INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Environment and Sustainable Development

ISSN online: 1478-7466 - ISSN print: 1474-6778
https://www.inderscience.com/ijesd

Environmental variable regulation and optimisation strategy in forest seedling cultivation based on reinforcement learning

Xianghong Liu

DOI: 10.1504/IJESD.2026.10077680

Article History:

Received:	29 October 2025
Last revised:	07 February 2026
Accepted:	11 March 2026
Published online:	18 June 2026

Environmental variable regulation and optimisation strategy in forest seedling cultivation based on reinforcement learning

Xianghong Liu

Qingyang Forestry and Grass Seedling Station,
Qingyang Forestry and Grassland Bureau,
Qingyang, 745000, Gansu China
Email: lxh_1017@163.com

Abstract: Addressing reinforcement learning's failure to model high-dimensional dynamic interactions in seedling cultivation, this paper proposes an intelligent regulation strategy integrating species embedding and proximal policy optimisation (PPO) for precise, dynamic environmental management of diverse seedlings. The study constructs a 12-dimensional state space encompassing initial biological traits, real-time environmental parameters, and species encoding, designs a 243-dimensional discrete action space encompassing temperature, humidity, light, water, and CO_2, and introduces a multi-objective reward function to co-optimise growth rate, survival rate, resource efficiency, and stress avoidance. The strategy improves the average daily growth rate, with a final survival rate of 95.2%, an average reduction in electricity consumption per unit biomass of approximately 0.4 kWh/g, and an average reduction in stress events of 19.86. Furthermore, a transfer learning mechanism enables fine-tuning for new tree species in just seven days. This study provides a generalisable, efficient, and personalised regulation paradigm for intelligent seedling cultivation.

Keywords: reinforcement learning; forest tree seedling cultivation; environmental variable regulation; environmental optimisation strategies; proximal strategy optimisation.

Reference to this paper should be made as follows: Liu, X. (2026) 'Environmental variable regulation and optimisation strategy in forest seedling cultivation based on reinforcement learning', *Int. J. Environment and Sustainable Development*, Vol. 25, No. 6, pp.63–86.

Biographical notes: Xianghong Liu graduated from Southwest University of Science and Technology, China, and was conferred a Master's degree. Presently, he is employed at Qingyang Forestry and Grass Seedling Station in Gansu Province. His research interests encompass the cultivation and management of forest and grass seedlings, along with the cultivation and breeding of medicinal plants.

1 Introduction

With the increasing urgency of global climate change and ecological restoration, intelligent technology has become a key component of modern forestry development (Mahmood et al., 2024; Wang et al., 2024a; Hoppen et al., 2024). Traditional seedling

cultivation often relies on empirical thresholds to control environmental factors, making it difficult to dynamically respond to the coupling effects of individual differences in seedlings and multidimensional environments, leading to resource waste, uneven growth, and even stress-induced death (Socha et al., 2022; Alotaibi et al., 2023; Augusto et al., 2025). In recent years, reinforcement learning has shown potential in agricultural greenhouse regulation. Still, it has not yet been systematically applied to forest seedlings, which have long growth cycles, strong species specificity, and high individual heterogeneity. In particular, it cannot model the high-dimensional interactions between 'species-initial traits-environment-stress' (Platero-Horcajadas et al., 2024; Maraveas et al., 2023; Ajagekar et al., 2023).

This paper focuses on the dynamic environmental regulation of intelligent tree seedling cultivation. The research subjects include typical afforestation species such as Masson pine, Chinese fir, *Larix gmelinii*, and *Pinus tabulaeformis*. These species not only have high ecological value and large seedling production requirements, but also have significant differences in physiological characteristics, placing higher demands on precise regulation (Li et al., 2024a; Liu et al., 2024; Han et al., 2025). The focus is on their differentiated responses to environmental factors, including temperature, humidity, light, water, and CO_2, during the early stages of seedling production. In recent years, many scholars have conducted in-depth exploration of species-specific response mechanisms: Wang et al. (2022a, 2024c) conducted experimental studies on the growth characteristics and utilisation of various factors in Masson pine; Li et al. (2023a) studied the absorption, accumulation, and distribution of mineral nutrients in Chinese fir under drought stress; Huang et al. (2023) analysed the effects of drought stress on non-structural carbohydrates in different organs of Chinese fir; Li et al. (2024a) and Jin et al. (2024) studied the relationship between larch development and soil water resources. These studies collectively reveal significant heterogeneity in the sensitivity of different tree species to environmental factors, and an urgent need for an adaptive regulatory strategy that integrates individual initial states with tree species-specific information to achieve 'policy based on species and control based on seedlings'.

