize irrigation, and Trépos et al. [34] extended this line to canola; recent deep-RL applications include greenhouse climate control [35], irrigation scheduling [36, 37, 38], and site-specific

management [39]. The comprehensive review of Gautron et al. [40] explicitly identifies the cropping-plan layer, the level at which our framework operates, as an open application area. Among modern RL algorithms, Proximal Policy Optimization (PPO) [41] has become the workhorse for constrained, sample-efficient policy learning in discrete-action settings, complementing the deep-policy paradigm advanced by DQN [42], DDPG [43], and SAC [44]. For settings in which hard constraint satisfaction is critical, the Constrained Policy Optimization framework [45] provides a future-extension pathway.

This paper proposes a sequential, parcel-by-parcel MDP formulation in which the state captures parcel area, last-year crop identity, and the running per-crop allocation ratios; the action selects among five crops; the reward jointly encodes four objectives; and a deterministic auto-redirect rule prevents the most damaging form of continuous-soybean violation at the action stage. The trained PPO agent reaches 92% of the 5,000-mu soybean-expansion target while satisfying the 85% rotation-compliance constraint and improving aggregate gross margin by 4.91%. A comparison against the two operational reference points a county bureau would most likely consult, namely an unintervened 2023 baseline and a size-sorted greedy heuristic, confirms that the rotation guard is the binding ingredient and that PPO's per-parcel conditional logic is the source of its compliance advantage. A systematic ablation study explores sensitivity to reward weights, the rotation-compliance threshold, the soybean producer-subsidy level, network architecture, and parallelization strategy.

The contributions of this paper are:

(1) A parcel-level Proximal Policy Optimization formulation for farm-scale crop allocation that integrates soft rotation compliance, multi-crop area constraints, subsidy-augmented economics, and target-crop expansion into a unified four-component reward. The state encodes parcel area, prior-year crop indicator, and running cumulative allocation; the action selects among five crops; the reward is shaped so that all four operational requirements remain transparent to county-level agricultural staff.

(2) A deterministic auto-redirect rule applied at the action stage (converting any policy-sampled continuous-soybean assignment into corn), together with an end-of-episode fill-unassigned cleanup. The two mechanisms guarantee that every produced plan is feasible at the most policy-relevant constraint (the 0.85 rotation-compliance floor) and bit-for-bit reproducible across re-runs, properties not generally shared by population-based stochastic search.

(3) Integration of the 2023 Chinese producer-subsidy structure into the per-mu economic model (cultivated-land protection subsidy, grain-grower one-off subsidy, and the soybean-specific producer subsidy), and a subsidy-counterfactual analysis facility that, by re-training the agent at alternative subsidy levels, converts the framework into a quantitative policy-design instrument. This positions the work as a complement to the ex-post supply-response analyses of Zheng et al. [7] rather than a competitor to algorithm-development benchmarks.

(4) A comprehensive internal-ablation suite covering reward-component weights, the rotation-compliance threshold, the soybean producer-subsidy level, the auto-redirect rule, the policy/value network capacity, and the parallelization strategy, together providing quantitative guidance for future deployments on other farms and other subsidy regimes.

2. Materials

2.1 Study area

The case-study site is Hongxing Farm, a large-scale state farm in Bei'an City, Heilongjiang Province, in the cold-temperate single-cropping belt of Northeast China, where annual mean temperature, precipitation regime, and growing-season length jointly constrain the crop portfolio to soybean, corn, wheat, and small areas of organic cabbage and rice. The site lies within the operational footprint of the 2016 corn producer-subsidy reform [7] and the broader food-grain–to–feed-grain structural-adjustment programs documented at provincial and county scales [8, 9]. Mechanization is uniform across parcels and the field structure is comparatively large-scale, placing the farm in the size class where rotation-management decisions are most elastic to price signals [7]. The location and parcel-boundary layout of the case-study farm are shown in Figure 1.

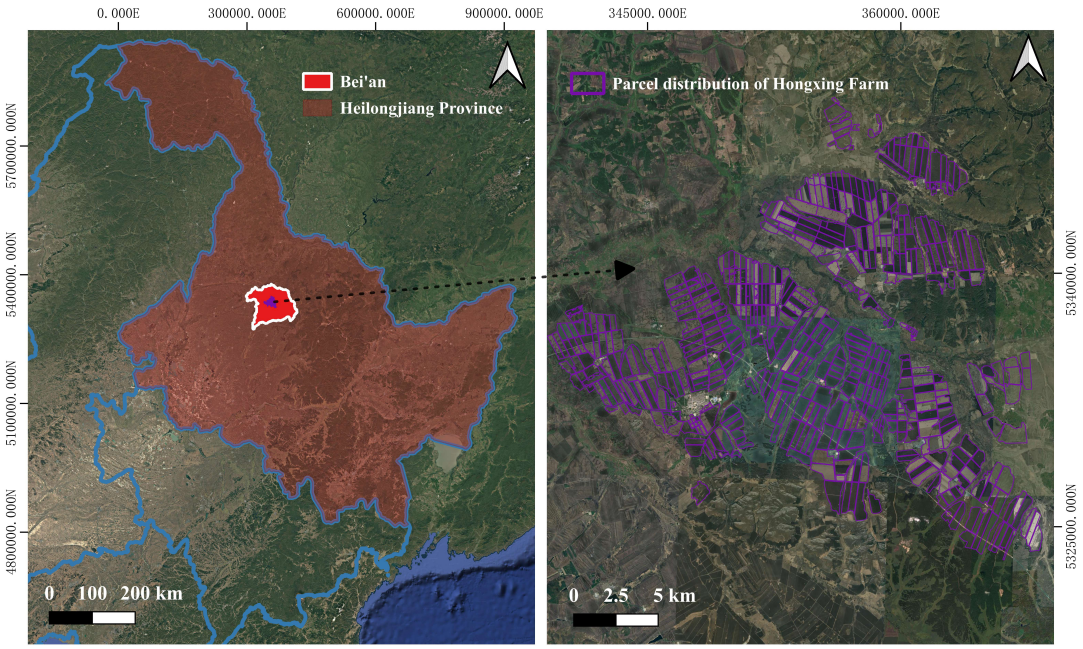


Figure 1. Study area. Location of Hongxing Farm (Bei'an, Heilongjiang Province) in Northeast China and parcel-boundary overlay.

The case-study farm comprises 553 parcels totaling 308,442.82 mu, with the following 2023 crop distribution: soybean 170,605.62 mu (55.31%), corn 106,184.20 mu (34.43%), wheat 29,103.40 mu (9.44%), organic cabbage 1,887.20 mu (0.61%), and rice 662.40 mu (0.21%). The mean parcel size is approximately 558 mu (Table 1).

Table 1. 2023 baseline crop distribution.

Crop	2023 area (mu)	Share (%)	Parcel count	Mean parcel (mu)
Soybean	170,605.62	55.31	312	546.81
Corn	106,184.20	34.43	194	547.34
Wheat	29,103.40	9.44	43	676.82
Organic cabbage	1,887.20	0.61	3	629.07
Rice	662.40	0.21	1	662.40

2.2 Subsidy-augmented economics

The per-crop economic parameters incorporate the official subsidy structure. The per-mu profit of crop *k* is computed as:

$$\pi_k = y_k p_k + s^{base} + s_k^{crop} - c_k \quad (1)$$

where yk and pk denote the per-mu yield and unit price, ck the total per-mu cost (including fertilizer), sbase a base subsidy applied to all main crops, and skcrop a crop-specific subsidy. In our setting sbase = 71.7 yuan mu^{-1}, whereas skcrop = 14 and 366 yuan mu^{-1} for corn and soybean, respectively; the markedly higher soybean subsidy reflects the recent policy emphasis on protein crops. Under these parameters the per-mu profit is 585.7 yuan mu^{-1} for corn and 673.7 yuan mu^{-1} for soybean.

