

LLM-Assisted Offline Reward Design for Reinforcement Learning-Based Irrigation Scheduling

Qiufen Wang*, Tianhao Wang, Ziyue Wang

School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, China

Abstract—This study develops an LLM-assisted offline workflow for reward design in reinforcement learning-based irrigation scheduling under a seasonal water allocation. Researchers provide agronomic knowledge and feedback on candidate-policy behavior, and the language model translates these inputs into executable rewards. Eight candidates are compared over four generations. The selected reward, aridity–soil curriculum, and FAO-56 exploration mask are held fixed during final policy training and transfer evaluation. Across 24 transfer scenarios comprising two withheld sites, three soil textures and four seasons, proximal policy optimization achieves a mean net benefit of 1943 USD ha⁻¹, 2.8 per cent above the strictest tuned rule and 9.1 per cent above the untuned FAO-56 rule. The paired mean gain is 52.5 USD ha⁻¹, with site means of +127.9 at Lubbock and –23.0 at Urbana. Grouping shared-weather outcomes into eight site–season blocks gives a two-sided sign-flip p-value of 0.117; joint resampling of sites, seasons, and training seeds gives an exploratory percentile range of –43.3 to 159.6 USD ha⁻¹. The combined curriculum and exploration mask increase convergence frequency from 5/10 to 10/10 runs in the evaluated ablation. The results connect offline reward design with stage-sensitive water allocation and show how its scheduling value varies between semi-arid and humid conditions.

Keywords—Smart agriculture; irrigation scheduling; deep reinforcement learning; large language models; reward design; curriculum learning; water use efficiency

I. INTRODUCTION

Irrigated agriculture accounts for roughly seventy per cent of global freshwater withdrawals, and depletion threatens major aquifer-dependent regions. Scanlon et al. documented groundwater depletion in the High Plains and Central Valley [1]. Deines et al. quantified irrigation adaptation following stakeholder-driven groundwater management [2]. Drysdale and Hendricks examined adaptation to locally imposed irrigation restrictions [3], while Deines et al. combined remote sensing and crop models to assess the sustainability of groundwater management [4]. These restrictions make the timing of a limited seasonal allocation an operational question. FAO-56 supplies the water-balance basis for estimating crop requirements [5]; scheduling must additionally distribute an insufficient volume across a season whose daily weather is uncertain.

Threshold rules are a strong operational reference because they connect root zone depletion to an interpretable irrigation trigger. FAO-33 specifies stage-dependent yield responses to water deficit [6], and Fereres and Soriano explain how deficit irrigation can preserve productivity by concentrating shortfalls in less sensitive stages [7]. Stage-specific thresholds incorporate part of this knowledge. A seasonal allocation adds an

intertemporal decision: an application today reduces the volume available at flowering, while a minimum interval between applications makes the next irrigation opportunity relevant to the current action. These management conditions motivate the sequential scheduling problem examined here.

Reinforcement learning offers a way to represent this sequential decision. Sun et al. formulated irrigation as a water-efficient reinforcement learning control problem [8]. Yang et al. extended the decision state to soil moisture, evapotranspiration, forecast precipitation, and crop stage with deep reinforcement learning [9]. Saikai et al. provided a framework for learning irrigation decisions from high-dimensional crop and soil feedback [10]. Three practical issues motivate the present study: a harvest objective supplies sparse feedback, each episode requires a complete growing-season simulation, and policies must cope with changes in soil and climate. Dense rewards provide daily learning signals, structured training experience supports convergence, and held-out scenarios assess transfer across soil and climate conditions.

Language models offer a means of translating textual domain knowledge into executable reward programs. Ma et al. developed evolutionary reward-code optimization in Eureka [18]. Xie et al. introduced dense reward generation and iterative refinement with human feedback in Text2Reward [19]. Yu et al. translated language instructions into reward specifications for robotic skill synthesis [20]. Kwon et al. investigated language models as proxy reward functions [21], and Wang et al. developed an embodied agent that expands a code library through feedback in Voyager [22]. This study applies the offline design idea to irrigation. Researchers provide agronomic inputs, interpret candidate-policy behavior, and guide revisions; the resulting reward, curriculum, and exploration mask are then held fixed during final policy training and transfer evaluation. The irrigation controller is a conventional policy network.

The contributions are as follows.

- We document a human-guided offline workflow that translates agronomic knowledge and qualitative policy feedback into eight reward programs, and compare their learning outcomes under a common simulator and training protocol.
- We evaluate the selected reward with an aridity–soil curriculum and a FAO-56 exploration mask, quantifying their effects on convergence frequency and the performance of converged runs.
- We examine transfer across withheld sites, soil textures and seasons, report the contrast between semi-arid and

*Corresponding author.

humid conditions, and account for shared weather and training-seed variation when interpreting the pooled policy margin.

Section II reviews irrigation control and language-assisted reward design. Section III describes the simulator, offline design workflow, and learning algorithms. Section IV sets out the experimental protocol. Section V presents the candidate comparison, transfer results, ablation, and limitations. Section VI concludes the study.

II. RELATED WORK

Irrigation reinforcement learning studies differ in their simulators, objectives, and evaluation protocols. Yang et al. developed deep irrigation scheduling from multiple agronomic observations and compared simulated net returns with conventional approaches [9]. Saikai et al. used APSIM-Wheat with phenology, canopy, and layered soil-water feedback to learn daily irrigation decisions [10]. Kelly et al. released AquaCrop-GYM and showed that the value of learned irrigation depends on rainfall variability and water restrictions [11]. Alkaff et al. combined AquaCrop-OSPy and PPO with daily water penalties and a harvest-yield reward for maize [12]. Kelly et al. separately assessed the economic value of adapting irrigation within the season [27]. Alibabaei et al. studied irrigation optimization at a site in Portugal [28], and Chen et al. incorporated weather forecasts into reinforcement learning irrigation decisions for rice [29]. Wu et al. treated nitrogen management as a crop-simulation control problem [30]. These studies motivate explicit agronomic objectives and strong tuned-rule comparisons.

Crop models determine the agronomic responses available to an irrigation agent. Jones et al. described the integrated DSSAT cropping-system model [23]. Keating et al. introduced APSIM as a modular farming-systems simulator [24], and Holzworth et al. described its subsequent development [25]. Steduto et al. presented the canopy and water-productivity principles of AquaCrop [26]. The present study uses a compact FAO-based water balance coupled to a single-cell nitrogen mechanism to represent stage-dependent water stress and nutrient loss through drainage.

Open simulation interfaces and field studies address different parts of the development problem. Overweg et al. introduced CropGym for reinforcement learning fertilization management with process-based crop models [15]. Gautron et al. exposed DSSAT through a Gym interface with maize-based fertilization and irrigation tasks [16]. Turchetta et al. introduced CyclesGym for long-term crop management [17]. Ding and Du proposed DRLIC for irrigation control [13] and subsequently studied deep reinforcement learning irrigation in the field [14]. These contributions respectively support reproducible simulation comparisons and the study of real actuation constraints. The present work evaluates frozen reward programs and transfer scenarios in simulation.

Existing language-assisted methods provide several design mechanisms. Eureka generates reward code and revises candidates using environmental fitness feedback [18]. Text2Reward produces dense executable rewards from a task description and permits human feedback during refinement [19].

Yu et al. specify robotic skills through language-generated rewards that are optimized downstream [20]. Kwon et al. investigate a language model as a proxy reward signal [21]. Huang et al. extract actionable plans from language models [31], while Voyager builds an expanding library of executable skills [22]. Madaan et al. study iterative self-refinement using self-feedback [32], and Shinn et al. use verbal reflection on prior outcomes in Reflexion [33]. Tao et al. combine reinforcement and imitation learning for crop management [34]. The workflow in Section III-B uses researcher-written behavioral feedback and numerical candidate scores. Its retained artifacts support reproduction of the downstream policy experiments.

Potential differences and curricula have an established theoretical basis. Ng et al. characterize reward transformations that preserve optimal policies under their stated assumptions [35]. Devlin and Kudenko extend potential-based shaping considerations to multi-agent systems [36]. Mannion et al. study policy invariance in multi-objective reinforcement learning [37], and Eck et al. examine finite-horizon potential-based shaping for online POMDP planning [38]. Bengio et al. introduce curriculum learning through ordered training examples [39]. Narvekar et al. organize the reinforcement learning curriculum design space [40], and Florensa et al. generate curricula by progressively expanding feasible starting states [41]. Here, a fixed score orders archived training scenarios by aridity and soil water capacity. Its deterministic implementation is the training intervention evaluated by the ablation.