To address the issue of environmental regulation of trees or crops, existing studies have introduced a variety of technologies and algorithms, but these still face significant limitations. Deep learning has been widely used in agricultural smart greenhouses, but it is difficult to distinguish individual needs (Akbar et al., 2024); Lee et al. (2024) proposed a hybrid deep learning model to achieve automated and customised greenhouse operations, but their strategy had poor stability in forestry systems; Iwasaki et al. (2022) conducted environmental control on *Arabidopsis*, and Javaid et al. (2022) conducted environmental control on ryegrass, but neither of them conducted cross-species migration; Xuan et al. (2024) proposed the application of multi-agent reinforcement learning in the coordinated optimisation of stand structure in Yunnan pine secondary forests, but the initial characteristics of seedlings have not been effectively modelled; Shi et al. (2024) used deep learning integrated with the sparrow search algorithm to optimise greenhouse microclimate prediction to adapt to the growth environment of seedlings, but it did not match the characteristics of forestry; Sun and Chang (2023) combined dynamic factors with machine learning to achieve microclimate prediction, but only used environmental variables as state inputs and did not encode seedling information. In general, existing methods have three major defects: first, they ignore biological traits and species identification, making it difficult to distinguish individual needs; second, they rely on continuous control, which does not match the characteristics of forestry; third,

they lack an efficient cross-species migration mechanism, and the deployment of new varieties requires training data from scratch, which makes it difficult to meet the actual needs of the tight forestry production cycle.

To address the above issues, this paper proposes a reinforcement learning control framework that integrates species embedding and proximal policy optimisation. We have chosen to implement the proximal policy optimisation algorithm because it is best suited to our problem among the available options, including deep Q-networks (DQN) and soft actor-critic (SAC). In our case, we have defined our action space as discrete and high-dimensional, with 243 unique composite actions. The issue with DQN learning in high-dimensional discrete action spaces is that each action has a Q-value, which can lead to instability and poor performance during training. Although SAC works well for continuous control problems, it does not readily apply to the discrete action requirements that are imposed by the operational requirements of forestry cultivation equipment. The advantage of using proximal policy optimisation is that it uses policy gradient methods to effectively apply to high-dimensional discrete action spaces. First, based on biological traits and species identification, a 16-dimensional state space is constructed, consisting of 4-dimensional initial traits, 7-dimensional environmental variables, and 5-dimensional learnable species embedding. Second, based on forestry characteristics, a 243-dimensional discrete action space is designed, accurately mapped to five types of actuators: temperature, humidity, light, water, and CO_2.

Furthermore, a multi-objective reward function that accounts for growth, survival, energy conservation, and stress avoidance is constructed. The policy stability is improved through a two-stage training process that combines simulation pre-training with real-world fine-tuning. To address cross-species issues, species embedding and transfer learning mechanisms are introduced. Experiments show that this method improves the average daily growth rate and final survival rate, reduces energy consumption per unit biomass, and, through an efficient parameter transfer mechanism, enables new tree species to achieve a high level with only fine-tuning, thereby verifying the strategy's efficiency, generalisation, and practical value.

2 Related work

The advancement of precision and automation in agriculture/horticulture has enabled improvements in data-driven environmental systems for management. There has been considerable research on the use of intelligent greenhouses to control environmental conditions during crop production; this research has primarily focused on how to automatically monitor and regulate the environment within the production system for optimal growth conditions. However, when attempting to convert these technologies for use in an intelligent greenhouse for the cultivation of forest tree seedlings, the challenges differ from those associated with the aforementioned technologies.

