Transformer-Enhanced Soft Actor-Critic with EV-Aware Reward Shaping for Maize Optimization

Xuan Lim[1], Hock Guan Goh[2], Shen Khang Teoh[3], Peh Chiong Teh[4], Ivan Andonovic[5]
Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia[1, 2, 3]
Faculty of Engineering and Green Technology, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia[4]
Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom[5]

Abstract—**Optimizing fertilization and irrigation strategies is essential for improving productivity and resource efficiency in precision agriculture. Artificial intelligence (AI), particularly reinforcement learning (RL), has been increasingly explored for adaptive crop management under uncertain environmental conditions. However, many existing approaches rely on single-action formulations that struggle with joint input control, leading to economically unstable outcomes and limited policy interpretability. This study proposes a Transformer-enhanced Soft Actor-Critic (SAC) framework with expected value (EV)-aware reward shaping for maize optimization in a Decision Support System for Agrotechnology Transfer (DSSAT) Gym environment, enabling simultaneous control of fertilization and irrigation under dynamic crop-environment interactions. Unlike standard SAC implementations, the proposed framework incorporates a transformer-based state encoder for richer agronomic state representation and an EV-aware reward shaping mechanism to guide economically stable long-horizon decision-making. The proposed AI-driven approach improves economic profitability and profit stability compared with the prior state-of-the-art (SOTA) large language model (LLM)-enhanced Deep Q-Network (DQN) baseline. Behavioral analysis shows that the learned policy exhibits temporally structured decision patterns characterized by smaller-magnitude, higher-frequency actions and an associated input-efficiency trade-off. Furthermore, Shapley Additive Explanations (SHAP)-based explainable AI (XAI) analysis identifies growth-stage and crop-development variables as dominant drivers of long-horizon control decisions. Overall, the results demonstrate that the Transformer-enhanced SAC with EV-aware reward shaping provides a more profitable, financially stable, and interpretable AI-based decision-making framework for maize optimization in the DSSAT Gym environment.**

Keywords—Precision agriculture; maize optimization; fertilization and irrigation management; reinforcement learning; Soft Actor-Critic; transformer; reward shaping; explainable artificial intelligence

I. INTRODUCTION

Maize is one of the world's most important crops, serving as a staple food and playing a critical role in livestock feed and biofuel production. By 2025, maize cultivation spans approximately 197 million hectares globally, with total production reaching 1.27 billion tonnes, making it the most widely grown cereal crop worldwide [1], [2]. As global demand rises, optimizing crop management practices is essential to maintain high yields while ensuring resource efficiency and environmental sustainability. Fertilizer and irrigation inputs significantly influence maize productivity. However, excessive or inefficient application leads to resource waste, increased costs, and environmental degradation.

Traditional crop management relies on static schedules and predefined heuristics, which lack flexibility under real-world variability. Weather patterns, soil heterogeneity, and irregular rainfall introduce substantial uncertainty into crop growth processes. Consequently, rule-based systems often result in inefficient resource allocation and reduced robustness.

Recent advances in AI, particularly RL, have enabled adaptive decision-making through interaction with complex environments. RL has been increasingly applied to fertilization and irrigation scheduling; however, many approaches rely on single-action formulations, limiting their ability to jointly optimize multiple agronomic inputs and often resulting in economically unstable outcomes. Prior studies also emphasize aggregate performance metrics such as yield or profit, with limited analysis of learned policy behavior and profit stability.

Transformer-based models extend RL systems by providing powerful representations of sequential and contextual information, and have demonstrated strong potential through architectures such as the Decision Transformer and other transformer-RL frameworks [3], [4]. Originally developed for natural language processing, transformer architectures model long-range dependencies and temporal patterns via self-attention mechanisms [5]. When applied to agriculture, transformers enhance state representations by capturing complex crop-environment interactions over time. Integrating transformer encoders with actor-critic RL algorithms enables context-aware control policies better suited for dynamic and uncertain agricultural environments.

Building on these advances, this study proposes a Transformer-enhanced SAC framework with EV-aware reward shaping for maize optimization in a DSSAT Gym environment. The model enables simultaneous control of fertilization and irrigation decisions using Transformer-based state encoding, with EV-aware reward shaping that prioritizes long-term stable returns. This design reduces profit variance and improves average economic performance.

Beyond performance improvement, this work provides a multi-stage analysis of learned policies. The proposed actor-critic framework is evaluated against the prior SOTA LLM-enhanced DQN baseline. Behavioral analysis examines action magnitude distribution, decision frequency, and input-efficiency trade-offs, while SHAP-based XAI analysis identifies growth-stage and environmental variables driving long-term control decisions.

By jointly addressing profitability, profit stability, and interpretability, this work advances RL-based crop management beyond static decision-making and opaque black-box optimization.

The main contributions are threefold: 1) A Transformer-enhanced SAC framework with EV-aware reward shaping for joint fertilization and irrigation control in a DSSAT Gym environment, demonstrating improved economic profitability and profit stability relative to the SOTA LLM-enhanced DQN baseline. 2) A detailed behavioral analysis revealing a consistent shift toward smaller-magnitude, higher-frequency actions and input-efficiency trade-offs. 3) A SHAP-based XAI pipeline identifying growth-related environmental and crop variables as dominant drivers of long-term fertilization and irrigation decisions.

The remainder of this study is organized as follows: Section II reviews related work on reinforcement learning for crop management and DSSAT-based optimization. Section III describes the proposed methodological framework, including the transformer-based state representation, actor-critic architecture, and EV-aware reward shaping mechanism. Section IV presents the experimental configuration and implementation details. Section V reports the experimental results and analysis across economic performance, policy behavior, and explainability evaluation. Finally, Section VI concludes the study and discusses future research directions.

II. Related Work

A. DSSAT as a Simulation Platform for RL-Based Crop Management

DSSAT is a widely validated crop simulation framework extensively used in agricultural research and agronomic decision support due to its validated physiological foundations and ability to simulate crop growth, development, and yield formation under diverse environmental and management conditions. Its long-standing application across crop types and geographic regions has established DSSAT as a reliable platform for evaluating crop management strategies and conducting controlled simulation-based studies.

In recent years, DSSAT has been increasingly adopted as a simulation environment for RL-based crop management. Beyond DSSAT, several RL-enabled crop simulation environments have emerged, supporting adaptive policy learning under stochastic environmental settings [6-8]. Within these frameworks, agricultural decision-making is naturally formulated as a Markov Decision Process (MDP), where the crop-environment state evolves over time and agents select fertilization and irrigation actions to maximize long-term agronomic or economic returns.

Integrating DSSAT with RL offers advantages over field trials, which are costly, time-consuming, and subject to uncontrollable variability. DSSAT enables large-scale, repeatable experimentation under controlled yet agronomically realistic conditions, allowing RL agents to explore a wide range of management strategies without real-world risk. This capability is particularly important for studying adaptive fertilization and irrigation under environmental uncertainty and delayed effects.

Recent developments further enhance DSSAT's suitability for RL research through interactive simulation interfaces such as DSSAT Gym [9]. These interfaces provide standardized observation, action, and reward application programming interfaces (APIs), enabling seamless integration with modern RL algorithms. By exposing crop and environmental states at each time step and allowing dynamic action application throughout the growing season, DSSAT Gym bridges traditional rule-based modeling with data-driven adaptive control, establishing DSSAT-based RL frameworks as a foundational benchmark for intelligent crop management strategies.

B. Prior Work on RL-Based Fertilization and Irrigation Control

RL has increasingly been adopted for agricultural decision-making, enabling adaptive fertilization and irrigation strategies that improve agronomic outcomes under uncertainty. Prior studies have applied RL to support crop management tasks such as nitrogen management, resource allocation, and decision support planning, demonstrating its potential to improve agricultural productivity and sustainability [10].

A substantial body of research has focused on irrigation optimization, reporting improvements in water-use efficiency, adaptive scheduling, and irrigation control under uncertain and variable field conditions [11-14]. Beyond irrigation-only applications, RL has also been employed for joint fertilizer-irrigation management, incorporating climate uncertainty and environmental considerations such as nitrous oxide gas emissions while improving economic and resource outcomes [15], [16]. RL has also been integrated into agricultural digital-twin environments to analyze and deploy management strategies across irrigation, greenhouse control, and crop production systems [17]. Comparative benchmarking against classical control methods such as Model Predictive Control (MPC) further demonstrates RL's competitiveness in dynamic agricultural decision-making [18].