The simulator draws its agronomic components from established sources. FAO-56 supplies the crop water balance [5]. Allen et al. detail its dual crop coefficient procedure [42], and Pereira et al. review its use in crop evapotranspiration estimation [43]. FAO-33 provides stage-dependent yield response factors [6]. Saxton and Rawls relate texture and organic matter to soil-water characteristics [44], and Raes et al. describe the AquaCrop algorithms used as the source of the drainage characteristic [45]. Burns provides a mixing-cell relation for solute redistribution [46], while Ciampitti and Vyn review maize nitrogen uptake and yield relationships [47]. Fereres and Soriano review deficit irrigation practice [7]. Geerts and Raes discuss deficit irrigation as a water-productivity strategy [48], and Payero et al. quantify the effect of timing a limited maize irrigation allocation [49]. Foster et al. release AquaCrop-OS for repeated crop-water simulations [50], and Perez-Quesada and Hendricks analyze collective efforts to manage irrigation withdrawals in Kansas [51]. These sources motivate the scenario mechanisms used here.

III. MATERIALS AND METHODS

The workflow links human-guided offline design to irrigation control. Researchers specify agronomic inputs and management conditions, inspect candidate-policy behavior, and supply numerical and qualitative feedback to the language model. Candidate PPO training provides the screening scores. The selected reward program, aridity-soil ordering, and FAO-56 exploration mask are frozen before final policy training. A conventional policy network then reads soil water, weather, crop stage, and management state and selects one of five irrigation depths. Fig. 1 presents the design, training and transfer stages.

The accompanying code contains the reward functions and training rules used throughout the policy experiments.

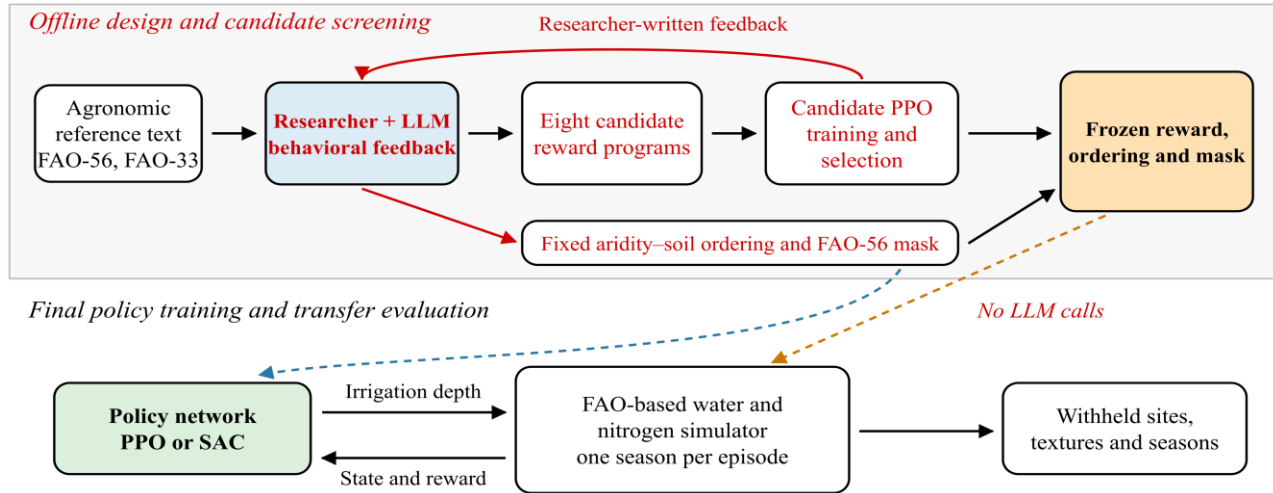


Fig. 1. Human-guided offline design followed by policy training and transfer evaluation with frozen artifacts.

A. Simulation Environment

The environment is a daily water balance of a single homogeneous soil profile under a maize crop, following the dual crop coefficient procedure of FAO-56 [5]. Reference evapotranspiration is computed by the FAO-56 Penman-Monteith equation from daily maximum and minimum temperature, incoming shortwave radiation, mean relative humidity, and wind speed at two metres. Crop evapotranspiration is partitioned into transpiration, through the basal crop coefficient K_{cb} , and evaporation from the surface layer, through the evaporation coefficient K_e , so that water lost from a wet surface after an irrigation event is accounted separately from that transpired by the canopy: $ET_c = (K_{cb}K_s + K_e)ET_0$, where K_s is the water stress coefficient, which falls below unity once root zone depletion D_r exceeds the readily available water $RAW = pTAW$. Total available water TAW is the difference between water content at field capacity and at the

wilting point over the current rooting depth, and the depletion fraction p for maize is 0.55. Soil hydraulic properties are assigned by texture class from FAO-56 Table XIX and the pedotransfer relations of Saxton and Rawls [44]; the six textures used span total available water from 70 to 160 mm per metre of profile. Two departures from the FAO-56 balance are needed for the decision problem to be well posed. First, water applied in excess of field capacity does not vanish within the day; a transient store above field capacity drains at the rate $\tau = 0.0866 K_{sat}^{0.35}$ of the AquaCrop balance [45], bounded by the saturated conductivity, so that a large application on a heavy soil produces percolation over several days rather than instantaneously. Second, surface runoff is computed by the SCS curve number method, with curve numbers assigned by texture. Fig. 2 names the state variables of the profile; the same notation is used in the reward listings of Section III-B.

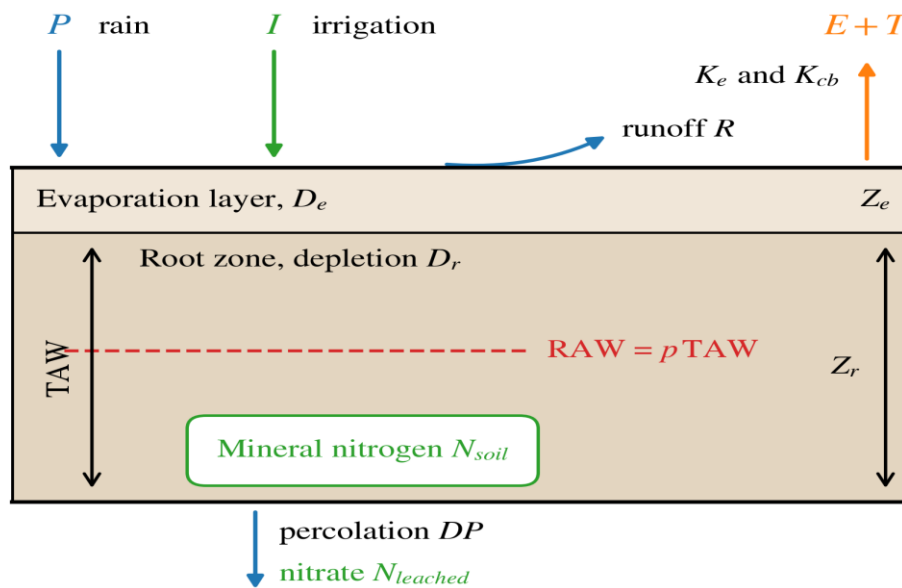


Fig. 2. Soil-water and nitrogen state variables in the FAO-based irrigation simulator.

Yield responds to water through the stage-wise multiplicative form of FAO-33 [6]. Each growth stage i contributes $1 - K_{y,i}(1 - \bar{K}_{s,i})$, where $\bar{K}_{s,i}$ is the mean stage stress coefficient and $K_{y,i}$ is the stage yield response factor. Relative water-limited yield is the product over four stages. The maize factors are 0.40, 1.50, 1.30 and 0.50, making reproductive-stage deficits more costly than early deficits. In the initial water-only configuration, coordinate ascent on daily irrigation depths matched the threshold rule. This development result motivated the addition of nitrogen dynamics to represent the trade-off between irrigation, drainage, and nutrient retention.