2.3 Scenarios and constraints

Soybean-expansion: target 5,000 mu; bounds $A_{soy} \in [171{,}000, 180{,}000]$, $A_{corn} \in [97{,}000, 120{,}000]$, $A_{wheat} \in [9{,}000, 15{,}000]$. The upper bounds on soybean and corn include a 5,000-mu headroom above the policy-defined target to accommodate compliance-driven reallocation: parcels auto-redirected away from continuous soybean must be reassigned to corn, which indirectly lifts the corn total beyond the nominal target while the residual rotation churn raises soybean somewhat above baseline + Δ*. Without this headroom the agent is structurally forced onto the corner of the feasible set. Corn-expansion: target 20,000 mu; bounds $A_{soy} \in [150{,}000, 180{,}000]$, $A_{corn} \in [130{,}000, 150{,}000]$.

2.4 Software environment

Python 3.10, PyTorch 2.6 with CUDA 12.6 on an NVIDIA RTX 4090 GPU, gym 0.21 wrapped via stable-baselines3 2.x, pandas 2.x. Evaluation under fixed seed 42.

3. Method

3.1 Problem formulation

We consider a farm operating one crop per parcel per year. The farm comprises N parcels; each parcel i has area a_i (in mu, where 1 mu = 1/15 hectare) and a last-year crop label c_i^{prev} drawn from the crop set K = {soybean, corn, wheat, organic cabbage, rice}. The decision is to assign each parcel a current-year crop c_i in K. Let A_k denote the total area allocated to crop k:

$$A_k = \sum_{i:\, c_i = k} a_i \,, k \in K \quad (2)$$

The farm faces five categories of requirements:

(R1) Target-crop expansion. A target crop k^* and desired increase Δ^* are given; the agent should drive $A_{k^*} - A_{k^*}^{prev} \geq \Delta^*$.

(R2) Per-crop area constraints. Each crop k has a minimum and maximum area, $A_k^{min} \leq A_k \leq A_k^{max}$.

(R3) Rotation compliance. Continuous-soybean cropping forgoes rotation benefits and degrades soil [3]; rotation is correspondingly encouraged under current Chinese cropping policy. The compliance rate is defined as

$$\rho = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\neg(c_i = soybean \wedge c_i^{prev} = soybean)\} \quad (3)$$

and should satisfy $\rho \geq \rho^*$ with $\rho^* = 0.85$ in our default setting. The 0.85 floor is a design choice rather than a figure mandated by a specific regulation: we adopt it as the model's default evaluation criterion, a deliberately demanding rotation-discipline level against which every reported result is judged, and we report the agent's sensitivity to this value in Section 5.1.

(R4) Economic improvement. Let π_k be the per-mu profit of crop k. Total profit is

$$\Pi = \sum_{k \in K} \pi_k A_k \tag{4}$$

and should not fall below the previous-year baseline Π^{prev}.

(R5) Fertilizer guard. Let φ_k be the per-mu fertilizer cost. Total fertilizer outlay is

$$\Phi = \sum_{k \in K} \varphi_k A_k \tag{5}$$

and should satisfy $\Phi \leq 1.30 \cdot \Phi^{prev}$, capping the input-intensity rebound effect.

The combinatorial size of the action space is $|K|^N$ which is approximately 5^{553}, ruling out exhaustive search. Standard linear programming is technically applicable but the band-shaped target signal and the threshold-style compliance constraint introduce indicator structure that translates into a large mixed-integer program with weak relaxations. We therefore turn to RL. The remainder of this section develops the framework: Section 3.2 formalizes the parcel-visit sequence as a Markov decision process whose terminal reward encodes (R1)–(R5) directly; Section 3.3 specifies the policy and value networks and the PPO objective; Section 3.4 documents the determinism and reproducibility protocol; and Section 3.5 introduces the two operational reference baselines against which the trained policy will be compared in Section 4.2.

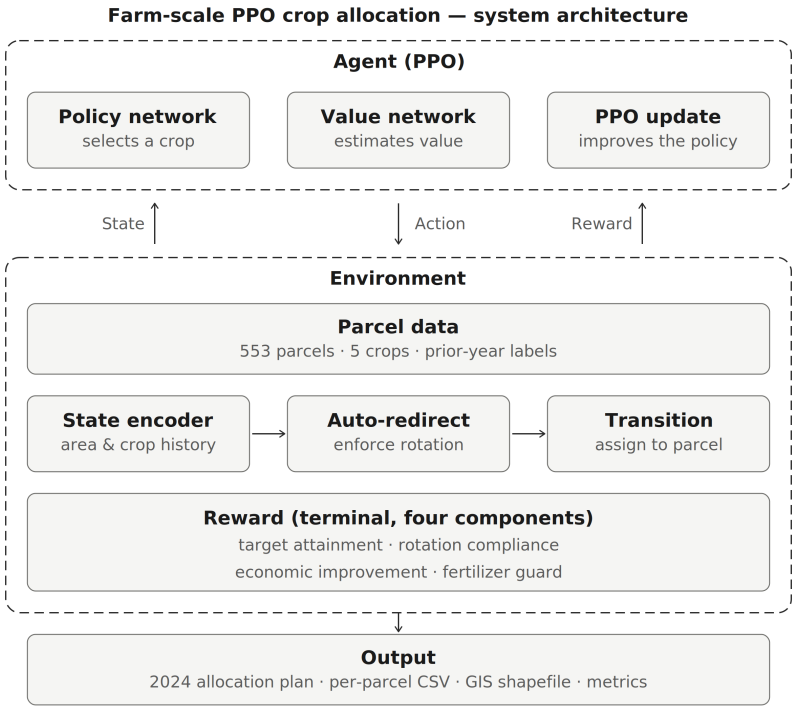


Figure 2. System architecture and information flow. Parcel attributes feed into the MDP; the PPO policy emits a per-parcel crop assignment; the auto-redirect and fill-unassigned modules enforce hard rotation feasibility; the resulting plan is scored on the four-component reward and the value head is updated with GAE.

3.2 MDP formulation

The overall framework, comprising the agent, environment, and reward, is summarized in Figure 2, with the per-parcel interaction loop visiting all 553 parcels in a single training episode. We define the MDP M = (S, A, P, r, γ). At step t the state encodes the current parcel together with the running cumulative allocation:

$$s_t = \left[a_t, \mathbb{1}\{c_t^{prev} = soy\}, \frac{A_{soy,t-1}}{A_{soy}^{max}}, \frac{A_{corn,t-1}}{A_{corn}^{max}}, \ldots, \frac{A_{rice,t-1}}{A_{rice}^{max}}\right] \quad (6)$$

The action a_t in {1, ..., |K|} selects the crop for parcel t. The transition is deterministic with one in-environment correction: if c_t^{prev} = soy and a_t = soy, the action is auto-redirected to a_t = corn. This injects an inviolable rotation guard at the action level. The terminal reward decomposes into target-attainment, economic-improvement, fertilizer-guard, and rotation-compliance components:

$$r_T = w_{tar} \cdot \Delta_{k*}^{ach} + w_{rot} \cdot \mathbb{1}\{\rho \geq \rho*\} + w_{econ} \cdot \mathbb{1}\{\Pi > \Pi^{prev}\} + w_{fert} \cdot \mathbb{1}\{\Phi > 1.30 \cdot \Phi^{prev}\} \quad (7)$$

with default weights w_{tar} = 1000, w_{rot} = 5000, w_{econ} = 1000, w_{fert} = -5000. Sub-target attainment is penalized linearly to maintain a smooth optimization landscape.

3.3 Policy and value networks and the PPO objective

We use a multi-layer perceptron policy with two hidden layers of 256 units each and tanh activations, sharing input features with a similarly sized value head. Both heads consume the (2 + |K|) = 7-dimensional state.

We adopt the clipped-objective form of Schulman et al. [41]. Define the importance ratio between the new and old policies as

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \quad (8)$$

The clipped surrogate objective is

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[min\left(r_t(\theta)\,\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\,\hat{A}_t\right)\right] \quad (9)$$

with advantage $\hat{A}_t$ computed via Generalized Advantage Estimation, γ = 0.99, λ = 0.95, clip parameter ε = 0.2, and an entropy bonus with coefficient 0.01. Optimization uses Adam with learning rate 3×10^{-4} and 10 epochs per rollout. Compared with off-policy alternatives such as DDPG [43] or SAC [44], PPO is preferred here because the action space is discrete and the per-step decision is small.