Early RL applications in crop management evolved from neural network models toward more advanced decision-making frameworks. A notable study combined DQN with imitation learning (IL) to optimize nitrogen fertilization and irrigation scheduling within DSSAT [19], demonstrating measurable economic gains compared with rule-based baselines. Experiments conducted in Florida and Zaragoza reported up to 45% and 55% improvements in economic profit, respectively, while reducing environmental impact through lower nitrogen leaching. The rule-based baselines followed practical agronomic standards derived from regional maize production guides and farmer survey data [20-22].

Building on these foundations, subsequent work introduced transformer-enhanced RL frameworks leveraging LLMs for improved state representation within DSSAT [23]. In this approach, DSSAT state variables were converted into natural language descriptions and encoded using DistilBERT, with embeddings provided to a DQN agent. This representation-learning strategy improved training stability, data efficiency, and economic performance across benchmark environments, achieving profit gains of 49% in Florida and 67% in Zaragoza relative to conventional baselines. These findings highlight the potential of LLM-based representations to enhance generalization and adaptability in crop management policies.

Collectively, prior studies demonstrate the effectiveness of RL-based approaches for fertilization and irrigation control within simulation environments. The progression from rule-based strategies to learning-based decision-making, and more recently to LLM-enhanced representations, establishes the LLM-enhanced DQN framework as a strong SOTA benchmark for DSSAT-based crop management.

C. Limitations of Existing Approaches

Despite promising progress in integrating DSSAT with RL, several important limitations remain. First, most prior work adopts single-action formulations, typically based on DQN, which inherently support only discrete action outputs. This design is ill-suited for simultaneous fertilizer and irrigation control in complex agricultural settings. To address this constraint, earlier studies discretized handcrafted action mappings that convert continuous inputs into fixed fertilizer-irrigation pairs representing coarse nitrogen and water levels [23]. While functional, this discretization restricts policy expressiveness, limits decision granularity, and lacks a principled mechanism for capturing nuanced input trade-offs.

Second, existing studies often employ weak evaluation protocols that inadequately assess robustness. Models are frequently evaluated under the same environmental conditions used during training, with conclusions based on a single best-performing policy. Performance variance and profit stability are rarely reported, leaving the reliability of learned policies under realistic agricultural uncertainty insufficiently examined.

Third, prior work largely focuses on aggregate performance metrics such as average yield or economic profit, with limited investigation into policy behavior or decision dynamics. Analyses of action magnitude distributions, temporal decision patterns, and input-efficiency trade-offs are typically absent. Moreover, the lack of explainability mechanisms prevents meaningful interpretation of how environmental and crop-growth variables influence long-horizon decisions. This black-box optimization limits transparency and practical adoption, particularly in real-world agricultural systems where economic stability and interpretability are critical.

Collectively, these limitations highlight the need for RL frameworks that support joint multi-input decision-making, adopt robust evaluation protocols, and integrate behavioral and explainable analyses to ensure policies are not only profitable but also stable and interpretable in real-world deployment.

III. Methodology

A. Problem Formulation in DSSAT Gym Environment

Maize fertilization and irrigation scheduling are formulated as a sequential decision-making problem under uncertainty, where management actions influence crop growth, yield formation, and economic outcomes across an entire growing season. The problem is modeled as a MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P})$, where $\mathcal{S}$ denotes the crop-environment state space, $\mathcal{A}$ represents management actions, $\mathcal{R}$ is the reward function, and $\mathcal{P}$ captures the environment transition dynamics.

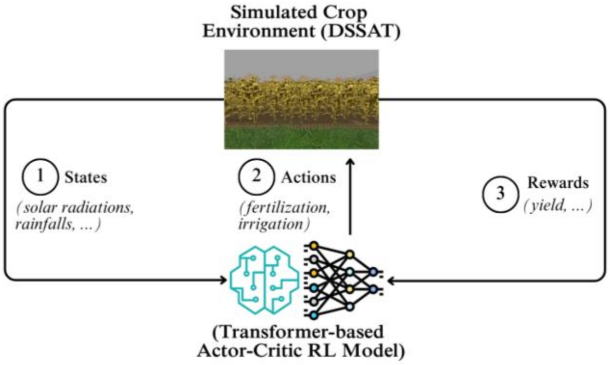
Simulated Crop Environment (DSSAT)
1 States
(solar radiations, rainfalls, ...)
2 Actions
(fertilization, irrigation)
3 Rewards
(yield, ...)
(Transformer-based Actor-Critic RL Model)

Fig. 1. System architecture of the transformer-enhanced actor-critic RL framework within the DSSAT Gym crop simulator.

Fig. 1 illustrates the overall architecture of the proposed transformer-enhanced actor-critic reinforcement learning framework. The DSSAT Gym provides environmental dynamics through a standardized RL interface. At each time step, the agent observes a structured multi-dimensional state comprising crop growth indicators, accumulated management inputs, soil moisture profiles, environmental conditions, and stress variables. Based on this state, the agent selects fertilization and irrigation actions. The DSSAT simulator advances the crop system and returns the next state and reward, forming a closed-loop interaction that captures delayed effects, seasonal dependencies, and cumulative impacts inherent to agricultural systems. Formally, the instantaneous reward at time step t is defined as:

$$r_t(s_t, a_t) = \{ \begin{matrix} w_1 Y - w_2 N_t - w_3 W_t, & \text{ιφ ηαρϖεστ ατ τ,} \\ -w_2 N_t - w_3 W_t, & \text{οτηερωισε,} \end{matrix} \tag{1}$$

where, $r_t(s_t, a_t)$ denotes the reward at time step t given state s_t and action a_t. The term Y denotes final crop yield (kg/ha), N_t is nitrogen fertilizer (kg/ha), and W_t is irrigation water applied (L/m^2). Coefficients w_1, w_2, and w_3 encode the economic value structure of maize production, where w_1 reflects the monetary return per unit yield, and w_2 and w_3 represent the respective costs of nitrogen fertilizer and irrigation water. In this study, w_1 = 0.158, w_2 = 0.79, and w_3 = 1.1, inherited from the prior SOTA LLM-enhanced DQN formulation [23], ensuring consistency with established agronomic and economic modeling assumptions.

This profit-oriented reward formulation links management decisions to long-term economic outcomes by rewarding yield while penalizing excessive input use. Intermediate rewards reflect input costs only, encouraging consideration of delayed effects and cumulative impacts throughout the growing season.

A summary of DSSAT state variables is provided in Table I. By combining an MDP formulation with a high-fidelity crop simulator, the environment provides a principled foundation for learning adaptive fertilization and irrigation strategies that optimize long-term economic performance under uncertainty.

TABLE I. SUMMARY OF DSSAT STATE VARIABLES USED IN THIS STUDY

Category	Variable	Description
Crop Growth and Phenology	xlai	Plant population leaf area index (m^2[leaf]/m^2[soil])
	grnwt	Grain weight dry matter (kg/ha)
	topwt	Above the ground population biomass (kg/ha)
	rtdep	Root depth (cm)
	vstage	Vegetative growth stage (number of leaves)
	istage	Growing stage
	dap	Duration after planting (day)
Management History	cumsumfert	Cumulative nitrogen fertilizer applications (kg/ha)
	totir	Total irrigated water (L/m^2)
Soil Moisture Profile	sw_0 - sw_8	Volumetric soil water content in soil layers (cm^3[water]/cm^3[soil])
Environmental Conditions	srad	Solar radiation during the current day (MJ/m^2/day)
	tmax	Maximum temperature for current day (°C)
	ep	Actual plant transpiration rate (L/ m^2/day)
	dtt	Growing degree days for current day (°C/day)
	wtdep	Depth to water table (cm)
Stress Indicators	nstres	Index of plant nitrogen stress
	swfac	Index of plant water stress

All experiments use the UFGA8201 maize dataset conducted in Florida (1982), a benchmark formally documented in DSSAT and widely adopted in RL-driven agricultural research [24]. The experimental setting follows prior SOTA RL-DSSAT studies to ensure comparability and rigorous evaluation.

B. Actor-Critic Framework

Many prior DSSAT-based RL studies rely on value-based methods such as DQN, which are inherently designed to output a single discrete action at each decision step. While effective for single-variable control problems, this formulation is ill-suited for crop management tasks requiring simultaneous optimization of multiple agronomic inputs, such as fertilization and irrigation. Discrete action mapping schemes are often required, limiting scalability and decision granularity.

To address this limitation, an actor-critic framework is adopted. In this paradigm, the actor directly outputs joint continuous management actions, while the critic evaluates the expected return of the resulting state-action pair. This structure naturally supports coordinated multi-input control and enables fertilizer and irrigation decisions to be optimized jointly.