The simulator was subsequently extended with a single mixing cell for root zone mineral nitrogen, following Burns [46]. Fertilizer and net mineralization supply nitrogen, crop uptake removes it, and percolating water displaces nitrogen according to $N_{leach} = N_{soil}[1 - \exp(-P_{erc}/W)]$, where W is the water volume resident in the profile plus the depth draining that day. Fertilizer applied at the surface is excluded from that day's drainage. Nitrogen-limited yield is computed from seasonal uptake relative to crop requirement with a response factor of 0.90 and combined non-compensably with water-limited yield. Fertilizer is fixed at 240 kg ha⁻¹, with 40 per cent before sowing and 60 per cent on day 35; crop requirement is 22 kg N per tonne of grain following Ciampitti and Vyn [47]. This mechanism assigns a yield cost to excessive drainage. The FAO-56 and FAO-33 routines were checked against twelve source worked examples distributed with the code.

One episode represents a 140-day maize season beginning on 1 May, with one daily irrigation decision. Gross application depths are 0, 10, 20, 30, and 40 mm, with application efficiency 0.85. The scenario imposes a four-day minimum interval between applications, an 8 USD cost per start, sensor noise with standard deviation 0.02 m³ m⁻³, and a three-day rainfall forecast with skill 0.6. The seasonal quota equals 0.70 of the unconstrained threshold rule's application for the same archived site–soil–season. This retrospective reference-based quota is computed before policy evaluation and applied identically to all policies. The evaluation objective is $NB = p_g Y - p_w I - c_e n_e$, where $p_g = 220$ USD t⁻¹, $p_w = 0.05$ USD m⁻³, $c_e = 8$ USD, and n_e counts irrigation starts. Irrigation depth is converted to volume at 10 m³ ha⁻¹ per millimetre. Fertilizer expenditure is constant across policies and omitted from their net benefit; leaching influences the objective through simulated yield.

The observation is a 19-dimensional vector containing sensor-based relative depletion and its margin to readily available water; basal crop coefficient, season progress and growth stage; reference evapotranspiration and current rainfall; the three-day rainfall forecast; rooting depth and surface dryness; remaining quota, days since the last application and machine availability; nitrogen sufficiency and profile mineral nitrogen; and total available water and seasonal aridity index. Total available water represents soil storage, while the aridity index summarizes climatic demand using the archived growing-season weather.

B. Human-Guided Offline Design

The offline workflow supplies the language model with agronomic summaries from FAO-56 and FAO-33, the reward

callback signature, accessible state variables and their units, and the management conditions in Section III-A. Researchers inspect candidate-policy outcomes and provide feedback on inadequate irrigation, excessive percolation, and early quota use. This feedback guides three of the four generations. The model translates the agronomic inputs and feedback into reward code, and researchers select the retained program using design-selection scores. The study follows one four-generation design trajectory. Its reward functions and candidate-level rationales are retained in the accompanying code for subsequent policy experiments.

The design specification uses the callback `fn(flux, action_mm, summary, env)`, with summary supplied at harvest. Inputs include daily transpiration and crop demand, growth stage, depletion and available water, leached nitrogen, remaining quota, and rainfall forecast. Agronomic summaries cover the RAW stress threshold, maize stage yield responses, nitrate displacement by drainage, nitrogen uptake, forecast-aware scheduling, surface evaporation, and the interval between centre-pivot applications. The output is an executable scalar reward. Candidate annotations record the behavioral feedback associated with each revision, linking the design rationale to the implemented reward functions.

Candidate training uses Garden City and Lincoln on sand, loam, and clay loam in 2006 and 2011, giving 12 environments. Candidate selection uses the same sites and textures in 2013, giving six environments. Each candidate receives 1600 training seasons without the curriculum or exploration mask. All eight candidates are screened with three seeds, and the two leaders receive two additional seeds each, giving 28 candidate runs. The hand-written reference receives three separate runs, so Table I contains 31 runs in total. The six selection environments also occur in the later policy-training pool; the final transfer set withholds sites, textures, and seasons from both stages.

Table I summarizes the eight candidates and the hand-written reference. Generation 1 supplies either terminal net benefit alone or a daily transpiration credit. Generation 2 introduces stage-dependent yield response weights, first as a stress penalty and then as attained transpiration, followed by a daily penalty on leached nitrogen. These candidates provide alternative dense learning signals for the same training procedure. The leaching term uses the simulator's daily nitrogen loss to guide the irrigation–drainage trade-off.

The third generation added potential-based shaping. The potential is the squared exceedance of depletion beyond the readily available water, weighted by the stage K_y ,

$$\Phi = -K_{y,i} \left(\frac{\max(0, D_r/TAW - p)}{1-p} \right)^2 \quad (1)$$

The shaping signal $\Phi_{t+1} - \Phi_t$ provides daily feedback on changes in stage-weighted depletion. Researcher feedback on early quota use led to a pacing term that penalizes expenditure ahead of the K_y -weighted seasonal demand profile,

$$r_{pace} = -1/2 \left[\max \left(0, \frac{I_t}{Q} - \frac{\sum_{d \leq t} K_{y,d} K_{cb,d}}{\sum_d K_{y,d} K_{cb,d}} \right) \right]^2 \quad (2)$$

where, I is cumulative gross irrigation, and Q is the fixed seasonal quota. The pacing coefficient is 1/2 in Eq. (2). Pacing

and the daily agronomic terms define a proxy training objective; candidate selection and policy comparisons use seasonal net benefit in USD ha⁻¹. The selected reward is

$$r_t = K_{y,i} \frac{T_a}{T_c} + 2(\Phi_{t+1} - \Phi_t) - 0.05 N_{leach,t} + r_{pace} + 1[t = T] \frac{NB}{200} \quad (3)$$

The coefficient on stage-weighted attainment sets the reference daily scale. The factor 2 scales the potential difference, 0.05 scales daily leached nitrogen, and the 1/2 coefficient in Eq. (2) scales excessive early quota use. Dividing terminal net benefit by 200 brings the harvest signal into the numerical range used for training. These fixed heuristic scales combine agronomic feedback with the economic terminal objective. The leaching coefficient assigns 0.05 reward units per kg N ha⁻¹. Candidate selection and policy evaluation use unshaped seasonal net benefit, keeping their economic comparison on a common scale.

The fourth generation adds two anticipation terms: a charge for applications made redundant by forecast rainfall and a charge for depletion projected to the next irrigation opportunity. Its design-selection score is below the paced third-generation candidate. Table I includes this outcome alongside all earlier candidates, documenting the complete search trajectory.

The second retained artifact is a deterministic difficulty score. The proposed ordering uses total available water and the seasonal aridity index $AI = P/ET_0$, ranking humid, high-storage scenarios before arid, low-storage ones. The implemented score is

$$\text{difficulty} = w_a \left[1 - \min \left(1, \frac{AI}{0.9} \right) \right] + w_s \left[1 - \frac{TAW - TAW_{\min}}{TAW_{\max} - TAW_{\min}} \right] \quad (4)$$

with $w_a = 0.65$ and $w_s = 0.35$. These fixed heuristic weights emphasize climatic scarcity and retain the buffering effect of soil water capacity. Training uses four cumulative stages, each lasting one quarter of the 4000-season budget and covering the easiest $(k+1)/4$ fraction at stage k . Earlier environments remain available as the pool expands. Direct computation of Eq. (4) from the same descriptors and tie-breaking rule reproduces the ordering used in training.

The third artifact is a FAO-56-based exploration rule. It triggers when depletion exceeds readily available water, refills toward relative depletion 0.30, and discounts forecast rainfall when estimating the required depth. A forecast discount of 0.70 is used, and the nearest available gross irrigation depth is selected after correction for application efficiency. During the first 500 training seasons, the mask admits actions within one index of the suggested action and retains zero irrigation. The mask is then removed. These settings implement an agronomic prior during early exploration.

C. Learning Algorithms

Two algorithms are trained on the frozen artifacts. Proximal policy optimisation [52] uses a clipped surrogate objective with generalised advantage estimation [53], hidden layers of 64 units, learning rate 3×10^{-4} , GAE parameter 0.95, clipping 0.2, ten epochs per update at minibatch 256, four episodes per rollout, entropy coefficient 0.01, and early stopping at a target KL

divergence of 0.03. Haarnoja et al. introduced the maximum-entropy framework for soft actor-critic [54], and Christodoulou adapted it to discrete actions [55]. The SAC implementation uses hidden layers of 128 units, target smoothing 0.005, a replay buffer of 200,000 transitions, 2000 initial random steps, and an automatically tuned temperature with target entropy $0.20 \log|A|$. Both algorithms use a discount factor of 1.0 to retain the full seasonal terminal objective; a factor of 0.995 would approximately halve its contribution over 140 steps.