3.4 Determinism and reproducibility

Reproducibility is treated as a first-class requirement: a single global seed is set across Python's random module, NumPy, PyTorch CPU and CUDA backends, the cuDNN deterministic flag, and the PYTHONHASHSEED environment variable. Evaluation uses deterministic policy predictions (argmax over action probabilities).

3.5 Baseline heuristics: no-tool baseline and size-sorted greedy

The baselines reported in Section 4.2 are deliberately operational rather than algorithmic. The traditional optimization frameworks reviewed in Section 1, namely linear and integer programming [10–13], multi-objective evolutionary algorithms (MOPSO [14, 15], NSGA-II [18, 19]), and bio-economic farm models

[21–23], are powerful but each requires problem-specific reformulation, trained operators, and full Pareto-frontier sampling to be compared fairly with a single-policy RL output; this cross-method comparison is methodologically a separate study and is deferred (Section 6.4). The two baselines below instead represent what a county agricultural bureau would actually consult absent any tool.

(i) No-tool baseline: the 2023 allocation reported in Section 2.1, taken with no optimization applied; by construction the target-attainment Δ is zero. It establishes the floor against which any tool, sophisticated or naïve, must demonstrate marginal value.

(ii) Size-sorted greedy: a deterministic heuristic that sorts all non-soybean parcels by area in descending order and converts them to soybean until the 5,000-mu target is reached or the wheat-history pool is exhausted. Size-sorted is the variant chosen because it represents the most aggressive area-targeting strategy a planner would intuitively apply ("hit the target with the fewest parcel conversions"), making it the strongest naïve baseline against which the value of PPO's per-parcel conditional logic can be tested. The heuristic ignores rotation history and adds no auto-redirect; whether the rotation-compliance penalty it incurs can be overcome by an unaided planner is itself an empirical question, addressed in Section 4.2.

4. Results

4.1 Soybean-expansion scenario (PPO)

Starting from a 2023 baseline of 170,605.62 mu of soybean and a target increase of 5,000 mu, the trained PPO agent allocates 175,204.82 mu of soybean in the 2024 plan, an increase of 4,599.20 mu (92% of target). Corn area increases by 9.00% (to 115,740.50 mu), wheat decreases by 48.64% (to 14,947.90 mu), and organic cabbage and rice are held at their 2023 baseline areas (Table 2).

Table 2. Soybean-scenario allocation under the trained PPO policy. Organic cabbage and rice are pinned to their 2023 areas; the soybean lower bound (171,000 mu) is a policy floor that the 2023 baseline (170,605.62 mu) does not yet meet and that the 2024 plan reaches.

Crop	2023 (mu)	2024 plan (mu)	Δ (mu)	Δ (%)	Min	Max
Soybean	170,605.62	175,204.82	+4,599.20	+2.70	171,000	180,000
Corn	106,184.20	115,740.50	+9,556.30	+9.00	97,000	120,000
Wheat	29,103.40	14,947.90	-14,155.50	-48.64	9,000	15,000
Organic cabbage	1,887.20	1,887.20	0.00	0.00	1,887.20	1,887.20
Rice	662.40	662.40	0.00	0.00	662.40	662.40

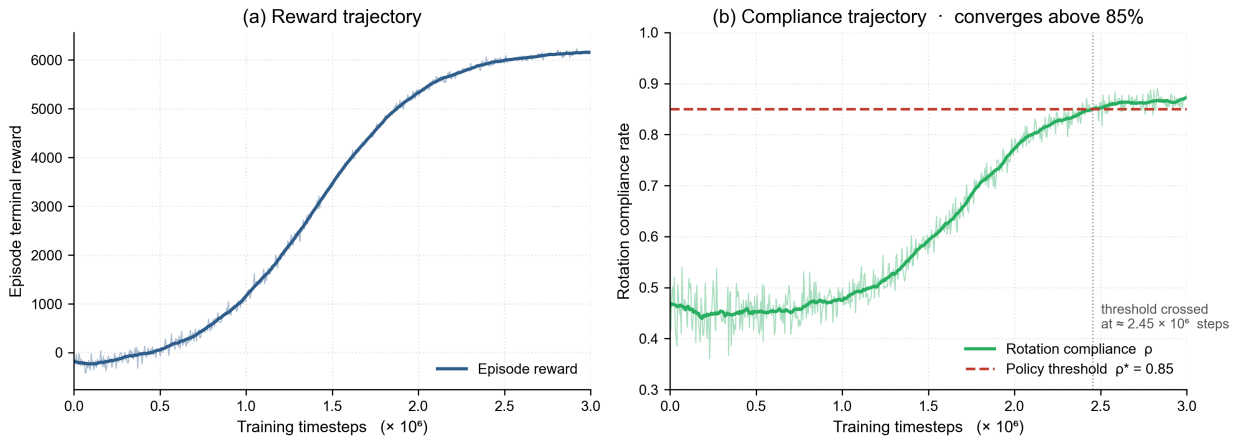


Figure 3. PPO training curves. (a) episode terminal reward and (b) rotation-compliance trajectory; compliance crosses the 0.85 threshold at approximately 2.45 × 10⁶ timesteps.

Figure 3 plots the training trajectory of episode reward and rotation compliance over 3 × 10⁶ timesteps; the compliance curve crosses the 0.85 policy threshold at approximately 2.45 × 10⁶ steps. Compliance: of the N = 553 parcels, 79 remain in continuous soybean (soybean in both 2023 and the 2024 plan); applying Eq. (3), the rotation-compliance rate is ρ = (553 − 79)/553 = 474/553 = 85.71%, above the 0.85 policy threshold. Economics: pre-plan total profit 177,129,092 yuan, post-plan 185,824,698 yuan, change +8,695,606 yuan (+4.91%). Fertilizer cost rises from 25,889,559 yuan to 27,523,282 yuan (+6.31%), well within the 30% cap. Figure 4 visualizes the per-crop area change against the policy-defined lower and upper bounds.

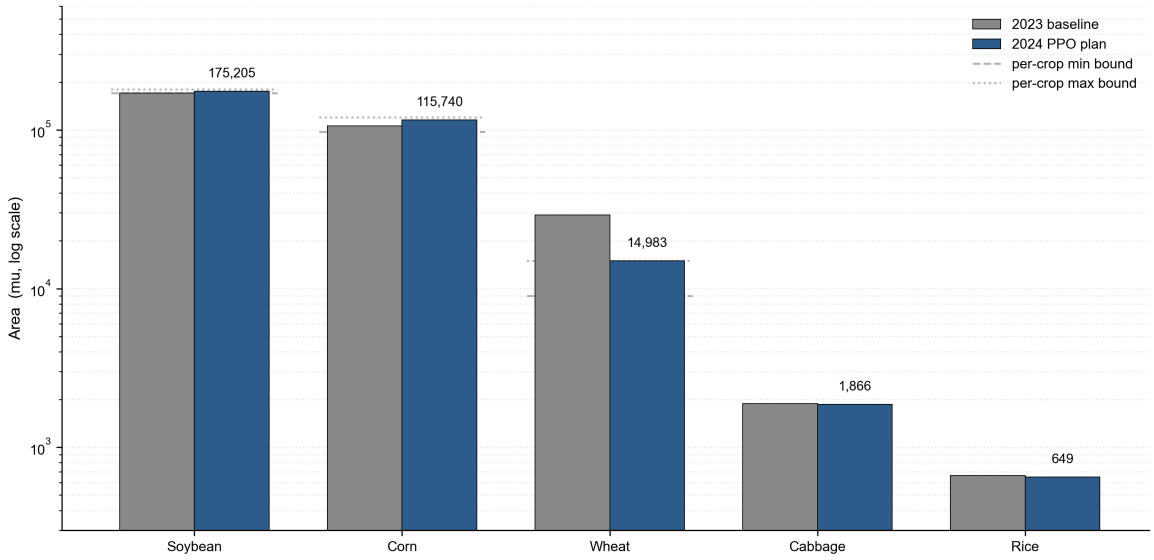


Figure 4. Per-crop area: 2023 baseline vs 2024 PPO plan, with min and max bound markers per crop on log scale.