Four representative actor-critic algorithms are evaluated as baselines: Soft Actor-Critic (SAC) [25], Truncated Quantile Critics (TQC) [26], Twin Delayed Deep Deterministic Policy Gradient (TD3) [27], and Proximal Policy Optimization (PPO) [28]. These methods represent widely adopted stochastic and deterministic policy-gradient approaches for continuous control. Among them, SAC is selected for further development due to its stable entropy-regularized learning objective and robustness in multi-dimensional action spaces.

Together, these algorithms provide a comprehensive foundation for evaluating joint fertilizer-irrigation control within the DSSAT Gym environment.

C. EV-Aware Reward Shaping

While the actor-critic framework enables joint optimization over multiple management actions, policy learning remains driven by instantaneous rewards along individual trajectories. In agricultural decision-making, management actions often produce delayed and uncertain outcomes due to stochastic weather patterns and season-long crop responses. Reliance on step-wise rewards may therefore bias learning toward short-term or episodic outcomes that do not fully reflect long-horizon agronomic objectives.

To better align learning with long-horizon decision-making under uncertainty, this study introduces an expected value (EV)-aware reward shaping mechanism. Rather than optimizing solely for immediate returns, EV-aware shaping incorporates an expectation-based perspective into the reward formulation, guiding policy updates toward decisions evaluated by anticipated long-term economic outcomes.

In RL, the EV represents the long-term average return of a policy under uncertainty. Formally, the EV of a policy π is defined as:

$$EV(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} r_t\right] \tag{2}$$

where, r_t denotes the reward at timestep t, T is the episode length, and $\mathbb{E}[\cdot]$ denotes the expectation over stochastic transitions and environmental variability. In this study, EV corresponds to the average season-level return generated by a policy across varying environmental conditions.

To encode acceptable economic performance targets, the EV-aware reward formulation incorporates a bounded range for expected returns. When the realized EV falls below a predefined lower threshold, a penalty proportional to the shortfall is applied. When EV exceeds the upper threshold, a bonus proportional to the degree of improvement is granted. This range-based shaping encourages consistently viable outcomes while discouraging unstable or underperforming strategies, thereby balancing risk sensitivity and long-term profitability.

By embedding this expectation-based perspective into the training signal, the agent is encouraged to internalize trade-offs among yield generation, input costs, and delayed effects across the growing season. Rather than incentivizing rare extreme yield realizations, EV-aware shaping promotes stable, agronomically feasible returns aligned with practical farming objectives. This design is consistent with recent reward-shaping studies emphasizing structured reward design and the mitigation of misaligned incentives in RL [29], [30].

Importantly, the EV-aware reward shaping mechanism modifies only the internal reward signal used for policy optimization and does not alter the underlying environment dynamics, state representation, or action space. As such, it functions as a guidance strategy within the actor-critic learning process, preserving modeling flexibility while enabling systematic evaluation of expectation-guided rewards in complex crop management settings.

D. Three-Stage Analysis Framework

Prior DSSAT-based RL studies primarily evaluate policies using aggregate performance metrics such as yield or economic profit. However, such evaluations alone are insufficient to assess economic reliability, profit stability, and decision consistency under stochastic environmental conditions. To address this gap, this study adopts a three-stage analysis framework that sequentially evaluates profitability and stability, examines learned action behaviors and decision patterns, and provides post-hoc explainability to identify key drivers of policy decisions.

1) Stage 1: Profitability and profit stability evaluation: Stage 1 establishes a robust baseline comparison among competing RL policies by evaluating economic performance under unseen stochastic weather realizations. Beyond average profitability, the evaluation assesses consistency across diverse environmental conditions to reflect practical deployment requirements. Each policy is evaluated using a common test set and a consistent set of economic and agronomic metrics, including:

- Average economic profit (expected return per episode)
- Maximum economic profit (peak achievable performance)
- Minimum economic profit (downside risk under unfavorable conditions)
- Profit variance metrics (%)
- Average crop yield (kg/ha)
- Average fertilizer usage (kg/ha) and average irrigation usage (L/m²)

To quantify profit stability independently of average profitability, normalized variance indicators are computed (all expressed as percentages):

- Average-maximum variance $V_{\text{max,avg}}$, capturing upside volatility
- Average-minimum variance $V_{\text{min,avg}}$, capturing downside risk
- Maximum-minimum variance $V_{\text{max,min}}$, capturing overall profit dispersion

These metrics are formally defined as:

$$V_{max,\alpha\varpi\gamma} = \frac{|P_{max} - P_{\alpha\varpi\gamma}|}{P_{\alpha\varpi\gamma}} \times 100 \tag{3}$$

$$V_{min,\alpha\varpi\gamma} = \frac{|P_{min} - P_{\alpha\varpi\gamma}|}{P_{\alpha\varpi\gamma}} \times 100 \tag{4}$$

$$V_{max,min} = \frac{|P_{max} - P_{min}|}{P_{\alpha\varpi\gamma}} \times 100 \tag{5}$$

where, P_{avg}, P_{max}, and P_{min} denote the average, maximum, and minimum economic profits observed during testing.

Together, these indicators provide a structured characterization of expected profitability, downside exposure, outcome dispersion, and resource efficiency, enabling economic stability to be evaluated alongside average return.

Within this framework, Stage 1.1 examines the transition from a value-based DQN baseline to a multi-action actor-critic architecture, while Stage 1.2 evaluates the effect of introducing EV-aware reward shaping on the best-performing actor-critic policy in Stage 1.1. All metrics defined in Stage 1 are reused under identical testing conditions to ensure that observed differences in profitability, variance, yield, or resource usage reflect learning behavior rather than experimental inconsistencies.

2) Stage 2: Policy behavior and action pattern analysis: While Stage 1 evaluates profitability and stability, it does not explain how policies generate their decisions. Stage 2 analyzes fertilization and irrigation behavior under stochastic environmental realizations using the same test episodes as Stage 1. Four strategies are compared:

- Expert-informed static management schedule.
- Prior SOTA LLM-enhanced DQN policy.
- Best-performing actor-critic policy (Stage 1.1).
- Corresponding actor-critic policy augmented with EV-aware reward shaping.

a) Stage 2.1: Action magnitude distribution: This sub-stage evaluates fertilizer and irrigation magnitude distributions across test episodes. Aggregated actions are analyzed to compare low- versus high-magnitude applications, characterizing decision granularity and intervention intensity.

b) Stage 2.2: Average temporal action patterns: Average daily fertilizer and irrigation actions are examined as a function of planting day to evaluate scheduling behavior and phase-dependent resource allocation aligned with crop development.

c) Stage 2.3: Spatio-temporal action density across episodes: To capture variability across environmental realizations, spatio-temporal action densities are analyzed across growing days and episodes. This evaluates consistency, dispersion, and adaptability of scheduling patterns under uncertainty.

d) Stage 2.4: Resource efficiency and yield-input trade-off: This sub-stage examines the relationship between average yield and total resource usage. Yield and inputs are normalized to compare relative efficiency across policies, defined as:

$$\text{Εφφιχιενχψ} = \frac{Y}{F+W} \tag{6}$$

where, Y is average crop yield, F fertilizer usage, and W irrigation usage. This formulation highlights the trade-off between resource efficiency and yield-oriented optimization.

3) Stage 3: Input feature importance and policy explainability analysis: Stage 3 evaluates the interpretability of the learned RL policy by identifying key state variables that drive fertilization and irrigation decisions. While Stages 1 and 2 assess performance and characterize action behavior, they do not explain why specific management actions are selected. To address this, post-hoc explainability is applied to reveal how environmental conditions, crop growth states, and management history influence long-horizon decision-making.

Explainability analysis is conducted using SHAP, a model-agnostic framework grounded in cooperative game theory, which attributes each input feature's contribution to the model output [31]. SHAP is applied to the best-performing policy identified in Stage 1 under the same stochastic test conditions used for performance evaluation, ensuring consistency across experimental stages.

Global feature importance is computed separately for fertilizer and irrigation actions using mean absolute SHAP values across evaluation states. Fertilizer- and irrigation-specific importance scores are then aggregated to obtain an overall ranking, providing a unified view of dominant decision drivers across both management dimensions.

Through this analysis, Stage 3 examines whether the learned policy relies on agronomically meaningful signals such as crop growth indicators, cumulative management inputs, and soil moisture conditions rather than spurious correlations. This final stage enhances transparency, interpretability, and real-world deployability by linking observed policy behavior to domain-relevant state variables, complementing the performance and behavioral findings from earlier stages.

IV. Experimental Setup

A. Datasets and Simulation Settings

All experiments are conducted using the DSSAT Gym simulation environment, based on the UFGA8201 maize field experiment conducted in Florida (1982), a benchmark formally documented within DSSAT and widely adopted in RL-based crop management research.