In their study, Xu and Jiang (2025) illustrated how AI can change from traditional forestry management systems to an advanced new way of managing forests by demonstrating the importance of incorporating AI into forestry management systems in the future; Castro et al. (2024) used drones and AI technologies for automatic and accurate forest seedling planting and reduced the economic and ecosystemic cost associated with planting and increased the survival rate of seedlings. The study conducted

by Zhao et al. (2024) examined the potential of using multi-agent reinforcement learning in forestry as well as the benefits that it offers. The findings from this research should provide new ideas and methodologies for obtaining sustainable management of forest ecosystems. In Wang et al. (2025b), the authors examined how artificial intelligence, machine learning, and deep learning could transform how we manage forests sustainably. The authors provided examples of how these technologies can be applied to the field of sustainable forestry through the use of predictive analytics and modelling approaches which provide insight into how forests will change over time related to carbon sequestration, distribution of various species within the forest and ongoing health of ecosystems; Fan et al. (2024) used a two-step process to create a system for identifying drought stress in the seedlings of the poplar tree. This system will also provide information on the drought-related strength of poplar trees; thus, it can support the management of poplar forests and irrigation management during drought. Wang et al. (2024b) developed a method to classify the different varieties of poplar seedlings and determine their drought stress levels using deep learning and multi-source time-series data, enabling better drought-stress response while maintaining a high degree of classification certainty. Binlajdam et al. (2025) focused on sustainable forest management practices, highlighting recent developments and advancements through the integration of technology-based innovations with policy-directed frameworks. By using deep learning technology, Ecke et al. (2024) helped facilitate species identification and the assessment of canopy conditions through their research, reducing (or even eliminating) many operations that would otherwise need to be performed in person and, as a result, creating dramatic savings in both time and money. Using machine learning algorithms, Caliskan et al. (2024) classified and clustered seeds and seedlings based on morphological characteristics. Their work provides a solid foundation and catalyst for afforestation, yield enhancement, and breeding programs. Ehrlich-Sommer et al. (2024) emphasised the human-machine collaboration approach in AI-powered smart forestry, which can enhance adaptability and effectiveness across various scenarios.

To sum up, while existing research provides an important foundation for forestry environmental control and monitoring and species-specific physiological studies, there is currently no unified, adaptive framework. This framework should dynamically integrate the initial biological characteristics of seedlings and real-time environmental and species-specific information, while operating effectively within the discrete, high-dimensional action spaces created by forestry infrastructure and leveraging transfer learning to rapidly and efficiently adapt to new tree species.

3 Algorithm and model design

3.1 State space construction

To achieve refined dynamic control of the tree seedling cultivation process, it is necessary to construct a high-dimensional state space that comprehensively characterises the coupling between the physiological state of individual seedlings and the environment. In this study, a single seedling is used as the basic control unit, and the state vector $\mathbf{s}_t \in R^{12}$ is defined as the observation input to the reinforcement learning agent at time t.

Initial biological characteristics of each seedling are collected. Seedling height h_0 (cm) and ground diameter d_0 (mm) are automatically measured using a laser rangefinder

on the day of transplanting. Chlorophyll relative content is measured using a SPAD-502 Plus chlorophyll meter, with the average value of three functional leaves taken as the average $SPAD_0$. Root developmental grade $r_{root} \in \{1, 2, 3, 4, 5\}$ is automatically assessed using scanned root images (resolution: 2,400 dpi) using a U-Net semantic segmentation model (Hui et al., 2023; Li et al., 2023b; Xu et al., 2022). The input is an red, green, blue (RGB) image $\mathbf{I} \in \mathrm{R}^{512\times512\times3}$, and the output is a pixel-level root mask $\hat{\mathbf{M}}$. Root volume percentages $\rho = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{M}_{ij}$ are mapped to grade intervals ($\rho < 0.05 \rightarrow 1$, $0.05 \leq \rho < 0.1 \rightarrow 2$, …, $\rho \geq 0.2 \rightarrow 5$).

Environmental variables are collected in real time every 10 minutes via a multi-parameter sensor network deployed within the nursery chamber. These include air temperature T_a, relative humidity RH, photosynthetically active radiation PAR, CO_2 concentration C_{CO_2}, and soil volumetric moisture content θ_v. The seven-dimensional environmental variables are thus fully enumerated as: air temperature T_a, relative humidity RH, photosynthetically active radiation PAR, CO_2 concentration C_{CO_2}, soil volumetric moisture content θ_v, soil temperature T_s, and vapour pressure deficit VPD. Soil temperature (T_s) is measured using a thermocouple probe inserted at a depth of 5 cm adjacent to the seedling root zone. Vapour pressure deficit is calculated from the measured air temperature and relative humidity using the standard Magnus formula. All seven parameters are sampled synchronously every 10 minutes by the sensor network, smoothed via the edge gateway's sliding window, and collectively constitute the environmental component of the state vector. All environmental data is smoothed using a sliding window on the edge computing gateway to suppress transient noise.