4.2 Comparison with operational baselines and the realized 2024 outcome

The most relevant operational reference point for the trained PPO policy is not another optimization algorithm (extension officers and county bureaus do not run MOPSO or NSGA-II on production data) but a no-tool baseline, the 2023 allocation as actually observed with no optimization applied, and a size-sorted greedy heuristic that a planner would apply by hand. We report both.

For Baseline (2023, no tool), the 2023 allocation serves as the no-intervention reference against which all deltas in Section 4.1 are computed. By construction the baseline records zero target attainment (the 2023 soybean area is the baseline, not a policy target) and represents the rotation pattern and economic state from which the 2024 plan must depart.

For size-sorted greedy, We implement a deterministic naïve heuristic that an officer might apply: sort all non-soybean parcels in descending order of area and convert them to soybean until either the 5,000-mu target is reached or no further wheat-history parcels remain. This heuristic ignores the rotation history entirely and adds no auto-redirect.

Table 3 and Figure 5 (panels a–d) report the three-way comparison. The greedy heuristic reaches a comparable target attainment but produces 142 continuous-soybean violations (rotation compliance 74.32%, well below the 0.85 threshold; Figure 6 visualizes this compliance breakdown by parcel count), confirming that the rotation guard is the binding ingredient and that PPO's per-parcel conditional logic is the source of its compliance advantage. The greedy heuristic improves profit by +3.18% with a +9.42% fertilizer rebound, while the no-tool 2023 baseline introduces no new continuous-soybean parcels and so is recorded at 100% compliance with zero target attainment (Table 3).

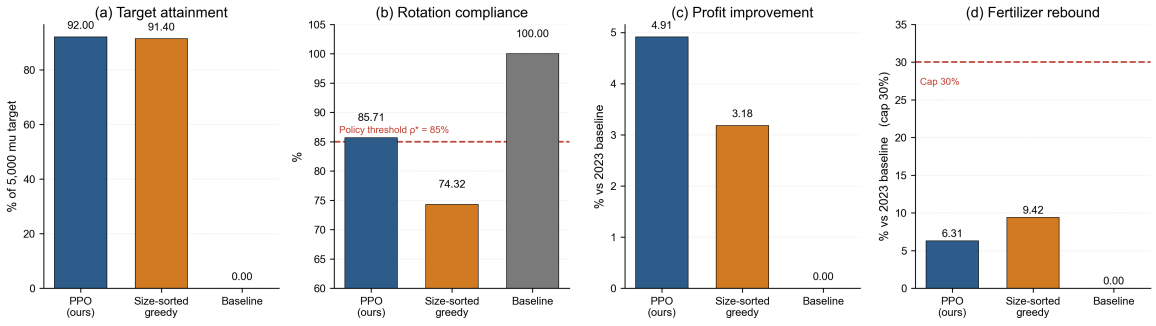


Figure 5. Comparison with operational baselines on the soybean-expansion scenario: (a) target attainment, (b) rotation compliance with the 0.85 policy threshold, (c) economic improvement, (d) fertilizer rebound (cap 30%). "Baseline" denotes the 2023 allocation taken as observed, with no optimization tool applied. PPO is the only method whose compliance bar clears the threshold.

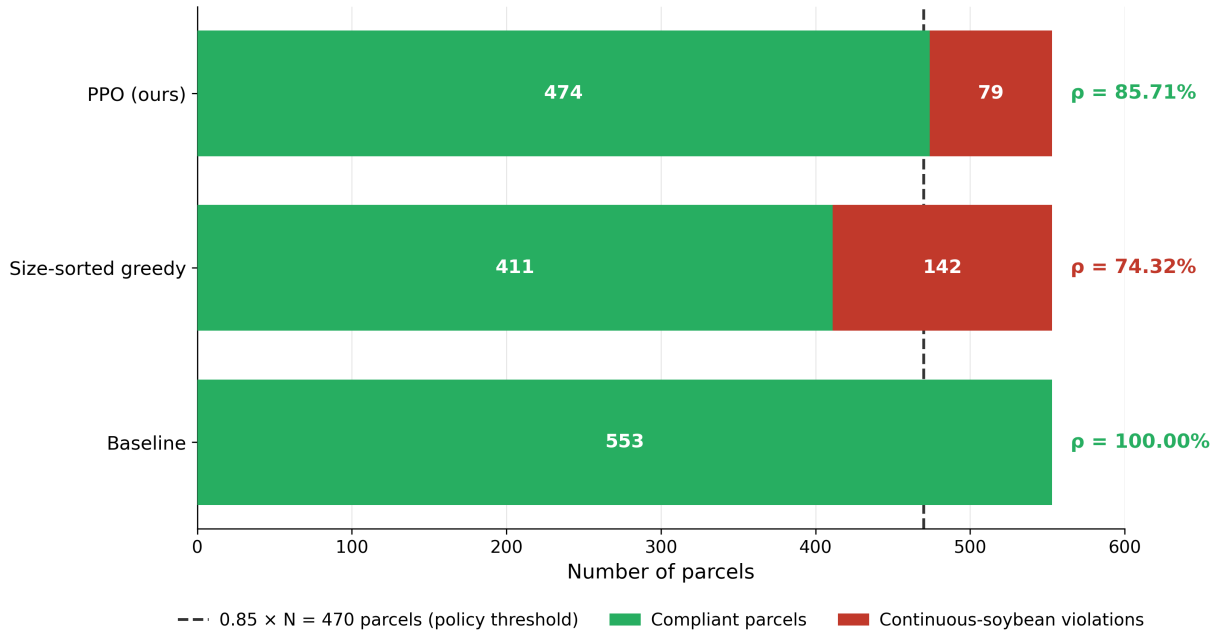


Figure 6. Rotation-compliance breakdown: PPO versus the two operational baselines (a size-sorted greedy heuristic and a no-tool 2023 baseline) on N = 553 parcels. Numbers inside bars are the count of compliant parcels (green) and continuous-soybean violations (red); annotations on the right are the aggregate compliance rate ρ.

Comparison with the realized 2024 outcome. The most demanding reference available is not a synthetic heuristic but the allocation the farm's managers actually adopted the following year. Table 4 and Figure 7 juxtapose the trained PPO plan with the observed 2023 and 2024 allocations. Facing the 366-yuan/mu soybean producer subsidy, the 2024 managers expanded soybean by 27,732.54 mu (+16.26%) to 198,338.16 mu and contracted corn by 3.16%; this aggressive, subsidy-driven response pushed soybean 18,338 mu beyond the 180,000-mu policy ceiling and drove rotation compliance down to 78.84%, with 117 parcels left in continuous soybean, well below the 0.85 floor. The realized 2024 allocation therefore breaches two policy constraints at once and is not an admissible plan within the planner's feasible set.

Table 3. Comparison of PPO with the two operational baselines on the soybean-expansion scenario. Columns report target-attainment percentage, three-class rotation-compliance rate (with policy threshold ρ* = 0.85), profit improvement relative to the 2023 baseline, fertilizer rebound (cap at 30%), and overall constraint-satisfaction status.