The DSSAT Gym simulates season-long maize growth under stochastic weather conditions, generating daily state transitions and end-of-season agronomic outcomes, including crop yield and economic profit determined by cumulative fertilization and irrigation decisions.

To isolate the effects of learning strategy and reward design, a fixed crop cultivar, soil profile, and planting configuration are maintained across all experiments. Environmental uncertainty is introduced solely through weather variability, enabling evaluation under diverse yet agronomically realistic conditions.

B. Baselines and Compared Algorithms Framework

Four representative actor-critic algorithms, SAC, TQC, TD3, and PPO, are evaluated as Stage 1.1 baselines due to their strong empirical performance in continuous multi-action control tasks and widespread adoption in modern RL research. In addition, a prior SOTA LLM-enhanced DQN approach [23] is included as a primary benchmark, representing transformer-based state encoding integrated with value-based RL for DSSAT-based crop management.

To ensure controlled comparison across learning paradigms, all learning-based agents share a unified architecture and observation preprocessing pipeline. DSSAT state variables are converted into standardized textual descriptions and encoded using a shared DistilBERT-based transformer encoder. The resulting embeddings are passed through a fully connected network with a 512-unit hidden layer followed by a 256-unit hidden layer. For actor-critic methods, this shared representation branches into separate policy and value heads to produce continuous fertilization and irrigation actions. For DQN, the same encoder and hidden layers are retained, with a single Q-value head for action-value estimation. This design preserves representational consistency while isolating algorithmic differences.

Building on the Stage 1.1 baseline comparison, Stage 1.2 introduces EV-aware reward shaping applied to the best-performing actor-critic method. Two EV-aware variants are evaluated: SAC-EV, corresponding to the checkpoint with the highest EV-shaped reward during training, and SAC-Econ, corresponding to the checkpoint with the highest realized economic return. Both variants share identical architectures, action spaces, and training protocols with the Stage 1.1 baseline, differing only in reward formulation, enabling isolation of the reward-shaping effect.

An expert-informed fixed management schedule is included as a non-learning baseline, following standard DSSAT documentation and commonly used benchmarking configurations. This rule-based strategy serves as a lower-bound reference for evaluating the benefits of adaptive RL-based decision-making.

For clarity, learning-based agents are referenced by algorithm name (DQN, SAC, TQC, TD3, PPO), while the fixed schedule is denoted as Expert.

C. Training and Testing Protocol

All learning-based agents are trained and evaluated under a unified protocol to ensure fair comparison across paradigms. Each agent is trained for 3000 episodes, where each episode corresponds to a complete maize growing season simulated within the DSSAT Gym environment. This training horizon follows the prior SOTA LLM-enhanced DQN setting [23], in order to ensure fair and reproducible comparison. Training interactions occur under stochastic weather realizations, enabling policies to learn long-horizon strategies that account for delayed agronomic effects.

Policy performance is evaluated over 100 independent test episodes generated from the same underlying weather process as training, but not encountered during optimization. This ensures an unbiased assessment of robustness under environmental stochasticity.

To ensure consistency across algorithms, identical environment configurations, state representations, and action bounds are used during training and evaluation. Fertilizer and irrigation actions are constrained to continuous ranges of [0, 60] kg/ha and [0, 30] L/m², respectively, for all learning-based agents.

For EV-aware reward shaping, EV is computed as the season-level crop yield return. During training, a bounded EV target range of [10,000, 12,000] kg/ha is specified to stabilize learning and guide optimization. This range is empirically derived from the Stage 1.1 baseline experiments, which consistently reveal a high-yield operating regime. Policies falling below the lower bound incur a proportional penalty, while those exceeding the upper bound receive a proportional bonus.

To ensure statistical reliability, fixed random seeds are used across all experiments. Performance metrics are averaged over test episodes, and variance-related measures are computed to quantify policy stability under weather variability.

D. SHAP Configuration for Input Feature Importance and Policy Explainability

Input feature importance and policy explainability analysis are conducted in Stage 3 using SHAP to interpret the decision-making behavior of the best-performing EV-aware actor-critic policy selected from Stage 1.2, evaluated under the same stochastic test conditions to ensure consistency.

To construct the SHAP background distribution, 800 environment states are sampled from rollout trajectories generated by the trained policy under stochastic test conditions. These states are compressed using K-means clustering into 20 representative centroids, which serve as the background dataset for Kernel SHAP. This reduces computational complexity while preserving representative state variability.

For evaluation, 30 states are independently sampled to compute SHAP values. Global feature importance is calculated separately for fertilizer and irrigation actions by averaging mean absolute SHAP values across evaluation states. Because policy outputs are continuous action magnitudes, only absolute SHAP values are used to quantify feature influence, avoiding directional ambiguity.

An overall feature ranking is obtained by aggregating fertilizer- and irrigation-specific importance scores, providing a unified view of dominant decision-driving variables across management dimensions.

To assess robustness, a stability check is performed by re-running Kernel SHAP with a different random seed. Stability is quantified using Top-10 feature overlap and Spearman's rank correlation between importance rankings. High overlap and strong rank correlation indicate consistent reliance on environmental and crop-related signals, supporting the reliability of the explainability analysis.

V. Results and Analysis

Results are structured into three stages: 1) economic profitability and stability evaluation, 2) behavioral pattern analysis, and 3) policy interpretability under stochastic environmental conditions.

A. Stage 1: Profitability and Profit Stability Evaluation

Stage 1 evaluates economic performance under stochastic weather realizations, examining both architectural upgrades and reward design choices to establish a stable economic baseline for subsequent behavioral analysis.

1) Stage 1.1: Actor-critic upgrade: Stage 1.1 assesses the transition from a value-based DQN to an actor-critic framework in terms of profitability and profit stability under environmental uncertainty.

TABLE II. Training Performance (Stage 1.1)

Model	Highest Training Profit
DQN	1464
SAC	1401
TQC	~500
TD3	N/A
PPO	N/A

Table II reports the highest training profit achieved by each model during learning. These results are presented for completeness and convergence reference only. Training performance reflects optimization under fixed environmental trajectories and is not used for comparative evaluation under stochastic environmental conditions. Under the experimental configuration, DQN and SAC exhibit stable convergence, while the remaining actor-critic baselines fail to converge reliably due to sensitivity in multi-action optimization.

TABLE III. Testing Performance (Stage 1.1)

Metric	DQN	SAC
Avg. Profit	979	974
Max. Profit	1489	1410
Min. Profit	370	623

Table III summarizes testing performance under unseen weather realizations over 100 independent maize-growing episodes simulated under stochastic environmental conditions. SAC achieves a comparable average economic profit relative to DQN, while attaining a substantially higher minimum realized profit. Although average profit levels are similar, the higher minimum profit achieved by SAC indicates improved downside risk protection under unfavorable environmental realizations.

TABLE IV. Profit Variance in Testing Performance (Stage 1.1)

Metric	DQN	SAC	Relative Reduction
$V_{max,\text{avg}}$ (%)	52.09	44.76	7.33
$V_{min,\text{avg}}$ (%)	62.21	36.04	26.17
$V_{max,min}$ (%)	114.3	80.8	33.5

To further quantify economic stability and sensitivity to environmental uncertainty, Table IV reports normalized profit variance and extreme-value deviation metrics.

The maximum-minimum profit deviation captures the spread between best- and worst-case outcomes. SAC achieves a 33.50% reduction in maximum-minimum variability compared to DQN, indicating a narrower profit range across stochastic weather realizations. Consistent reductions across variance metrics confirm improved economic stability under stochastic environmental conditions.

TABLE V. YIELD AND INPUT UTILIZATION IN TESTING (STAGE 1.1)

Metric	DQN	SAC
Avg. Yield (kg/ha)	7864	8567
Avg. Fertilizer (kg/ha)	154	194
Avg. Water (L/m^2)	128	206

Table V reports average crop yield and input utilization during testing. The SAC policy applies higher fertilizer and irrigation inputs on average, resulting in increased crop yield relative to DQN. Since the optimization objective is economic profit rather than input minimization, higher input usage is not inherently suboptimal if it leads to improved and more stable economic returns. These results reflect a trade-off between yield enhancement and input cost.

Overall, Stage 1.1 demonstrates that while average profitability between SAC and DQN is comparable, upgrading to an actor-critic framework yields substantially improved profit stability and reduced downside risk under stochastic environmental uncertainty. SAC achieves a 33.50% reduction in maximum-minimum profit variability, indicating a narrower spread between worst- and best-case outcomes across stochastic weather realizations and establishing a stable foundation for subsequent enhancements.