Validation experiments informed the entropy and learning-rate settings. The SAC target-entropy sweep returned net benefits of 1255, 1486 and 1246 USD ha⁻¹ at coefficients of 0.10, 0.20 and 0.60, respectively. At 0.60 $\log \text{verticallog}|A|$, the learned temperature reached 2370, and the entropy bonus outweighed the daily agronomic reward. The 0.20 coefficient was therefore retained. Linear annealing of the PPO learning rate and entropy coefficient reduced validation net benefit from about 1950 to 1340 USD ha⁻¹. Its final-quarter step size was one tenth of the initial value, coinciding with the introduction of the hardest curriculum environments. Constant learning-rate and entropy settings are used in the reported experiments.

IV. EXPERIMENTAL PROTOCOL

Daily weather for six sites in the United States maize belt was obtained directly from the NASA POWER daily point service, release v9.0.1, agroclimatology community, for 2005–2023 [56]. Sparks describes access to this data source through the nasapower R client [57]; the present code downloads the daily records directly. The variables are maximum and minimum temperature at two metres, corrected precipitation, all-sky downward shortwave radiation, relative humidity at two metres, and wind speed at two metres. Each site record contains 6939 days with no missing values. Zhang et al. describe solar-irradiance processing in the NASA Surface Meteorology and Solar Energy data lineage [58]. Mean growing-season aridity index is 0.024 at Fresno, 0.301 at Lubbock, 0.389 at Garden City, 0.676 at Lincoln, 0.814 at Urbana and 0.822 at Ames.

An environment is a site–soil–season combination. Final policy training uses Fresno, Garden City, Lincoln and Ames on sand, loam and clay loam in 2007, 2010, 2013 and 2016, giving 48 environments. Validation uses the same sites and textures in 2018, giving 12 environments. Transfer uses Lubbock and Urbana on sandy loam, silt loam, and clay in 2020–2023, giving 24 scenarios. Transfer sites, textures, and seasons are withheld from candidate design and final policy training. Candidate selection uses six environments from the final training pool. This protocol assesses transfer across location, soil, and season under a common simulator, crop parameterization, and quota mechanism.

Five non-learning policies are evaluated on the transfer set. Rainfed applies zero irrigation. The fixed threshold rule irrigates above relative depletion 0.50. The FAO-56 rule implements the readily available water trigger, a target depletion, and a forecast discount with fixed settings. The tuned stage threshold fits vegetative and flowering triggers and an application depth. The tuned FAO-56 rule fits a trigger scale, target depletion, and forecast discount. Both fitted rules use grid search on seven training environments and transfer their selected settings unchanged: (0.70, 0.70, 40 mm) and (1.2, 0.4, 1.0), respectively.

All policies use paired sensor and forecast realizations within each transfer scenario.

PPO is trained for 4000 seasons and SAC for 2000, with SAC performing one gradient update per environment step. The behavior prior masks actions during the first 500 seasons, and validation is performed every 100 seasons. The main comparison uses ten PPO seeds and five SAC seeds; each of the four ablation conditions uses ten seeds. Policies are selected at their best validation snapshot and then evaluated on the transfer set. In the initial five-seed PPO comparison, validation-best snapshots returned 1952 ± 20 USD ha⁻¹, compared with 1863 ± 113 for final weights. Two runs peaked near season 2500 and later shifted toward fewer, heavier applications: at 364 mm of irrigation, yield fell by 0.96 t ha⁻¹ and nitrate loss rose 1.6-fold. Validation-based selection accommodates this late-training variability while keeping transfer outcomes separate from checkpoint selection.

Seasonal net benefit is the primary metric, accompanied by grain yield, applied water, total-input water productivity (WP), irrigation water productivity, application count, leached nitrogen, and stress days. Policy values are averages over the 24 transfer scenarios. The $\pm$ terms for learned policies and error bars in Fig. 4 show population standard deviations across training seeds. All irrigation policies apply nearly their full quota, with transfer means of 360–364 mm; scheduling and yield therefore account for most of the net-benefit differences. Sensor and forecast draws are paired between policies within each scenario. The final column of Table II reports relative net benefit against the tuned FAO-56 rule.

For inference, the three textures at a given site and season are grouped because they share weather. Paired PPO–rule differences are averaged over textures and the ten PPO seeds, giving eight site–season means. The two-sided sign-flip test enumerates their 256 sign assignments using the absolute mean as its statistic and assumes symmetry and independence of the blocks. A site-level sensitivity analysis averages seasons within each site and enumerates the four sign assignments of the two site means.

Training and environmental variation are propagated through 100000 hierarchical bootstrap replicates. Each replicate samples ten policy seeds with replacement as a common seed set across environments, samples two sites with replacement, and samples four seasons within each selected site. Texture outcomes remain grouped. The 2.5th and 97.5th percentiles provide an exploratory sensitivity range for the pooled mean under this two-site design. The resampling generator uses seed 3407. Ablation mean differences use one-sided enumerated permutation tests, and convergence frequencies use one-sided Fisher exact tests over ten seeds per condition.

V. RESULTS AND DISCUSSION

A. Reward Candidate Comparison

Table I and Fig. 3 compare eight reward candidates with the hand-written reference on six design-selection environments. Three seeds screen each candidate, with two additional seeds for each leader. Mean net benefits span approximately a factor of four and a half, showing the influence of the daily learning signal under a common training protocol. Table I reports reward-screening outcomes, and Table II reports the subsequent transfer evaluation.

The terminal-only and daily-transpiration candidates return 313 and 304 USD ha⁻¹, compared with 967 for the hand-written reference. The stage-weighted stress candidate returns 364, stage-weighted attainment returns 526, and attainment with a leaching penalty returns 1186. Potential-difference shaping gives 715; allocation pacing gives 1398; and the anticipatory candidate gives 1296. The two leading candidates return 1398 ± 239 and 1296 ± 264 over five seeds. The paced program has the higher selection mean and is retained for final policy training.

The paced candidate exceeds the untuned hand-written reference by approximately 45 per cent on the design-selection set. The reference uses a fixed conventional moisture-tracking reward. The four-generation workflow progresses from sparse harvest feedback to daily attainment, nitrogen loss and allocation pacing, with researcher-written behavioral feedback guiding three generations. The selection outcome identifies a useful reward configuration for the subsequent policy experiments.

TABLE I. REWARD CANDIDATES AND THE HAND-WRITTEN REFERENCE ON THE DESIGN-SELECTION ENVIRONMENTS

Gen.	Candidate	Content	Seeds	Net benefit (USD ha ⁻¹)
Ref.	Hand-written reference	Target tracking, linear water price, terminal yield bonus	3	967 ± 352
1	Terminal net benefit	The objective stated literally; one scalar per 140 decisions	3	313 ± 58
1	Daily transpiration credit	Credit for unstressed transpiration on every day	3	304 ± 42
2	K_y -weighted stress penalty	FAO-33 stage weights, applied as a penalty on stress	3	364 ± 63
2	K_y -weighted attainment	Same weights, rewarding attained transpiration instead	3	526 ± 248
2	Attainment less leaching	Adds a heuristic penalty on nitrogen displaced below the root zone	3	$1\,186 \pm 379$
3	Potential-based shaping	Adds the difference of the depletion potential, Eq. (1)	3	715 ± 174
3	Shaping and allocation pacing	Adds the budget pacing term, Eq. (2); selected	5	$1\,398 \pm 239$
4	Anticipation terms	Charges redundant applications and projected depletion	5	$1\,296 \pm 264$

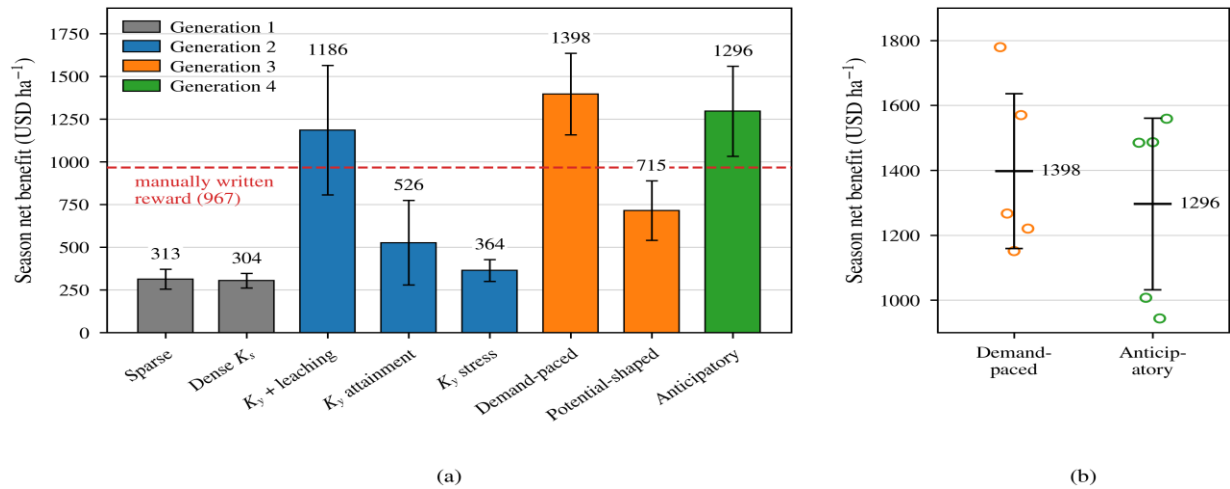


Fig. 3. Design-selection net benefit of the eight reward candidates and the two leading candidates.