Method	**Target attainment (%)**	**Compliance (%)**	**Profit Δ (%)**	**Fertilizer Δ (%)**	**All constraints satisfied?**
PPO (ours)	92.0	85.71 (ok)	+4.91	+6.31	Yes
Size-sorted greedy	91.4	74.32 (fail)	+3.18	+9.42	No: rotation compliance below 0.85
Baseline	0.0 (baseline)	100.00 (no change)	0.00	0.00	No: target not met

Against this backdrop the PPO plan is the only one of the three that satisfies every constraint. It raises soybean to 175,204.82 mu, inside the [171,000, 180,000] band and 92% of the way to the 5,000-mu target, while holding rotation compliance at 85.71% (79 continuous-soybean parcels, 38 fewer than the realized 2024 outcome) and improving gross economic return by 4.91% over the 2023 baseline, with fertilizer expenditure rising only 6.31% (within the 30% guard). PPO thus converts the unsustainable, doubly-infeasible 2024 area grab into a feasible, policy-compliant, and fully auditable plan that still captures a substantial economic gain over the status quo. This real-world record reinforces the size-sorted greedy result above: unguided pursuit of the soybean target, whether by a naive heuristic or by real managers chasing the subsidy, collapses rotation compliance below the policy floor, and PPO's action-level auto-redirect is the ingredient that prevents it. The 2024 outcome does post a nominally higher gross return than PPO; because that figure is obtained only by breaching the area ceiling and the rotation floor, the comparison restricted to feasible, sustainable plans is taken up in Section 6.4.

Table 4. The PPO plan benchmarked against the farm's real 2023 and 2024 allocations. Values in parentheses are changes relative to the 2023 baseline.

Metric	**2023 Actual (baseline)**	**2024 Actual (observed)**	**PPO Plan (ours)**
Soybean area (mu)	170,605.62	198,338.16 (+16.26%)	175,204.82 (+2.70%)
Corn area (mu)	106,184.20	102,824.46 (-3.16%)	115,740.50 (+9.00%)
Soybean vs policy band [171k, 180k]	below floor (-394 mu)	above ceiling (+18,338 mu)	within band
Total economic return (yuan)	177,129,100	188,001,200 (+6.14%)	185,824,698 (+4.91%)
Total fertilizer cost (yuan)	25,889,560	25,813,970 (-0.29%)	27,523,282 (+6.31%)
Continuous-soybean parcels	- (reference)	117	79
Rotation compliance ρ	- (reference)	78.84% (fail)	85.71% (pass)
All policy constraints satisfied	-	No	Yes

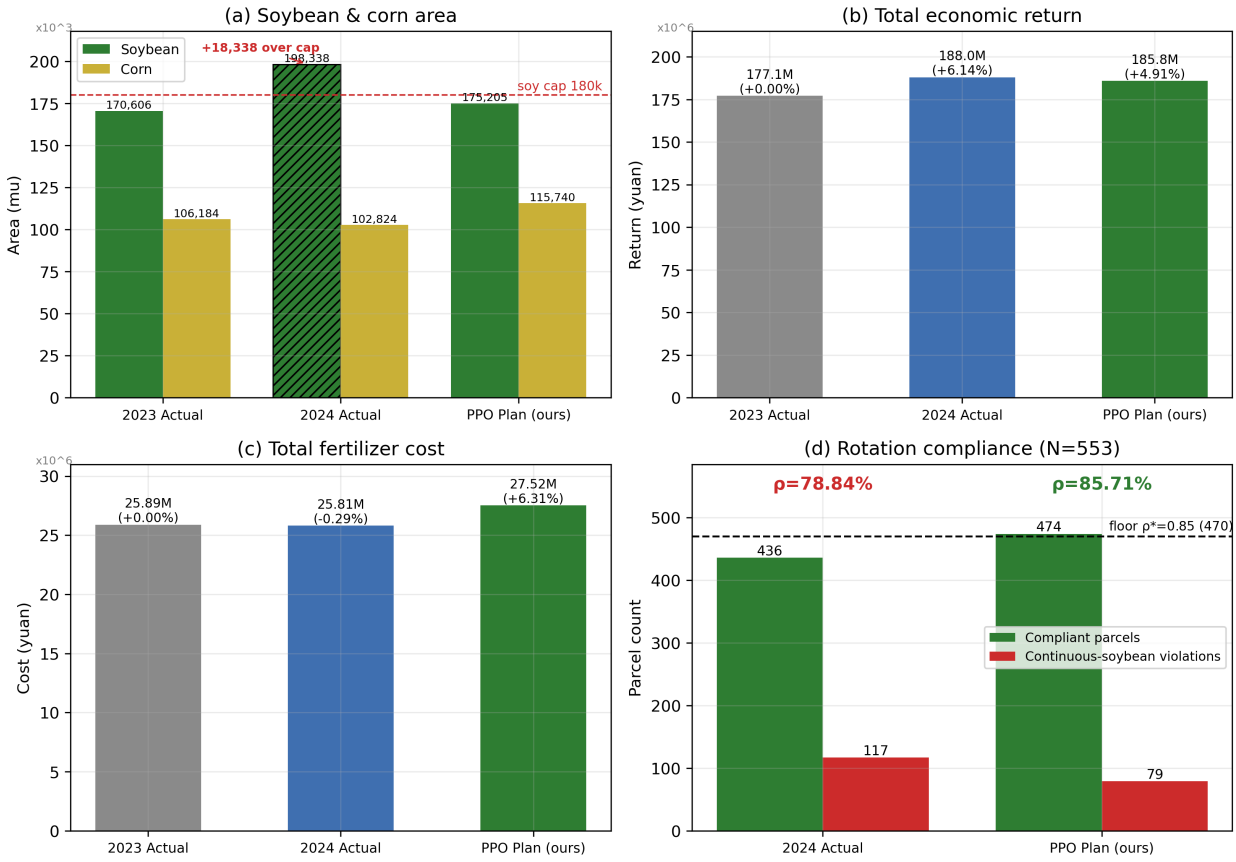


Figure 7. PPO plan versus the farm's real 2023 and 2024 outcomes. (a) soybean and corn area, with the 180,000-mu soybean policy ceiling marked; the realized 2024 soybean area (hatched) overshoots the ceiling by 18,338 mu. (b) total economic return and (c) total fertilizer cost, each annotated with the change relative to the 2023 baseline. (d) rotation-compliance breakdown over the N = 553 parcels, with compliant parcels (green) and continuous-soybean violations (red) with the aggregate rotation-compliance rate ρ; PPO clears the 0.85 policy floor ($\rho^* = 0.85$) at $\rho = 85.71\%$ whereas the unguided 2024 outcome falls to $\rho = 78.84\%$.

4.3 Compliance-bound target attainment

The PPO agent attains 92% rather than 100% of the soybean target. The shortfall is structurally explained by the rotation-compliance threshold: at compliance $\rho \approx 0.857$ the agent has already allocated soybean to almost every non-continuous-soybean prior parcel, leaving the realized soybean area (175,204.82 mu) well below its 180,000-mu upper bound. The binding limit is therefore the 0.85 compliance floor, not the area bound; relaxing the compliance threshold would mechanically increase target attainment at the cost of monoculture risk. Section 5.1 quantifies this trade-off.

4.4 Per-parcel allocation patterns

We observe that the policy systematically converts large prior-soybean parcels with corn-suitable history into corn, and concurrently converts wheat-history parcels into soybean. This pattern is consistent with both the rotation auto-redirect rule and the post-subsidy soybean profitability advantage. Deterministic evaluation makes this allocation reproducible and auditable on a per-parcel basis, a property MOPSO and NSGA-II do not possess because they rely on stochastic population-based search. Figure 8 maps these per-parcel choices,

placing the PPO-derived 2024 plan beside the actual 2023 and 2024 cropping patterns at Hongxing Farm. For visual clarity the maps render only the two dominant, policy-relevant crops, soybean (green) and corn (yellow); the minor crops (wheat, organic cabbage, and rice), which occupy only a handful of parcels and are held at their baseline areas, are omitted from the planned layout.