2) Stage 1.2: EV-aware reward shaping: Stage 1.2 evaluates the effect of incorporating EV-aware reward shaping on top of the best-performing actor-critic policy identified in Stage 1.1 (SAC). Two additional SAC variants are trained: SAC-Econ, selected based on the highest economic profit during training, and SAC-EV, selected based on the highest EV-shaped reward. Both are evaluated against the baseline SAC policy under identical testing conditions to isolate the impact of EV-aware reward shaping.

TABLE VI. TRAINING PERFORMANCE (STAGE 1.2)

Model	Highest Economic Profit	Corresponding Internal Score
SAC (Stage 1.1)	1401	1401
SAC-Econ	1588	2102
SAC-EV	1561	2179

Table VI reports training performance for the three SAC variants, including the highest achieved economic profit, and the corresponding internal training score. SAC-Econ and SAC-EV attain higher peak training profit than the baseline SAC policy, reflecting stronger optimization toward their respective economic objectives.

Table VII summarizes testing performance under unseen weather realizations. EV-aware reward shaping improves average economic profitability over the Stage 1.1 SAC baseline. SAC-EV achieves the highest average profit, corresponding to a 17.56% improvement relative to the Stage 1.1 SAC baseline and a 16.96% improvement over the previous SOTA DQN policy. In addition to improved average profitability, SAC-EV maintains a balanced performance profile, avoiding extreme downside degradation under unfavorable environmental realizations.

TABLE VII. TESTING PERFORMANCE (STAGE 1.2)

Metric	SAC (Stage 1.1)	SAC-Econ	SAC-EV
Avg. Profit	974	1112	1145
Max. Profit	1410	1531	1524
Min. Profit	623	478	574

TABLE VIII. PROFIT VARIANCE IN TESTING PERFORMANCE (STAGE 1.2)

Metric	SAC (Stage 1.1)	SAC-Econ	SAC-EV
$V_{max,\text{avg}}$ (%)	44.76	37.68	33.10
$V_{min,\text{avg}}$ (%)	36.04	57.01	49.87
$V_{max,min}$ (%)	80.80	94.69	82.97

To further examine economic stability, Table VIII reports normalized profit variance and extreme-value deviation metrics. SAC-EV achieves the lowest maximum-average profit variance among all evaluated methods. Although minimum-average and maximum-minimum variance exhibit slight increases relative to the Stage 1.1 SAC policy, SAC-EV consistently maintains substantially lower variability than the previous SOTA DQN across all reported measures, while simultaneously achieving higher economic profit.

TABLE IX. YIELD AND INPUT UTILIZATION IN TESTING (STAGE 1.2)

Metric	SAC (Stage 1.1)	SAC-Econ	SAC-EV
Avg. Yield (kg/ha)	8567	10202	10654
Avg. Fertilizer (kg/ha)	194	241	346
Avg. Water (L/m^2)	206	281	241

Table IX reports average crop yield and input utilization during testing. Both EV-aware variants apply higher fertilizer and irrigation inputs compared to the baseline SAC policy, resulting in increased crop yield. Among the evaluated methods, SAC-EV achieves the highest average yield, accompanied by higher fertilizer and water usage. These differences indicate that EV-aware reward shaping influences both economic outcomes and resource deployment strategies.

Overall, Stage 1 demonstrates that the two design components address complementary aspects of economic performance under uncertainty. The upgrade to an actor-critic framework substantially improves profit stability and reduces exposure to extreme outcomes, while EV-aware reward shaping further enhances average economic profitability. The proposed SAC-EV framework achieves a 16.96% improvement in average economic profit and a 31.33% reduction in maximum-minimum profit variability relative to the previous SOTA DQN baseline, indicating improved profitability and enhanced economic stability under stochastic environmental conditions.

3) Ablation Study 1: Action range sensitivity: To evaluate the impact of action space design on the proposed SAC-EV policy, three fertilizer-irrigation ranges were tested under identical training and evaluation settings. The expert-informed configuration (0-60 kg/ha fertilizer, 0-30 L/m² irrigation) used in Stages 1.1 and 1.2 was compared with a wider range (0-100 kg/ha, 0-50 L/m²) and a narrower range (0-30 kg/ha, 0-15 L/m²).

Variants are denoted as SAC-EV (F, W), where F and W represent the maximum allowable fertilizer (kg/ha) and irrigation (L/m²). Accordingly, SAC-EV (60, 30) corresponds to the baseline configuration, SAC-EV (100, 50) to the wider range, and SAC-EV (30, 15) to the narrower range.

TABLE X. TESTING PERFORMANCE (ABLATION STUDY 1)

Metric	SAC-EV (60, 30)	SAC-EV (100, 50)	SAC-EV (30, 15)
Avg. Profit	1145	1046	894
Max. Profit	1524	1467	1232
Min. Profit	574	631	482
$V_{max,avg}$ (%)	33.10	40.25	37.81
$V_{min,avg}$ (%)	49.87	39.67	46.09
$V_{max,min}$ (%)	82.97	79.92	83.89
Avg. Yield (kg/ha)	10654	10598	10277
Avg. Fertilizer (kg/ha)	346	334	371
Avg. Water (L/m^2)	241	331	436

Table X shows that SAC-EV (60, 30) achieves the highest average and maximum profit while maintaining a competitive minimum profit. Expanding the action bounds does not improve profitability, whereas constraining the range substantially reduces returns. Variance metrics remain broadly comparable across configurations, indicating limited sensitivity of stability to action bounds. Overall, agronomically grounded limits provide the best balance between profitability and resource utilization.

4) *Ablation study 2: EV threshold sensitivity:* To evaluate the sensitivity of EV-aware reward shaping, three configurations were tested under identical training and evaluation settings: the baseline (12k, 10k) used in Stage 1.2, a higher lower-bound variant (12k, 11k), and a higher upper-bound variant (13k, 10k).

Variants are denoted as SAC-EV (U, L), where U and L represent the upper and lower EV thresholds. Accordingly, SAC-EV (12k, 10k) corresponds to the baseline configuration, SAC-EV (12k, 11k) increases the lower threshold, and SAC-EV (13k, 10k) increases the upper threshold.

According to Table XI, the baseline SAC-EV (12k, 10k) achieves the highest average, maximum, and minimum profit while maintaining the lowest variance. Increasing either threshold substantially reduces average and minimum profit and sharply increases profit variability. Although yield and input usage change under modified thresholds, economic performance does not improve, indicating that EV-aware reward shaping is highly sensitive to threshold design and that misaligned EV targets degrade both profitability and stability under stochastic conditions.

TABLE XI. TESTING PERFORMANCE (ABLATION STUDY 2)

Metric	SAC-EV (12k, 10k)	SAC-EV (12k, 11k)	SAC-EV (13k, 10k)
Avg. Profit	1145	691	652
Max. Profit	1524	1241	1253
Min. Profit	574	111	72
$V_{max,avg}$ (%)	33.10	79.59	92.18
$V_{min,avg}$ (%)	49.87	83.94	88.96
$V_{max,min}$ (%)	82.97	156.53	181.13
Avg. Yield (kg/ha)	10654	10961	8588
Avg. Fertilizer (kg/ha)	346	627	204
Avg. Water (L/m^2)	241	497	494

B. Stage 2: Policy Behavior and Action Pattern Analysis

Stage 2 analyzes the behavioral characteristics of the selected policies to explain how economic performance is achieved beyond aggregate profit metrics. In addition to the Expert baseline and the prior SOTA LLM-enhanced DQN, the SAC and SAC-EV policies are included as the best-performing learning-based approaches identified in Stage 1.

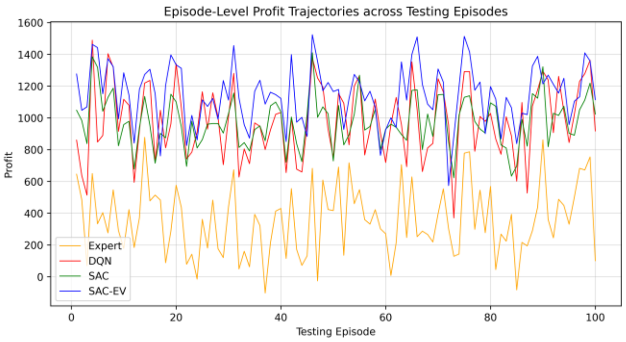

Fig. 2. Episode-level profit trajectories across 100 testing episodes for Expert, DQN, SAC, and SAC-EV.