Seedling species information is incorporated into the state space via one-hot encoding. Let the set of tree species be S = *Pinus massoniana Lamb.*, *Cunninghamia lanceolata (Lamb.) Hook.*, *Larix gmelinii (Rupr.) Kuzen.*,}, then the species encoding vector $\mathbf{e}_{sp} \in \{0, 1\}^3$ satisfies $\sum_{k=1}^{3} e_{sp,k} = 1$. For example, $\mathbf{e}_{sp} = [1, 0, 0]^\top$ is the corresponding for Masson pine.

Finally, the above variables are concatenated into a 12-dimensional state vector:

$$\mathbf{s}_t = \left[h_0, d_0, SPAD_0, r_{root}, T_a(t), RH(t), PAR(t), C_{CO_2}(t), \theta_v(t), e_{sp,1}, e_{sp,2}, e_{sp,3}\right]^\top \quad (1)$$

In formula (1), the first four terms are fixed initial traits, and the last eight terms are time-varying environmental and species identities. This state-space design explicitly models the three-way interaction among 'individual differences, environmental responses, and species specificity', providing the policy network with discriminative inputs to distinguish among different seedling regulation requirements. The state vector s_t defined in equation (1) has a fixed dimensionality of 12. This dimensionality is determined by concatenating four initial seedling traits, seven real-time environmental variables, and a one-hot species-encoding vector of length 3, yielding 4 + 7 + 1 = 12 dimensions. The design of the dimensions is based on principles; that is, each dimension has a unique, quantifiable value (physical or biological) identified in previous studies as integral to growth dynamics. This representation of a state will be used across the proximal policy optimisation framework. Rather, a hyperparameter, such as the dimensionality, is not subject to sensitivity testing for algorithms that use deep Q-learning; this study instead used policy gradient techniques, allowing the direct use of

the specified fixed-length state vector as input to the policy network. All raw data are Z-score normalised and then fed into the reinforcement learning model:

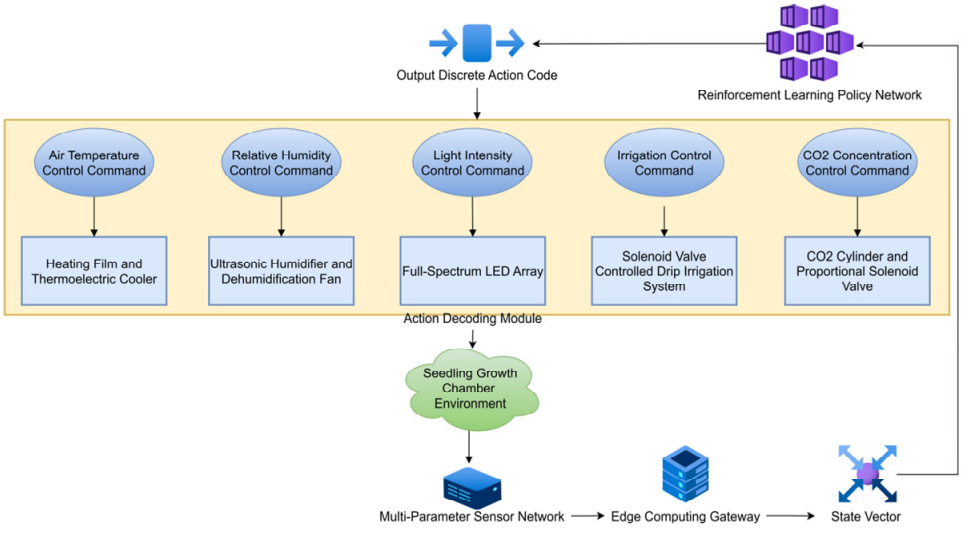
$$\tilde{s}_t^{(i)} = \frac{s_t^{(i)} - \mu^{(i)}}{\sigma^{(i)}} \tag{2}$$

In formula (2), $s_t^{(i)}$ represents the raw value of the i^{th} dimension of the state vector, and $\mu^{(i)}$ and $\sigma^{(i)}$ represent the mean and standard deviation of that dimension over the historical training dataset. This standardisation ensures consistent dimensioning of all features, accelerating the convergence of the policy network. The state vector is updated every 10 minutes and serves as the basis for the PPO agent's decision making, fully supporting the generation of subsequent dynamic control strategies.