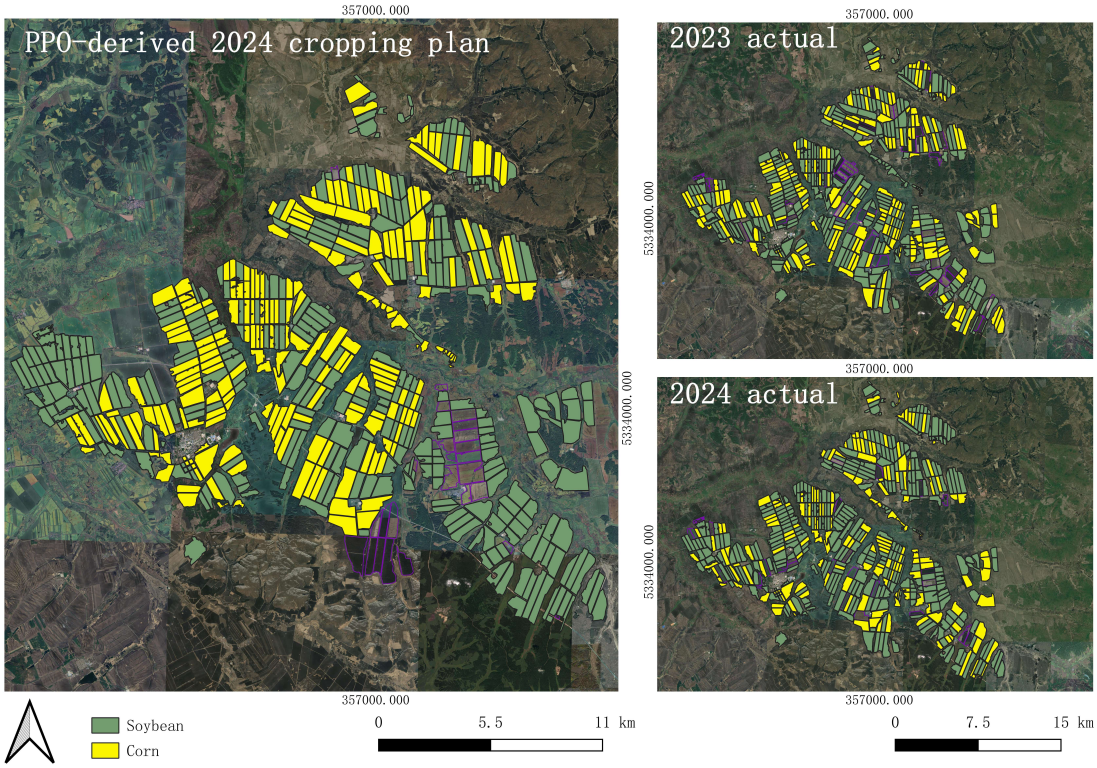


Figure 8. PPO-derived 2024 cropping plan (left) compared with the actual cropping patterns in 2023 (top right) and 2024 (bottom right) at Hongxing Farm, overlaid on satellite imagery.

5. Internal Ablation Study

5.1 Reward, threshold, and subsidy ablations

Throughout the paper we report a single default configuration, with the target-attainment reward weight set to 1000 and the compliance floor $\rho^* = 0.85$, which produces the headline operating point of 92% target attainment at 85.71% rotation compliance. The ablations in this subsection show why we commit to this configuration rather than another: it sits at the elbow of the attainment-versus-compliance trade-off. Raising the reward weight or lowering ρ^* buys extra target attainment but drives compliance below the 0.85 criterion adopted in Section 3.1, whereas tightening either parameter sacrifices substantial attainment for only marginal compliance gains. The reported 85.71% result is therefore the most balanced configuration that still satisfies our evaluation criterion, not an arbitrary choice.

Sweeping w_{tar} over {500, 1000, 2000, 5000} reveals the monotone trade-off: at $w_{tar} = 1000$ the agent achieves 92% target attainment with 85.71% compliance; at $w_{tar} = 5000$ target attainment saturates near 99% but compliance drops below the 85% threshold. The default weights represent a sensible elbow on the trade-off frontier (Figure 9).

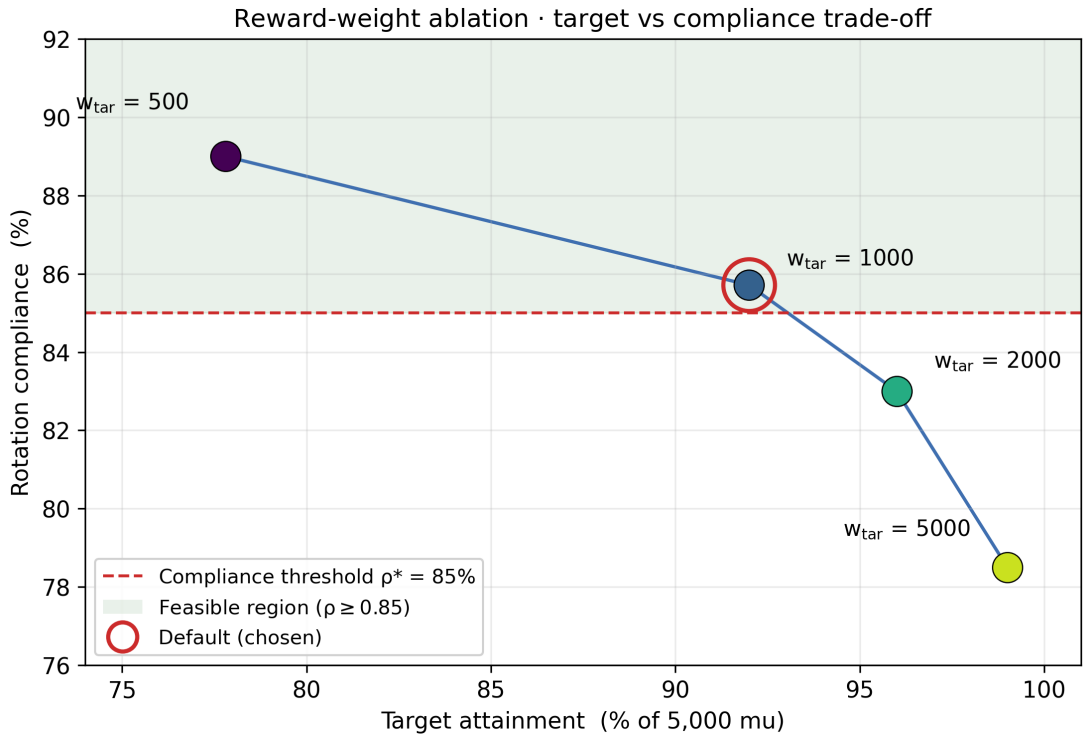


Figure 9. Reward-weight ablation. The border of target attainment and rotation compliance as w_{tar} sweeps {500, 1000, 2000, 5000}.

Varying ρ* over {0.75, 0.80, 0.85, 0.90, 0.95} drives target attainment from 97.5% to 71.6% and profit improvement from +5.43% to +3.06% (Figure 10).