TABLE II. POLICY PERFORMANCE ON THE 24 SIMULATED TRANSFER SCENARIOS

Policy	Net benefit (USD ha ⁻¹)	Yield (t ha ⁻¹)	Water (mm)	WP (kg m ⁻³)	Events	N leached (kg ha ⁻¹)	Stress (d)	vs. rule (%)
Rainfed	944	4.29	0	0.975	0.0	13.1	83.3	-50.1
Fixed threshold	1 812	9.51	362	1.468	12.5	20.5	38.0	-4.2
FAO-56 rule	1 782	9.32	364	1.426	10.9	22.9	37.7	-5.8
Tuned threshold stage	1 865	9.64	360	1.481	9.5	26.9	44.8	-1.4
Tuned FAO-56 rule	1 891	9.80	362	1.501	10.6	16.8	43.0	—
PPO (n = 10)	1 943 ± 53	10.03 ± 0.23	362	1.526 ± 0.035	10.3	18.8	36.9	+2.8
SAC (n = 5)	1 803 ± 39	9.46 ± 0.17	363	1.438 ± 0.022	12.0	27.4	44.5	-4.7

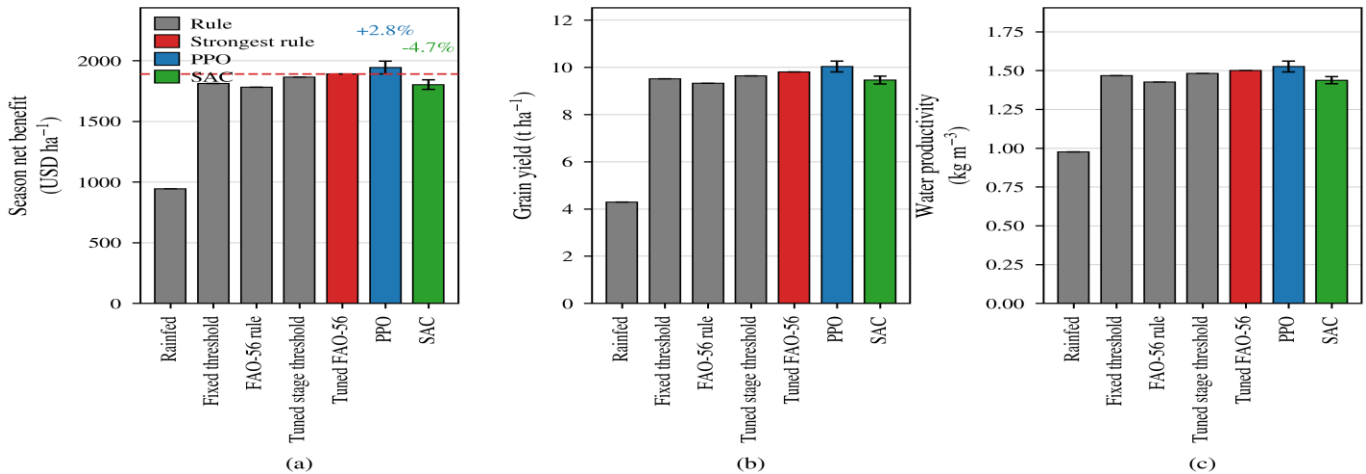


Fig. 4. Transfer net benefit, grain yield and total-input water productivity; error bars show training-seed standard deviations.

B. Transfer Performance

Table II and Fig. 4 report every policy on the 24 transfer environments.

The tuned FAO-56 rule is the strongest non-learning policy at 1891 USD ha⁻¹; the tuned stage rule returns 1865, the fixed threshold 1812, the untuned FAO-56 rule 1782, and rainfed 944. PPO achieves 1943 ± 53 USD ha⁻¹ over ten seeds, 2.8 per cent above the strictest rule and 9.1 per cent above the untuned FAO-56 rule. At approximately 362 mm of irrigation for both, PPO

yields 10.03 ± 0.23 t ha⁻¹ against 9.80, raises WP from 1.50 to 1.53 kg m⁻³, and reduces stress days from 43.0 to 36.9. Leached nitrogen rises from 16.8 to 18.8 kg ha⁻¹.

SAC returns 1803 ± 39 USD ha⁻¹, 4.7 per cent below the strictest rule. It uses 12.0 applications against 10.6 for the tuned rule, leaches 27.4 kg N ha⁻¹ against 16.8, and records 44.5 stress days against 43.0. The entropy sensitivity in Section III-C links exploration intensity to its scheduling performance.

C. Site-Specific Performance

Fig. 5 groups the 24 paired differences by site, showing the location-specific structure of the pooled margin.

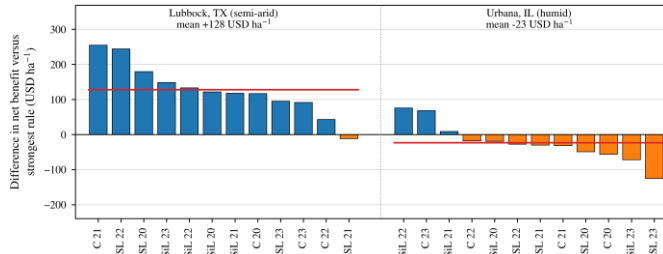


Fig. 5. Paired PPO-tuned-rule net-benefit differences across transfer scenarios, grouped by site.

PPO's paired mean gain is 52.5 USD ha⁻¹, with a median of 55.8 and positive differences in 14 of 24 scenarios. Grouping the shared-weather textures into eight site-season blocks gives a two-sided sign-flip p-value of 0.1172. Joint resampling of sites, seasons, and policy seeds gives an exploratory percentile range of -43.3 to 159.6 USD ha⁻¹, while the two-site sign-flip sensitivity analysis gives $p = 1.000$. The site means are +127.9

USD ha⁻¹ at Lubbock and -23.0 at Urbana, with positive differences in 11 of 12 and 3 of 12 scenarios, respectively. The scheduling margin is concentrated at the semi-arid site. In Fig. 5, SL, SiL, and C denote sandy loam, silt loam, and clay, and the final two digits identify the season.

Fig. 6 illustrates the daily schedules and crop-water states for one transfer season at each site.

At Lubbock on silt loam in 2020, with aridity index 0.126, both policies apply the 672 mm quota. PPO reaches 7.98 t ha⁻¹ against 7.24 for the tuned rule, uses 17 starts against 19, and records 49 stress days against 82. It accepts more vegetative stress when K_y is 0.40, and preserves water through flowering when K_y is 1.50. This schedule follows the stage-sensitive allocation principle described by Fereres and Soriano [7] and represented by the attainment and pacing terms. At Urbana on silt loam in 2020, both policies apply 112 mm in three starts and receive 454 mm of rainfall. PPO returns 12.21 t ha⁻¹ against 12.39 and leaches 24.7 kg N ha⁻¹ against 16.9. The contrasting schedules show the value of water redistribution under scarcity and its drainage consequences in a wetter season.