Fig. 2 illustrates episode-level profit trajectories across 100 test episodes. While average profitability differences were established in Stage 1, these trajectories reveal variations in stability and episode-level behavior across policies. To understand how these outcomes arise, Stage 2 examines action magnitude distributions, temporal scheduling patterns, spatio-temporal input localization, and resource allocation trade-offs under identical stochastic test conditions.

1) *Stage 2.1: Action magnitude distribution:* Stage 2.1 analyzes the distribution of fertilizer and irrigation action magnitudes across testing episodes to characterize policy aggressiveness and resource allocation patterns under stochastic conditions.

Fig. 3 and Fig. 4 illustrate distinct action magnitude distributions across policies. The Expert baseline applies sparse, discrete interventions consistent with a fixed schedule. The DQN policy exhibits limited action diversity with occasional higher-magnitude applications, indicating reactive and coarse adjustments. In contrast, SAC and SAC-EV concentrate on frequent low-to-moderate magnitude actions, reflecting incremental and state-responsive management, with high-magnitude inputs applied selectively.

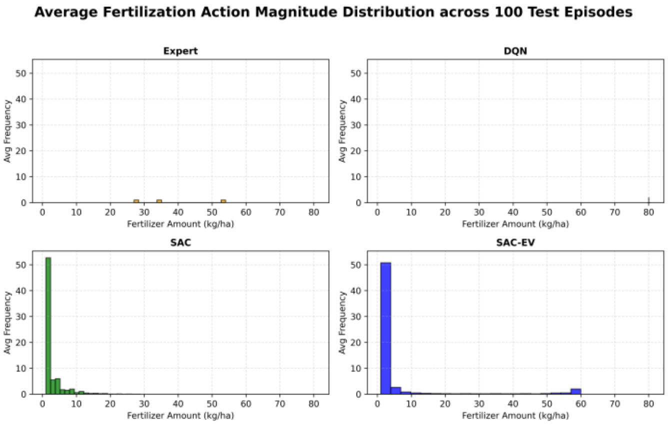

Fig. 3. Average fertilization action magnitude distribution across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

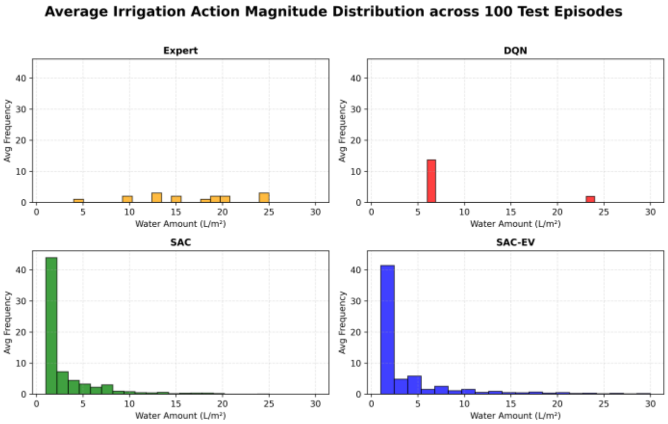

Fig. 4. Average irrigation action magnitude distribution across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

This shift toward finer-grained, high-frequency interventions indicates that actor-critic policies balance responsiveness with stability. SAC-EV shows the strongest concentration of controlled low-magnitude actions, consistent with structured and adaptive long-horizon resource management.

2) Stage 2.2: Average temporal action patterns: Stage 2.2 examines the average temporal structure of fertilizer and irrigation actions across growing seasons to evaluate policy timing and responsiveness under stochastic conditions.

Fig. 5 and Fig. 6 show that inputs are not uniformly distributed over time. Learning-based policies concentrate fertilization and irrigation applications during mid-season growth phases corresponding to elevated crop demand, whereas the Expert and DQN baselines exhibit sparse or irregular timing patterns.

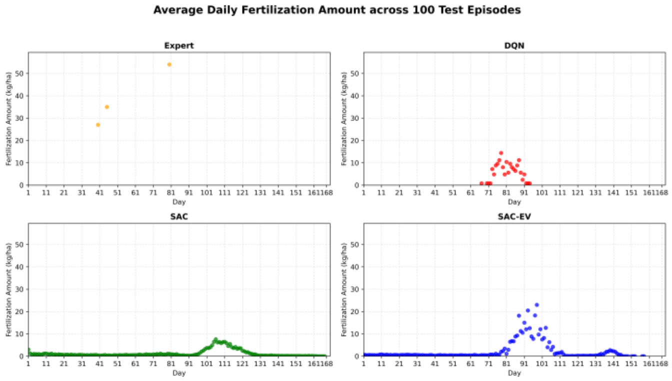

Fig. 5. Average temporal pattern of daily fertilizer application across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

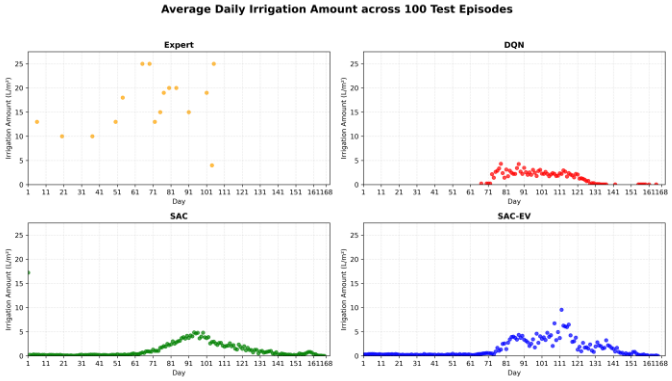

Fig. 6. Average temporal pattern of daily irrigation application across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

Actor-critic methods (SAC and SAC-EV) display smoother temporal profiles characterized by frequent low-to-moderate magnitude adjustments, with larger interventions applied selectively during critical developmental stages. Irrigation follows a similar trajectory, increasing toward peak growth and declining thereafter.

Overall, the learned policies align input timing with crop developmental dynamics, applying high-magnitude actions at agronomically critical phases while maintaining incremental adjustments for routine management, consistent with structured long-horizon control.

3) Stage 2.3: Spatio-temporal action density across episodes: Stage 2.3 analyzes episode-level spatio-temporal action densities to examine how policies allocate fertilizer and irrigation inputs across individual growing days and stochastic realizations.

Compared to the averaged temporal profiles in Stage 2.2, the heatmaps shown in Fig. 7 and Fig. 8 reveal stronger, more localized and episode-specific application patterns. Learning-based policies apply low-magnitude actions across most growing days, with higher-magnitude interventions concentrated within narrow agronomically critical windows.

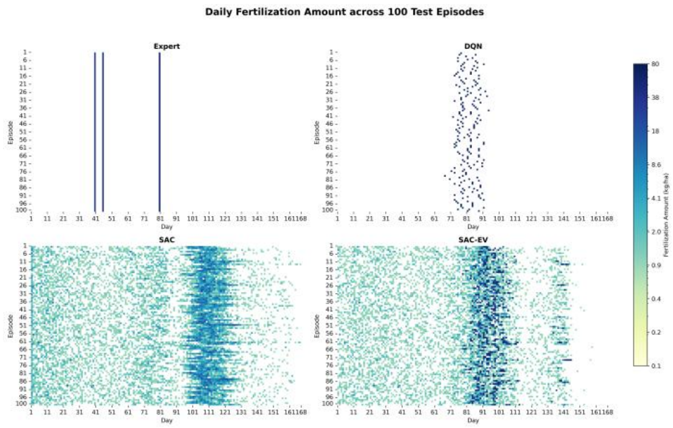

Fig. 7. Spatio-temporal density of daily fertilizer application across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

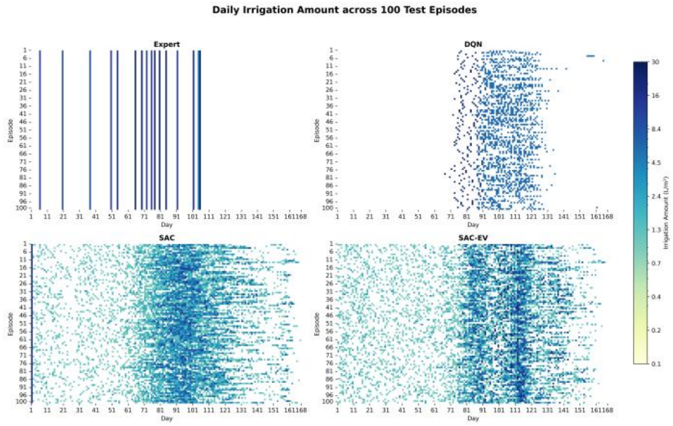

Fig. 8. Spatio-temporal density of daily irrigation application across 100 test episodes for Expert, DQN, SAC, and SAC-EV.