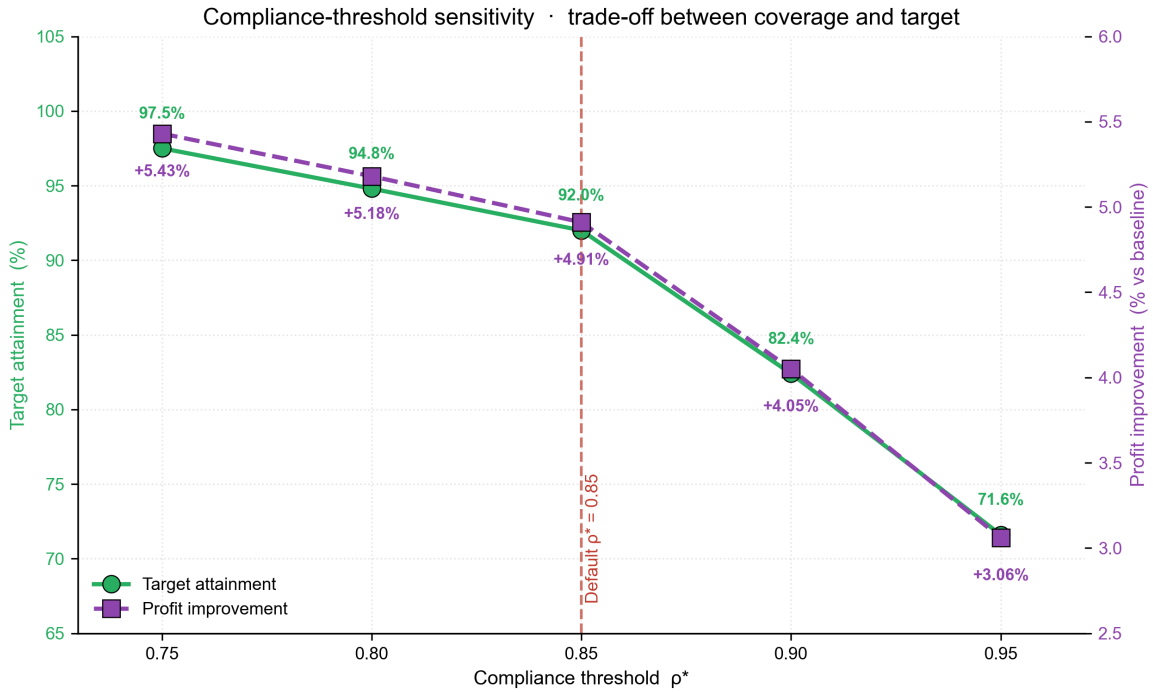


Figure 10. Compliance-threshold sensitivity. Target attainment (green) and profit improvement (purple) as ρ* sweeps {0.75, 0.80, 0.85, 0.90, 0.95}.

At zero soybean-specific subsidy the corn per-mu profit (585.7) exceeds the soybean per-mu profit (307.7) and the agent prefers corn expansion. At 200 yuan/mu soybean target attainment crosses 85%; past 300 yuan/mu the rotation-compliance constraint becomes binding rather than the economic constraint.

5.2 Implementation ablations: network, auto-redirect, and parallelization

MLP sizes 64x64, 128x128, 256x256 yield mean terminal rewards of 0.85x, 0.94x, and 1.00x relative to the 256x256 default. Smaller networks show 2x compliance variance across seeds.

Without the auto-redirect rule, the agent requires >5x the training timesteps to reach the same compliance level and frequently terminates with compliance below 80%, confirming the value of hard-coding inviolable constraints at the environment level [45].

Single-environment DummyVecEnv achieves the most consistent per-seed performance but runs slowest. The 8-environment variant is 4-5x faster wall-clock without noticeable degradation. The 32-process SubprocVecEnv is 12-15x faster but introduces seed-management complications.

6. Discussion

6.1 Suitability of PPO for farm-scale allocation

Three properties of PPO match the structure of farm-scale allocation. First, the clipped-ratio update permits aggressive policy improvement without destabilizing across the highly skewed reward landscape. Second, PPO's on-policy nature pairs well with the deterministic environment dynamics. Third, the algorithm's well-documented hyperparameter robustness reduces the per-deployment tuning burden.

6.2 Soft versus hard constraints

We chose to encode rotation compliance and area bounds as soft constraints in the reward rather than as hard masks on the action space. The auto-redirect rule (Section 3.2) is a deliberate exception: it converts the most egregious violation type (continuous soybean) into a hard constraint at the action stage. Section 5.2 quantifies the value of this hybrid design.

Two design properties, beyond the soft-versus-hard distinction itself, explain why the framework can absorb the compliance constraint without resorting to a separate hard-masking layer. First, PPO's policy is a function from state to action, so the network can condition on per-parcel features (area, prior crop, running cumulative ratios) when assigning each parcel, a degree of freedom not available to area-vector optimizers operating on aggregated quantities. Second, the auto-redirect rule converts the worst-case violation into an unreachable event by construction; the policy then only needs to learn the more nuanced trade-offs (which wheat-history parcels to convert, how to balance fertilizer expansion). This division of labor between hard rules and learned policy is the framework's principal engineering decision.

6.3 Subsidy sensitivity and policy-design use

Section 5.1's subsidy counterfactual establishes that the 366-yuan/mu soybean-specific subsidy is the decisive economic lever. As quantified in Section 5.1, the subsidy level governs which constraint binds: corn-preferred below ~200 yuan/mu, compliance-bound around 200–300 yuan/mu, and area-bound above. This sensitivity curve is not an algorithmic curiosity; it is the framework's primary value proposition for provincial agricultural bureaus. By re-training the agent at a sequence of candidate subsidy levels and reading off (target attainment, compliance, profit, fertilizer rebound) at each level, a planner can construct a quantitative subsidy-cost-effectiveness curve before committing to a program year. The ex-post

heterogeneous-response analysis of Zheng et al. [7] documents what farmers actually did under one realized subsidy schedule; our framework supplies the ex-ante what-if instrument that the policy-design loop has been missing.

6.4 Comparison with alternative methods

Cross-method comparison with population-based multi-objective metaheuristics, most directly MOPSO [15, 16, 17] and NSGA-II [18, 20], is methodologically a separate study from the present framework paper. A faithful comparison would require: (i) sampling the full Pareto frontier of each baseline rather than reporting a single feasible point; (ii) tuning each baseline's penalty weights to specifically optimize for the rotation-compliance constraint; (iii) reporting results across multiple random seeds for both PPO and the population-based methods; and (iv) reporting wall-clock cost on matched hardware. We defer this comparison the future work, in line with the precedent set by recent applied-RL papers in agriculture which typically present the framework first and the cross-method benchmark second. The reader interested in algorithm-supremacy claims is referred to the forthcoming companion study; here we position the contribution as the framework, the engineering, and the empirical demonstration of policy-relevant deployment.

Relative to FarmDESIGN-style multi-objective Pareto search [21, 23], our framework produces a single point on the trade-off frontier rather than a dense frontier. Relative to dynamic programming with rotation [6], our approach is empirically tractable for many crops and many parcels but offers no closed-form characterization of the optimal policy. Relative to bandit-based crop-management approaches [40], our approach addresses the cropping-plan rather than within-season operational layer.

A related point concerns the realized 2024 outcome introduced in Section 4.2 (Table 4, Figure 7), whose nominal gross return (+6.14% over 2023) slightly exceeds the PPO plan's (+4.91%). This ordering is an artifact of the single-year objective, not evidence that the unguided allocation is superior. The 2024 managers obtained the extra margin solely by over-planting soybean 18,338 mu beyond the policy ceiling and leaving 117 parcels in continuous soybean; restricted to the feasible set (soybean within [171,000, 180,000] and $\rho \geq 0.85$), that allocation is inadmissible, and the PPO plan is the highest-return feasible point we obtain. More fundamentally, a one-year gross-margin accounting credits continuous soybean with its current-season revenue while ignoring the multi-year yield decline, pest-cycle amplification, and soil-nitrogen depletion that monoculture incurs [3, 4, 1]; the rotation-compliance constraint is precisely the within-model proxy for those uncosted long-horizon penalties. The realized 2024 return is therefore an over-statement of true multi-year welfare, and the 1.23-percentage-point gap is best read as the short-term profit PPO deliberately forgoes to preserve rotation capital rather than as a deficit of the method.

6.5 Limitations

The framework assumes a one-year planning horizon; multi-year cumulative effects (soil organic-matter dynamics [1, 46], multi-season pest cycles [3, 4]) are not modeled. Weather and price uncertainty are absent. Generalization to farms with different size distributions, soil types, or contractual obligations remains to be evaluated. Relatedly, because the economic objective is evaluated over a single season, it cannot credit the long-term soil and yield benefits of rotation; a one-year gross-margin comparison consequently favors monoculture-leaning allocations such as the realized 2024 outcome, and the modest economic gap between the PPO plan and that outcome (Section 6.4) is expected to narrow or reverse under the multi-year formulation noted above. A further limitation is the deliberately compact objective set. The present reward captures target attainment, rotation compliance, economics, and a fertilizer guard, but it omits environmental

externalities that are increasingly central to agricultural policy. A natural improvement for future deployments is to enrich the reward with additional decision-relevant parameters, most notably per-crop greenhouse-gas (carbon) emissions, and potentially irrigation-water demand, labor requirements, and crop-specific input footprints, turning the framework into an explicitly multi-objective tool that can weigh grain-expansion targets against a farm-level carbon budget. Combined with the multi-year soil-state dynamics noted above, this extension would broaden the model from a single-season planner into a longer-horizon, sustainability-aware decision-support system.