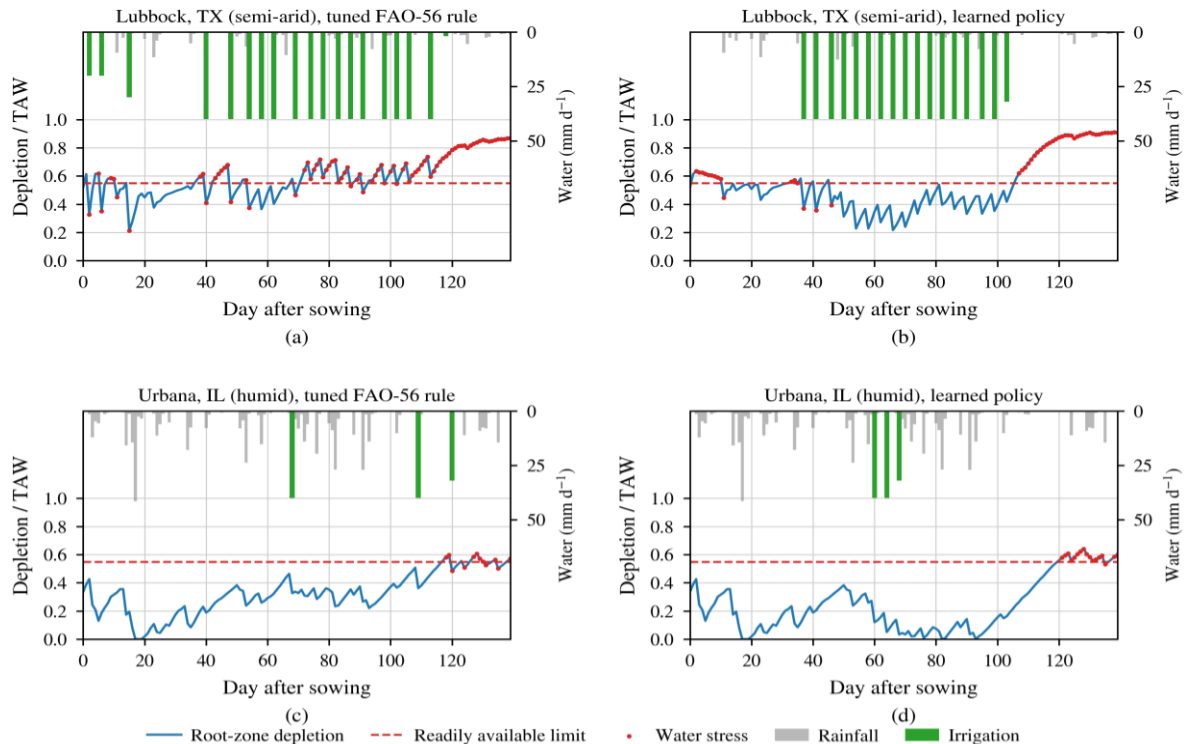


Fig. 6. Daily irrigation and crop-water states for illustrative Lubbock and Urbana transfer seasons in 2020.

TABLE III. TEN-SEED ABLATION OF THE CURRICULUM AND EXPLORATION MASK, SEPARATING CONVERGENCE FREQUENCY FROM CONVERGED PERFORMANCE

Condition	All seeds (USD ha ⁻¹)	Converged	Converged only (USD ha ⁻¹)	p (mean)	p (freq.)	p (converged only)
Full method	1 943 ± 53	10/10	1 943 ± 53	—	—	—
Without the curriculum	1 815 ± 220	7/10	1 957 ± 38	0.073	0.105	0.696
Without the behavior prior	1 768 ± 222	7/10	1 911 ± 47	0.013	0.105	0.123
Without either	1 660 ± 274	5/10	1 928 ± 56	0.005	0.016	0.311

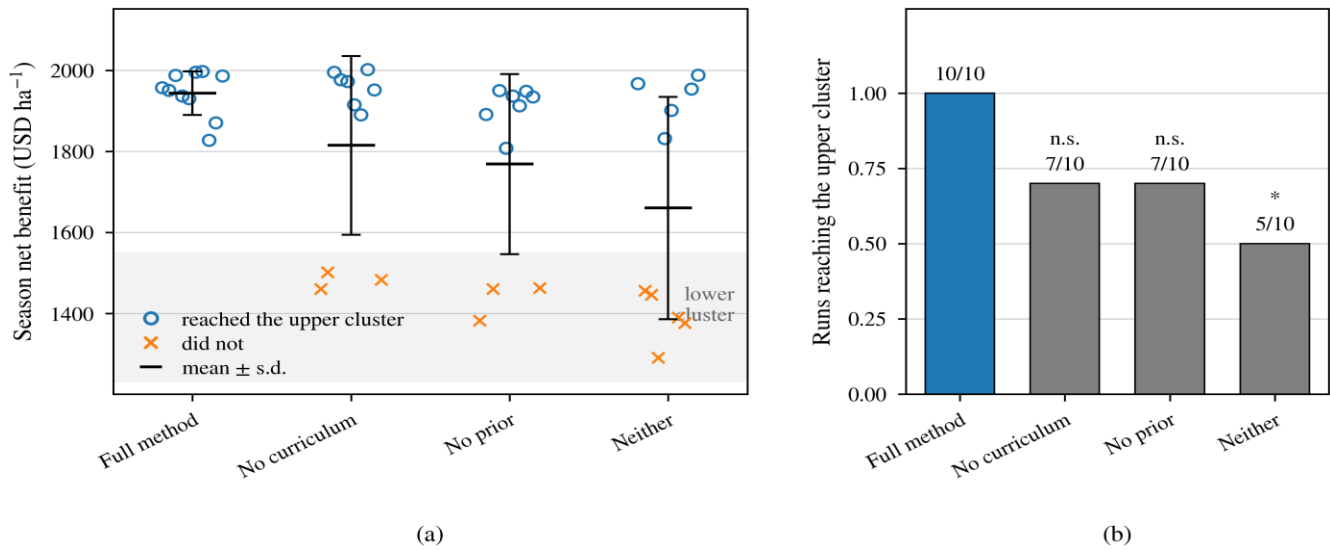


Fig. 7. Seed-level ablation outcomes and convergence frequencies for the four training configurations.

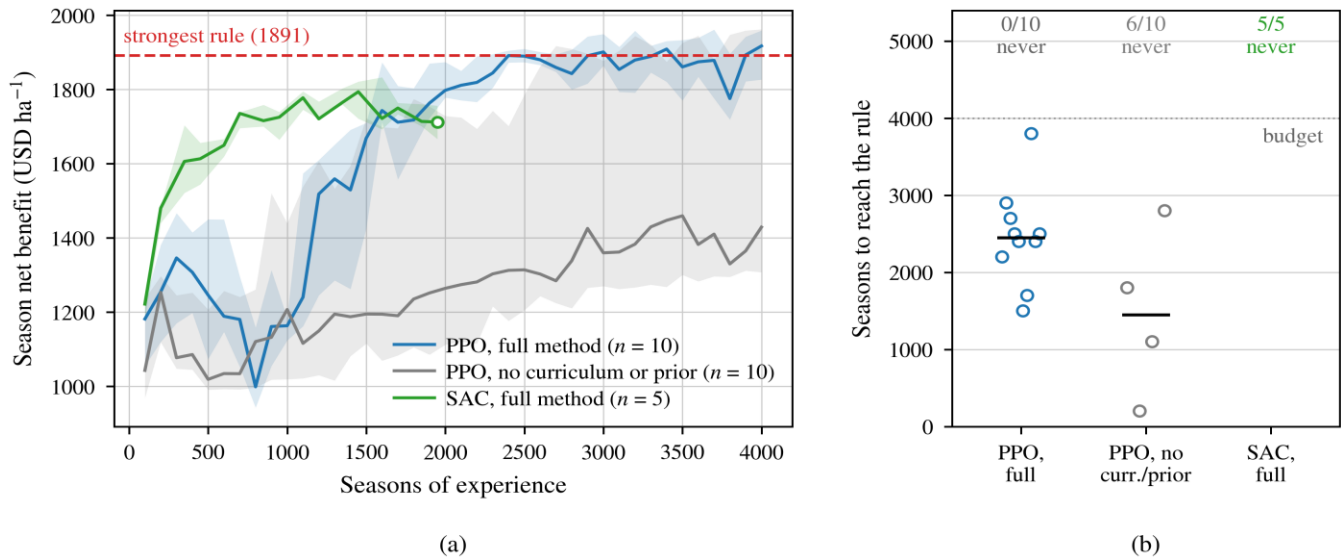


Fig. 8. Validation learning curves and seasons to a common net-benefit reference level.