The Expert baseline exhibits rigid vertical bands corresponding to fixed schedule-driven interventions, while DQN displays dispersed high-magnitude actions with weaker temporal alignment. In contrast, SAC and SAC-EV show more continuous distributions of low-magnitude actions, with higher-magnitude inputs strategically concentrated near peak crop demand.

These patterns indicate that larger interventions are temporally localized rather than uniformly distributed, while routine management relies on frequent smaller adjustments, supporting stable and structured long-horizon control under uncertainty.

4) Stage 2.4: Resource efficiency and yield-input trade-off: Stage 2.4 examines the relationship between crop yield, total input usage, and normalized resource efficiency to evaluate how policies balance productivity and input expenditure.

Fig. 9 and Fig. 10 indicate that learning-based policies improve efficiency relative to the Expert baseline through adaptive fertilizer and irrigation allocation. DQN achieves the highest normalized efficiency due to comparatively restrained input usage. However, transitioning from DQN to SAC and SAC-EV increases total input more sharply than normalized efficiency, reflecting a shift in resource deployment strategy.

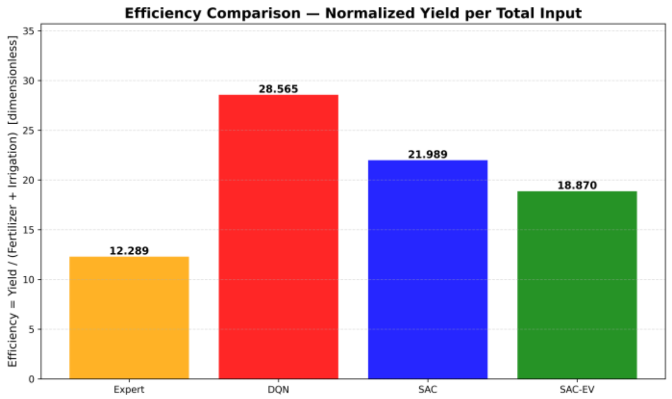

Fig. 9. Resource efficiency comparison measured as normalized yield per unit of total input across Expert, DQN, SAC, and SAC-EV.

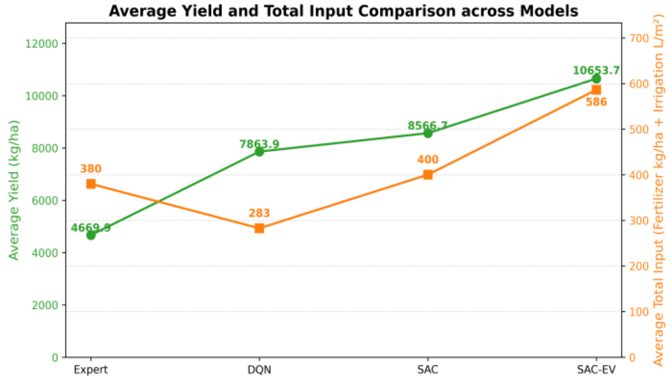

Fig. 10. Comparison of average crop yield and total input usage across Expert, DQN, SAC, and SAC-EV.

This pattern represents a deliberate trade-off aligned with the economic optimization objective. Rather than maximizing yield-per-input in isolation, advanced policies allocate additional resources when the expected economic return exceeds marginal cost. SAC-EV achieves the highest yield and profitability, accompanied by lower normalized efficiency, indicating a strategic shift from conservative input usage toward profit-maximizing allocation. Overall, Stage 2.4 highlights the explicit trade-off between input efficiency and long-horizon economic performance.

C. Stage 3: Input Feature Importance and Policy Explainability Analysis

Stage 3 examines the decision drivers of the best-performing policy (SAC-EV) using SHAP-based feature importance analysis. By identifying which crop, soil, and environmental variables most strongly influence fertilizer and irrigation actions, this stage explains the behavioral patterns observed in Stage 2 and evaluates the stability of the learned decision structure.

1) Global feature importance for fertilizer and irrigation decisions: Fig. 11 and Fig. 12 report mean absolute SHAP values for fertilizer and irrigation actions, respectively. For fertilizer decisions, the most influential features are cumulative fertilizer application (cumsumfert), days after planting (dap), root depth (rtdep), and growth-stage indicators (vstage, istage), indicating that fertilizer control is primarily governed by crop developmental status and accumulated management history.

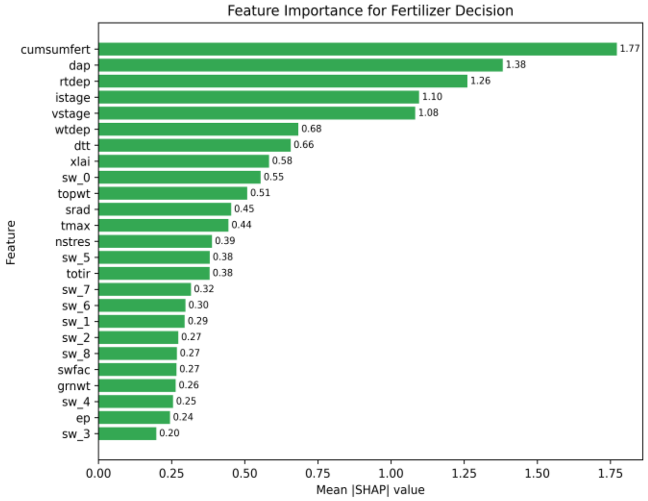

Fig. 11. Global feature importance for fertilizer application decisions based on mean absolute SHAP values.

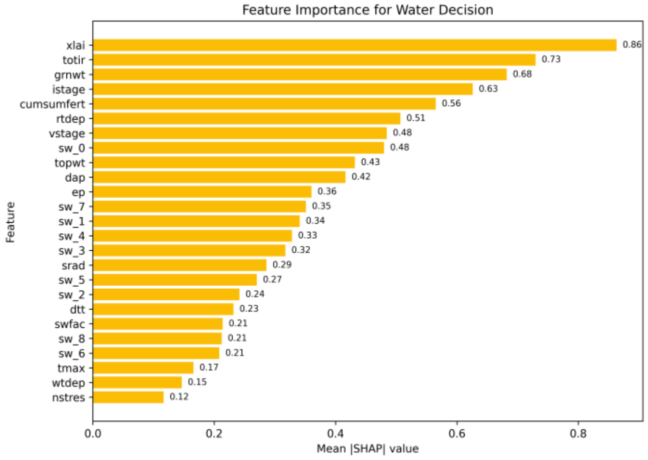

Fig. 12. Global feature importance for irrigation application decisions based on mean absolute SHAP values.

In contrast, irrigation decisions are dominated by leaf area index (xlai), total irrigation applied (totir), biomass growth (grnwt), and growth-stage indicators, reflecting responsiveness to canopy development and immediate water demand. Across both action types, short-term environmental fluctuations and deeper soil moisture layers exhibit comparatively lower influence.

Overall, the results indicate that both fertilizer and irrigation decisions are driven primarily by persistent growth-related state variables and cumulative management history rather than transient environmental signals, supporting a unified and state-dependent decision structure.

2) Overall global feature importance across management decisions: Fig. 13 aggregates SHAP values across fertilizer and irrigation actions. The combined ranking confirms that cumulative fertilizer, days after planting, root depth, and growth-stage indicators consistently dominate feature importance. This consolidated view reinforces that input decisions are conditioned on crop developmental context and prior management state, providing a coherent explanation for the temporally structured and stage-aware behaviors identified in Stage 2.

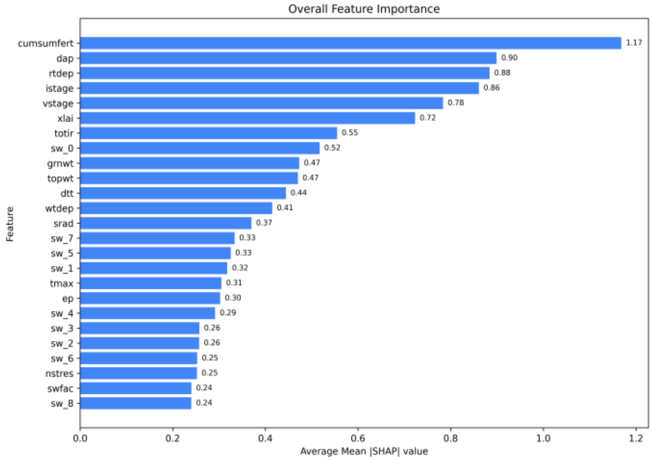

Fig. 13. Overall global feature importance across fertilizer and irrigation decisions based on aggregated mean absolute SHAP values.