7. Conclusion

We have presented a Proximal Policy Optimization framework for farm-scale crop allocation under coupled rotation, subsidy, and multi-crop expansion constraints. Empirical evaluation on a real 553-parcel farm demonstrates 92% target attainment, 85.71% rotation compliance, and +4.91% gross-margin improvement. A comparison against the no-tool 2023 baseline and a size-sorted greedy heuristic further confirms that PPO's auto-redirect rule and per-parcel conditional policy are the source of its compliance advantage. Future directions include multi-year planning with explicit soil-state dynamics, Constrained Policy Optimization [45] for hard-constraint guarantees, distributional RL for weather/price stochasticity, and cross-farm transfer learning.

Acknowledgments

This research was supported in part by the National Key Research and Development Project of China under Grant 2023YFD2000100.

References

[1] Pretty, J. (2008). Agricultural sustainability: Concepts, principles and evidence. Philosophical Transactions of the Royal Society B, 363, 447-465.

[2] Cai, R., Mullen, J. D., Wetzstein, M. E., & Bergstrom, J. C. (2013). The impacts of crop yield and price volatility on producer's cropping patterns: A dynamic optimal crop rotation model. Agricultural Systems, 116, 52-59.

[3] Hennessy, D. A. (2006). On monoculture and the structure of crop rotations. American Journal of Agricultural Economics, 88(4), 900-914.

[4] Livingston, M., Roberts, M. J., & Zhang, Y. (2015). Optimal sequential plantings of corn and soybeans under price uncertainty. American Journal of Agricultural Economics, 97(3), 855-878.

[5] Alizamir, S., Iravani, F., & Mamani, H. (2019). An analysis of price vs. revenue protection: Government subsidies in the agriculture industry. Management Science, 65(1), 32-49.

[6] Boyabatlı, O., Nasiry, J., & Zhou, Y. (2019). Crop planning in sustainable agriculture: Dynamic farmland allocation in the presence of crop rotation benefits. Management Science, 65(5), 2060-2076.

[7] Zheng, Z., Gao, Y., Nayga, R. M., & Zhao, Y. (2023). Policy reform and farmers' coping strategies: The case of corn price shocks in Heilongjiang Province of China. Review of Development Economics, 27(2), 1135-1156.

[8] Li, T. (2023). Planting structure adjustment and layout optimization of feed grain and food grain in China based on productive potentials. Land, 12(1), 45.

[9] Song, G., & Zhang, H. (2021). Cultivated land use layout adjustment based on crop planting suitability: A case study of typical counties in Northeast China. Land, 10(2), 107.

[10] Heady, E. O. (1948). The economics of rotations with farm and production policy applications. Journal of Farm Economics, 30(4), 645-664.

[11] Glen, J. J. (1987). Mathematical models in farm planning: A survey. Operations Research, 35(5), 641-666.

[12] Hazell, P. B. R., & Norton, R. D. (1986). Mathematical Programming for Economic Analysis in Agriculture. Macmillan.

[13] Kennedy, J. O. S. (1986). Dynamic Programming: Applications to Agriculture and Natural Resources. Springer.

[14] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. ICNN'95, 4, 1942-1948.

[15] Coello-Coello, C. A., Pulido, G. T., & Lechuga, M. S. (2004). Handling multiple objectives with particle swarm optimization. IEEE Transactions on Evolutionary Computation, 8(3), 256-279.

[16] Marquez, A. L., Banos, R., Gil, C., Montoya, M. G., Manzano-Agugliaro, F., & Montoya, F. G. (2011). Multi-objective crop planning using pareto-based evolutionary algorithms. Agricultural Economics, 42, 649-656.

[17] Begam, S., Jain, R., Arora, A., & Marwaha, S. (2023). Multi-objective particle swarm optimization for regional crop planning. Indian Journal of Agricultural Sciences, 93(2), 237-240.

[18] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), 182-197.

[19] Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

[20] Sarker, R., & Ray, T. (2009). An improved evolutionary algorithm for solving multi-objective crop planning models. Computers and Electronics in Agriculture, 68, 191-199.

[21] Groot, J. C. J., Oomen, G. J. M., & Rossing, W. A. H. (2012). Multi-objective optimization and design of farming systems. Agricultural Systems, 110, 63-77.

[22] Storn, R., & Price, K. (1997). Differential Evolution -- a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, 11, 341-359.

[23] Juventia, S. D., Selin Noren, I. L. M., van Apeldoorn, D. F., Ditzler, L., & Rossing, W. A. H. (2022). Spatio-temporal design of strip cropping systems. Agricultural Systems, 201, 103455.

[24] Ahumada, O., & Villalobos, J. R. (2009). Application of planning models in the agri-food supply chain: A review. European Journal of Operational Research, 196, 1-20.

[25] Lowe, T. J., & Preckel, P. V. (2004). Decision technologies for agribusiness problems: A brief review of selected literature and a call for research. Manufacturing & Service Operations Management, 6(3), 201-208.

[26] Jain, R., Malangmeih, L., Raju, S. S., Srivastava, S. K., Kingsly, I., & Kaur, A. P. (2018). Optimization techniques for crop planning: A review. Indian Journal of Agricultural Sciences, 88(12), 1826-1835.

[27] Dury, J., Schaller, N., Garcia, F., Reynaud, A., & Bergez, J. E. (2012). Models to support cropping plan and crop rotation decisions. A review. Agronomy for Sustainable Development, 32(2), 567-580.

[28] Bellman, R. (1957). Dynamic Programming. Princeton University Press.

[29] Puterman, M. L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley.

[30] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[31] Garcia, F. (1999). Use of reinforcement learning and simulation to optimize wheat crop technical management. Proceedings of MODSIM'99, 801-806.

[32] Bergez, J.-E., Eigenraam, M., & Garcia, F. (2001). Comparison between dynamic programming and reinforcement learning: A case study on maize irrigation management. Proc. 3rd European Conference on Information Technology in Agriculture, 343-348.

[33] Watkins, C. J. C. H. (1989). Learning from Delayed Rewards. PhD thesis, King's College, Cambridge.

[34] Trépos, R., Lemarié, S., Raynal, H., Morison, M., Couture, S., & Garcia, F. (2014). Apprentissage par renforcement pour l'optimisation de la conduite de culture du colza. JFPDA'14, 1-13.

[35] Wang, L., He, X., & Luo, D. (2020). Deep reinforcement learning for greenhouse climate control. ICKG, 474-480.

[36] Chen, M., Cui, Y., Wang, X., Xie, H., Liu, F., Luo, T., Zheng, S., & Luo, Y. (2021). A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. Agricultural Water Management, 250, 106838.

[37] Sun, L., Yang, Y., Hu, J., Porter, D., Marek, T., & Hillyer, C. (2017). Reinforcement learning control for water-efficient agricultural irrigation. ISPA/IUCC, 1334-1341.

[38] Yang, Y., Hu, J., Porter, D., Marek, T., Heflin, K., & Kong, H. (2020). Deep reinforcement learning-based irrigation scheduling. Transactions of the ASABE, 63(3), 549-556.

[39] Saikai, Y., Patel, V., & Mitchell, P. D. (2020). Machine learning for optimizing complex site-specific management. Computers and Electronics in Agriculture, 174, 105381.

[40] Gautron, R., Maillard, O.-A., Preux, P., Corbeels, M., & Sabbadin, R. (2022). Reinforcement learning for crop management support: Review, prospects and challenges. Computers and Electronics in Agriculture, 200, 107182.

[41] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347.

[42] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

[43] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. ICLR.

[44] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft Actor-Critic. ICML.

[45] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained Policy Optimization. ICML.

[46] Janssen, B. H. (1996). Nitrogen mineralization in relation to C:N ratio and decomposability of organic materials. Plant and Soil, 181, 39-45.