D. Ablation and Sample Efficiency

Table III and Fig. 7 report the ablation over ten seeds in each of four conditions: the full method, without the curriculum, without the behavior prior, and without either.

The ablation outcomes form two clusters: transfer yields of 9.4–10.25 t ha⁻¹ for stronger runs and 6.9–8.3 for weaker runs. A 9.0 t ha⁻¹ threshold separates these modes and defines convergence in Table III. Convergence occurs in 10/10 full-method runs, 7/10 without the curriculum, 7/10 without the prior, and 5/10 without either. Among converged runs, mean net benefits are 1943, 1957, 1911 and 1928 USD ha⁻¹; one-sided permutation p-values against the full method are 0.696, 0.123 and 0.311. Raw-mean comparisons give $p = 0.073$, 0.013, and 0.005. Convergence-frequency comparisons give $p = 0.105$ for each removal and $p = 0.016$ for joint removal. The combined training aids are associated with more frequent convergence, while successful runs reach comparable net-benefit levels.

The ablation evaluates the curriculum ordering and exploration mask as fixed training interventions. Eq. (4) and the FAO-56 rule are deterministic: direct computation from the same descriptors and tie-breaking rule reproduces the ordering and admitted actions. This equivalence identifies the implemented training components as the objects of the ablation.

Fig. 8 reads sample efficiency off the same runs. Fig. 8(a) shows the median and interquartile range of validation net benefit across seeds. Fig. 8(b) records the first validation evaluation reaching 1891 USD ha⁻¹, the strictest rule's transfer mean, as a common reference level. All ten full-method runs reach it, at a median of 2450 seasons and a range of 1500–3800. The crossing counts are 7/10 without the curriculum, 6/10 without the prior, and 4/10 without either; SAC reaches it in 0/5 runs within 2000 seasons. These validation crossing counts complement the transfer-yield convergence criterion in Table III. Crossing times are conditional on reaching the reference and are interpreted together with the corresponding success counts.

E. Limitations

The study follows one human-guided reward-design trajectory and compares its retained programs under a common simulation protocol. The 45 per cent selection gain uses an untuned reward reference, and the ablation evaluates fixed curriculum and masking components. Equal-budget random or Bayesian reward search and a random-order curriculum would quantify the contribution of the design process and the chosen ordering. Repeated design trajectories with specified generating models, prompts, and decoding settings would assess generation repeatability.

The simulator couples FAO-based water dynamics to a single-cell nitrogen mechanism developed after the initial water-only comparison. Its twelve worked-example checks cover the source equations; validation of the coupled model against DSSAT, APSIM, AquaCrop or field observations remains an important next step. The quota uses a retrospective threshold-rule reference, and seasonal aridity is calculated from archived weather. Prospective evaluation requires independently specified allocations and descriptors available before or during the growing season.

The transfer set spans eight site–season weather realizations across two sites, with three textures sharing each realization. This design supports the reported site comparison and exploratory uncertainty analysis. Additional sites, quota levels, and crops would establish how consistently the semi-arid/humid contrast appears across agricultural settings.

VI. CONCLUSION

This study develops an LLM-assisted offline reward-design workflow for irrigation reinforcement learning. Agronomic knowledge and researcher feedback guide four generations of executable rewards, followed by policy training with a fixed curriculum and exploration mask. The selected reward improves design-selection net benefit by approximately 45 per cent over the conventional reference. Across 24 transfer scenarios, PPO returns 1943 USD ha⁻¹ and a mean margin of 2.8 per cent over the strictest tuned rule. The site means of +127.9 USD ha⁻¹ at Lubbock and -23.0 at Urbana reveal the concentration of scheduling gains under semi-arid conditions. The combined curriculum and mask raise convergence frequency from 5/10 to 10/10 in the evaluated ablation. Together, the reward comparison, transfer analysis, and ablation connect agronomic reward structure to learned seasonal water allocation.

The practical implication is to align irrigation rewards with crop-stage sensitivity and the timing of a limited water allocation. Preserving water through flowering improves the illustrated water-constrained schedule, while humid-season outcomes emphasize drainage and nutrient retention. The retained reward programs provide a basis for extending this analysis to independently validated crop models and broader irrigation regimes.

REFERENCES

- [1] Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L., McMahon, P. B. (2012). Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. In: Proceedings of the National Academy of Sciences. <https://doi.org/10.1073/pnas.1200311109>.
- [2] Deines, J. M., Kendall, A. D., Butler, J. J., Hyndman, D. W. (2019). Quantifying irrigation adaptation strategies in response to stakeholder-driven groundwater management in the US High Plains Aquifer. *Environmental Research Letters*, 14(4), 044014.
- [3] Drysdale, K. M., Hendricks, N. P. (2018). Adaptation to an irrigation water restriction imposed through local governance. *Journal of Environmental Economics and Management*, 91, 150-165.
- [4] Deines, J. M., Kendall, A. D., Butler, J. J., Basso, B., Hyndman, D. W. (2021). Combining Remote Sensing and Crop Models to Assess the Sustainability of Stakeholder-Driven Groundwater Management in the US High Plains Aquifer. *Water Resources Research*, 57(3), e2020WR027756.
- [5] Allen, R. G., Pereira, L. S., Raes, D., Smith, M. (1998). Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. Food and Agriculture Organization of the United Nations, FAO Irrigation and Drainage Paper No. 56, Rome, Italy. <https://www.fao.org/4/x0490e/x0490e00.htm>.
- [6] Doorenbos, J., Kassam, A. H. (1979). Yield Response to Water. Food and Agriculture Organization of the United Nations, FAO Irrigation and Drainage Paper No. 33, Rome, Italy. <https://agris.fao.org/search/en/records/64724f9353aa8c896305d1b7>.
- [7] Fereres, E., Soriano, M. A. (2007). Deficit irrigation for reducing agricultural water use. *Journal of Experimental Botany*, 58(2), 147-159.
- [8] Sun, L., Yang, Y., Hu, J., Porter, D., Marek, T., Hillyer, C. (2017). Reinforcement Learning Control for Water-Efficient Agricultural Irrigation. In: *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pp. 1334-1341. <https://doi.org/10.1109/ISPA/IUCC.2017.00203>.
- [9] 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.
- [10] Saikai, Y., Peake, A., Chenu, K. (2023). Deep reinforcement learning for irrigation scheduling using high-dimensional sensor feedback. *PLOS Water*, 2(9), e0000169.
- [11] Kelly, T. D., Foster, T., Schultz, D. M. (2024). Assessing the value of deep reinforcement learning for irrigation scheduling. *Smart Agricultural Technology*, 7, 100403.
- [12] Alkaff, M., Basuhail, A., Sari, Y. (2025). Optimizing Water Use in Maize Irrigation with Reinforcement Learning. *Mathematics*, 13(4), 595.
- [13] Ding, X., Du, W. (2022). DRLIC: Deep Reinforcement Learning for Irrigation Control. In: *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 41-53. <https://doi.org/10.1109/ipsn54338.2022.00011>.
- [14] Ding, X., Du, W. (2024). Optimizing Irrigation Efficiency using Deep Reinforcement Learning in the Field. *ACM Transactions on Sensor Networks*, 20(4), 1-34. <https://doi.org/10.1145/3662182>.
- [15] Overweg, H., Berghuijs, H. N. C., Athanasiadis, I. N. (2021). CropGym: a Reinforcement Learning Environment for Crop Management. arXiv preprint. arXiv:2104.04326. <https://arxiv.org/abs/2104.04326>.
- [16] Gautron, R., Padrón, E. J., Preux, P., Bigot, J., Maillard, O. A., Emukpere, D. (2022). gym-DSSAT: a crop model turned into a Reinforcement Learning environment. arXiv preprint. arXiv:2207.03270. <https://arxiv.org/abs/2207.03270>.
- [17] Turchetta, M., Corinzia, L., Sussex, S., Burton, A., Herrera, J., Athanasiadis, I., Buhmann, J. M., Krause, A. (2022). Learning Long-Term Crop Management Strategies with CyclesGym. In: *Advances in Neural Information Processing Systems (NeurIPS 2022)*, Datasets and Benchmarks Track, pp. 11396-11409. <https://doi.org/10.52202/068431-0828>.
- [18] Ma, Y. J., Liang, W., Wang, G., Huang, D. A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., Anandkumar, A. (2024). Eureka: Human-Level Reward Design via Coding Large Language Models. International Conference on Learning Representations (ICLR 2024). arXiv:2310.12931. <https://openreview.net/forum?id=IEduRUO55F>.
- [19] Xie, T., Zhao, S., Wu, C. H., Liu, Y., Luo, Q., Zhong, V., Yang, Y., Yu, T. (2024). Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. International Conference on Learning