3) SHAP stability check: To assess robustness, Kernel SHAP was recomputed using a different random seed, and feature rankings were compared via Top-10 overlap and Spearman's rank correlation (ρ). Fertilizer decisions showed 8/10 overlapping features with ρ = 0.515, while irrigation decisions showed 9/10 overlap with ρ = 0.855. The high overlap and positive correlations indicate that the importance structure is largely preserved under stochastic sampling, confirming that dominant decision drivers are stable rather than artifacts of sampling variability and supporting the reliability of the learned policy's interpretable structure.

Overall, Stage 3 confirms that the behavioral patterns identified in Stage 2 are supported by stable and consistent decision drivers. Growth-stage indicators and cumulative input variables dominate feature importance rankings, explaining the observed stage-aware and temporally structured action patterns. The robustness of feature rankings further supports the reliability and interpretability of the learned policy.

VI. Conclusion

This study addressed the joint optimization of fertilizer and irrigation scheduling for maize under stochastic environmental conditions using a DSSAT-based simulation framework. A Transformer-enhanced SAC model with EV-aware reward shaping was proposed to improve profitability and economic stability in long-horizon crop management.

The SAC-EV framework achieves a 16.96% increase in average profit and a 31.33% reduction in maximum-minimum profit variance compared with the prior SOTA LLM-enhanced DQN baseline. Behavioral analysis shows that the learned policy adopts smaller-magnitude, higher-frequency actions, reflecting temporally structured and stage-aware control rather than fixed or heuristic scheduling. Fertilizer and irrigation inputs are strategically timed according to evolving crop-state trajectories under stochastic environmental variability.

SHAP-based interpretability analysis further reveals that growth-stage indicators, cumulative fertilizer history, and root development variables consistently dominate action selection. This establishes a coherent link between economic performance, policy behavior, and agronomically meaningful state variables, demonstrating that the controller operates on persistent crop-development signals rather than opaque black-box optimization.

Overall, SAC-EV integrates profitability, risk stability, and interpretability within a unified decision-support framework, providing a practically deployable approach for adaptive fertilizer and irrigation management. Future work will extend the framework to multi-crop and multi-region settings and evaluate its effectiveness in real-world agricultural environments through field experiments. Additional extensions may incorporate dynamic economic factors, such as price variability, to further enhance robustness under diverse agronomic and economic conditions.

Availability of Data and Material

Data will be made available on request.

Declaration on Generative AI

The author acknowledges the use of OpenAI's ChatGPT to assist with language polishing and formatting in this manuscript.

References

[1] O. Brenstein, M. Jales, K. Sonder, K. Mottaleb, B.M. Prasanna, Global maize production, consumption and trade: Trends and R&D implications, Food Security 14 (2022) 10–27. https://doi.org/10.1007/s12571-022-01288-7.

[2] United States Department of Agriculture, Foreign Agricultural Service, Corn Explorer, 2025. Available online: https://ipad.fas.usda.gov/cropexplorer/cropview/commodityView.aspx (accessed 29 December 2025).

[3] L. Chen et al., Decision transformer: Reinforcement learning via sequence modeling, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021.

[4] P. Agarwal, A. Rahman, P.-L. St.-Charles, S.J.D. Prince, S.E. Kahou, Transformers in reinforcement learning: A survey, arXiv:2307.05979, 2023. https://arxiv.org/abs/2307.05979.

[5] A. Vaswani et al., Attention is all you need, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[6] H.C. Overweg, H.N.C. Berghuis, I.N. Athanasiadis, CropGym: A reinforcement learning environment for crop management, arXiv:2104.04326, 2021. https://arxiv.org/abs/2104.04326.

[7] Q.-A. Mallard, M. Mathieu, J. Regier, Farm-gym: A modular reinforcement learning platform for stochastic agronomic games, arXiv:2307.07188, 2023. https://arxiv.org/abs/2307.07188.

[8] W. Solow, S. Saissabramanian, A. Fern, WOFOSTGym: A crop simulator for learning annual and perennial crop management strategies, Reinforcement Learning Journal, 2025.

[9] R. Cauton et al., gym-DSSAT: A crop model turned into a reinforcement learning environment, arXiv:2207.03270, 2022. https://arxiv.org/abs/2207.03270.

[10] M.G.J. Kallensberg, D. Ubbens, R. Boone, M. Corbeels, I.N. Athanasiadis, Nitrogen management with reinforcement learning in crop growth models, Environmental Data Science 2 (2023) e34. https://doi.org/10.1017/eds.2023.34.

[11] M. Alkaff, A. Bashuai, Y. Sari, Optimizing water use in maize irrigation with reinforcement learning, Mathematics 13 (2025) 559. https://doi.org/10.3390/math13040595.

[12] X. Ding, W. Du, Optimizing irrigation efficiency using deep reinforcement learning in the field, ACM Transactions on Sensor Networks 20 (2024). https://doi.org/10.1145/3659931.

[13] J. Liu et al., Deep reinforcement learning for irrigation optimization: Advantages, Opportunities, and Challenges, World Bank, Washington, DC, USA, 2023.

[14] J.P. Padilla-Nates, L.D. Garcia, C. Lozoya, L. Orona, A. Cortez-Perez, Greenhouse irrigation control based on reinforcement learning, Agronomy 15 (2025) 2781. https://doi.org/10.3390/agronomy15122781.

[15] R. Gautron, Q.-A. Mallard, P. Preux, M. Corbeels, R. Sabbadin, Reinforcement learning for crop management support: Review, prospects and challenges, Computers and Electronics in Agriculture 200 (2022) 107182. https://doi.org/10.1016/j.compag.2022.107182.

[16] L. Wang, S. Xiao, J. Wang, A. Parab, S. Patel, Reinforcement learning-based agricultural fertilization and irrigation considering NO emissions and uncertain climate variability, Agriculture 7 (2025) 252. https://doi.org/10.3390/agriculture7080252.

[17] G. Goldstein, K. Mallinger, S. Rubitzek, T. Neuber, Current applications and potential future directions of reinforcement learning-based Digital Twins in agriculture, Smart Agricultural Technology 8 (2024) 100512. https://doi.org/10.1016/j.atech.2024.100512.

[18] B. Morogoe, W. Yin, S. Boersma, E. van Henten, V. Puig, C. Sun, Reinforcement learning versus model predictive control of greenhouse climate control, Computers and Electronics in Agriculture 215 (2023) 108372. https://doi.org/10.1016/j.compag.2023.108372.

[19] R. Tao et al., Optimizing crop management with reinforcement learning and imitation learning, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 2023, pp. 6228–6236.

[20] D. Wright et al., Field corn production guide, University of Florida IFAS Extension, Gainesville, FL, USA, 2022, Rep. SS-AGR-85.

[21] W. Malik, R. Isla, F. Deschmi, DSSAT-CERES-maize modelling to improve irrigation and nitrogen management practices under Mediterranean conditions, Agricultural Water Management 213 (2019) 298–308. https://doi.org/10.1016/j.agwat.2018.10.002.

[22] A. Skhiri, F. Dechmi, Impact of sprinkler irrigation management on the Del Reguero River (Spain), Agricultural Water Management 110 (2012) 120–129. https://doi.org/10.1016/j.agwat.2011.11.003.

[23] Z. Wu et al., The new agronomists: Language models as experts in crop management, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 5346–5356.

[24] L.A. Huston, K.J. Boote, Data for model operation, calibration, and evaluation, in: Systems Approaches for Sustainable Agricultural Development, Springer, Dordrecht, 1998, pp. 39–49.

[25] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, arXiv:1801.01290, 2018. https://arxiv.org/abs/1801.01290.

[26] A. Kuznetsov, S. Shevchuk, A. Grishin, D. Vetrov, Controlling observational bias with a truncated mixture of continuous distributions, arXiv:2005.04269, 2020. https://arxiv.org/abs/2005.04269.

[27] S. Fujimoto, H. van Hoof, D. Meger, Addressing function approximation error in actor–critic methods, arXiv:1802.09477, 2018. https://arxiv.org/abs/1802.09477.

[28] J. Schulman et al., Proximal policy optimization algorithms, arXiv:1707.06347, 2017. https://arxiv.org/abs/1707.06347.

[29] S. Ibrahim, M. Mostafa, A. Jandik, H. Salioum, P. Osinenko, Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications, arXiv:2408.10215, 2024. https://arxiv.org/abs/2408.10215.

[30] C. Musilmani, M.L. Littman, Y. Bengio, A reward alignment metric for RL practitioners, Reinforcement Learning Journal, 2025.

[31] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.