- Representations (ICLR 2024). arXiv:2309.11489. <https://openreview.net/forum?id=tUM39YTRxH>.
- [20] Yu, W., Gileadi, N., Fu, C., Kirmani, S., Lee, K. H., Gonzalez Arenas, M., Chiang, H. T. L., Erez, T., Hasenclever, L., Humplik, J., Ichter, B., Xiao, T., Xu, P., Zeng, A., Zhang, T., Heess, N., Sadigh, D., Tan, J., Tassa, Y., Xia, F. (2023). Language to Rewards for Robotic Skill Synthesis. In: *Proceedings of the 7th Conference on Robot Learning (CoRL 2023)*, PMLR vol. 229, pp. 374-404. <https://proceedings.mlr.press/v229/yu23a.html>.
- [21] Kwon, M., Xie, S. M., Bullard, K., Sadigh, D. (2023). Reward Design with Language Models. International Conference on Learning Representations (ICLR 2023). arXiv:2303.00001. <https://openreview.net/forum?id=10uNUg15Kl>.
- [22] Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A. (2024). Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research (TMLR)*. arXiv:2305.16291. <https://openreview.net/forum?id=ehfRiF0R3a>.
- [23] Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, 18(3-4), 235-265.
- [24] Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N. G. (2003). An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3-4), 267-288.
- [25] Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., Chenu, K., van Oosterom, E. J., Snow, V., Murphy, C. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62, 327-350.
- [26] Steduto, P., Hsiao, T. C., Raes, D., Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles. *Agronomy Journal*, 101(3), 426-437.
- [27] Kelly, T. D., Foster, T., Schultz, D. M. (2023). Assessing the value of adapting irrigation strategies within the season. *Agricultural Water Management*, 275, 107986.
- [28] Alibabaei, K., Gaspar, P. D., Assunção, E., Alirezazadeh, S., Lima, T. M. (2022). Irrigation optimization with a deep reinforcement learning model: Case study on a site in Portugal. *Agricultural Water Management*, 263, 107480.
- [29] 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.
- [30] Wu, J., Tao, R., Zhao, P., Martin, N. F., Hovakimyan, N. (2022). Optimizing Nitrogen Management with Deep Reinforcement Learning and Crop Simulations. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1711-1719. <https://doi.org/10.1109/cvprw56347.2022.00178>.
- [31] Huang, W., Abbeel, P., Pathak, D., Mordatch, I. (2022). Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In: *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, PMLR vol. 162, pp. 9118-9147. <https://proceedings.mlr.press/v162/huang22a.html>.
- [32] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoy, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P. (2023). Self-Refine: Iterative Refinement with Self-Feedback. In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. <https://arxiv.org/abs/2303.17651>.
- [33] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. <https://arxiv.org/abs/2303.11366>.
- [34] Tao, R., Zhao, P., Wu, J., Martin, N. F., Harrison, M. T., Ferreira, C., Kalantari, Z., Hovakimyan, N. (2023). Optimizing Crop Management with Reinforcement Learning and Imitation Learning. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pp. 6228-6236. <https://doi.org/10.24963/ijcai.2023/691>.
- [35] Ng, A. Y., Harada, D., Russell, S. J. (1999). Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In: *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, pp. 278-287.
- [36] Devlin, S., Kudenko, D. (2011). Theoretical considerations of potential-based reward shaping for multi-agent systems. In: *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, pp. 225-232. <https://doi.org/10.65109/vdid3904>.
- [37] Mannion, P., Devlin, S., Mason, K., Duggan, J., Howley, E. (2017). Policy invariance under reward transformations for multi-objective reinforcement learning. *Neurocomputing*, 263, 60-73.
- [38] Eck, A., Soh, L. K., Devlin, S., Kudenko, D. (2016). Potential-based reward shaping for finite horizon online POMDP planning. *Autonomous Agents and Multi-Agent Systems*, 30(3), 403-445.
- [39] Bengio, Y., Louradour, J., Collobert, R., Weston, J. (2009). Curriculum learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pp. 41-48. <https://doi.org/10.1145/1553374.1553380>.
- [40] Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., Stone, P. (2020). Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. *Journal of Machine Learning Research*, 21(181), 1-50.
- [41] Florensa, C., Held, D., Wulfmeier, M., Zhang, M., Abbeel, P. (2017). Reverse Curriculum Generation for Reinforcement Learning. In: *Proceedings of the 1st Annual Conference on Robot Learning (CoRL 2017)*, PMLR vol. 78, pp. 482-495. <https://proceedings.mlr.press/v78/florensa17a.html>.
- [42] Allen, R. G., Pereira, L. S., Smith, M., Raes, D., Wright, J. L. (2005). FAO-56 Dual Crop Coefficient Method for Estimating Evaporation from Soil and Application Extensions. *Journal of Irrigation and Drainage Engineering*, 131(1), 2-13.
- [43] Pereira, L. S., Allen, R. G., Smith, M., Raes, D. (2015). Crop evapotranspiration estimation with FAO56: Past and future. *Agricultural Water Management*, 147, 4-20.
- [44] Saxton, K. E., Rawls, W. J. (2006). Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions. *Soil Science Society of America Journal*, 70(5), 1569-1578.
- [45] Raes, D., Steduto, P., Hsiao, T. C., Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agronomy Journal*, 101(3), 438-447.
- [46] Burns, I. G. (1974). A model for predicting the redistribution of salts applied to fallow soils after excess rainfall or evaporation. *Journal of Soil Science*, 25(2), 165-178.
- [47] Ciampitti, I. A., Vyn, T. J. (2012). Physiological perspectives of changes over time in maize yield dependency on nitrogen uptake and associated nitrogen efficiencies: A review. *Field Crops Research*, 133, 48-67.
- [48] Geerts, S., Raes, D. (2009). Deficit irrigation as an on-farm strategy to maximize crop water productivity in dry areas. *Agricultural Water Management*, 96(9), 1275-1284.
- [49] Payero, J. O., Tarkalson, D. D., Irmak, S., Davison, D., Petersen, J. L. (2009). Effect of timing of a deficit-irrigation allocation on corn evapotranspiration, yield, water use efficiency and dry mass. *Agricultural Water Management*, 96(10), 1387-1397.
- [50] Foster, T., Brozović, N., Butler, A. P., Neale, C. M. U., Raes, D., Steduto, P., Fereres, E., Hsiao, T. C. (2017). AquaCrop-OS: An open source version of FAO's crop water productivity model. *Agricultural Water Management*, 181, 18-22.
- [51] Perez-Quesada, G., Hendricks, N. P. (2021). Lessons from local governance and collective action efforts to manage irrigation withdrawals in Kansas. *Agricultural Water Management*, 247, 106736.
- [52] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv preprint. arXiv:1707.06347. <https://arxiv.org/abs/1707.06347>.
- [53] Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P. (2016). High-Dimensional Continuous Control Using Generalized Advantage

- Estimation. arXiv preprint (also ICLR 2016). arXiv:1506.02438. <https://arxiv.org/abs/1506.02438>.
- [54] Haarnoja, T., Zhou, A., Abbeel, P., Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, PMLR vol. 80, pp. 1861-1870. <https://proceedings.mlr.press/v80/haarnoja18b.html>.
- [55] Christodoulou, P. (2019). Soft Actor-Critic for Discrete Action Settings. arXiv preprint. arXiv:1910.07207. <https://arxiv.org/abs/1910.07207>.
- [56] NASA Langley Research Center 2024. POWER: Prediction of Worldwide Energy Resources, Agroclimatology Daily Point Data. NASA Langley Research Center, Applied Sciences Program. Data accessed 2026-08-26. <https://power.larc.nasa.gov>.
- [57] Sparks, A. H. (2018). nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. *Journal of Open Source Software*, 3(30), 1035. <https://doi.org/10.21105/joss.01035>.
- [58] Zhang, T., Stackhouse, P. W., Chandler, W. S., Westberg, D. J. (2014). Application of a global-to-beam irradiance model to the NASA GEWEX SRB dataset: An extension of the NASA Surface meteorology and Solar Energy datasets. *Solar Energy*, 110, 117-131.