3.2 Action space design

To achieve precise dynamic control of the tree seedling cultivation environment, the decision output of the reinforcement learning agent needs to be mapped into executable physical control instructions. This study defines the control action as a discrete adjustment combination of five independent environmental factors, constructs a finite discrete action space (Williams et al., 2022), and establishes a deterministic mapping relationship between it and the underlying actuator. The design flowchart for the mapping between the action space and the environmental control actuator is shown in Figure 1.

Figure 1 Flowchart of mapping between the design action space and the environmental control actuator (see online version for colours)



Five control dimensions are defined and divided into three discrete levels: air temperature control: three action levels $a_T \in \{-2, 0, +2\}$ (unit: °C) corresponding to cooling by 2°C, maintaining, and increasing by 2°C; relative humidity control: $a_H \in \{-10, 0, +10\}$ (unit: %) corresponding to dehumidification by 10%, maintenance, and humidification by 10%; light intensity control: $a_L \in \{-200, 0, +200\}$ (unit: μmol/m^2/s) corresponding to reducing

the fill light by 200, maintaining, and increasing by 200; irrigation control: $a_I \in \{0, 1, 2\}$, representing off, low flow (50 mL/min), and high flow (150 mL/min), respectively; CO_2 concentration control: $a_C \in \{0, 50, 100\}$ (unit: ppm) corresponding to off, injection of 50 ppm, and injection of 100 ppm. These particular step sizes were chosen for two reasons. These reasons are that they provide step sizes for both the physiological response sensitivity of seedlings and a significant range of incremental adjustments, operating at the same level of resolution as actuators commonly used. In terms of temperature, the ±2°C step size corresponds to a range of air temperatures at which tree seedlings experience dramatic stimulation of photosynthesis and respiration rates. For relative humidity, the ±10% step size covers a typical range of changes in stomatal conductance and transpiration. It falls within the operational range of commercially available ultrasonic humidifiers, dehumidifiers, and HVAC equipment used in many commercial nurseries. With respect to light, the ±200 μmol/m^2/s is an incremental saturation point for photosynthetically active radiation, as established for conifer seedlings grown in supplemental light, and has been developed in concert with the dimming capability of commercial-grade LED grow lights.

Regarding water, flow rates have been set at 0, 50, and 150 mL/min, which correspond to the range of maintenance through considerable soil moisture replenishment for conifer seedlings and are based on the common range of flow control segments for solenoid valves in drip irrigation. Finally, CO_2 enrichment steps of 0, 50, and 100 ppm were created to allow fine-tuning CO_2 above ambient levels to promote photosynthesis, while also being operationally achievable with common solenoid-valve-controlled CO_2 dosing systems. Each dimension takes three discrete values, forming a Cartesian product action space:

$$A = \left\{ (a_T, a_H, a_L, a_I, a_C) \middle| a_T \in D_T, \ldots, a_C \in D_C \right\} \tag{3}$$

In formula (3), $D_T = \{-2, 0, +2\}$, $D_H = \{-10, 0, +10\}$, and so on. The total size of the action space is $|A| = 3^5 = 243$. To facilitate the output of the policy network, each compound action is uniquely encoded as an integer $a_t \in \{1, 2, \ldots, 243\}$, using a quinary expansion mapping:

$$a_t = 1 + \sum_{k=1}^{5} d_k \cdot 3^{k-1} \tag{4}$$

In formula (4), $d_k \in \{0, 1, 2\}$ is the index of the k^{th} control dimension. This encoding ensures that actions are assigned unique integer identifiers and are reversible.

3.3 Multi-objective reward function design

To guide reinforcement learning agents to balance multiple objectives such as growth promotion, survival assurance, resource conservation, and stress avoidance during the cultivation of forest seedlings, this paper designs a quantifiable multi-objective weighted reward function R_t, the calculation of which is entirely based on measurable physical quantities and automated recognition results to ensure that the strategy optimisation direction is consistent with forestry production goals (Wang et al., 2022b; Icarte et al., 2022).

A growth rate bonus, $r_{growth,t}$ is defined. The LDM71 laser rangefinder is used to automatically measure the height, h_t (in cm), at 08:00 every day and calculate the daily increment:

$$r_{growth,t} = \frac{h_t - h_{t-1}}{\Delta t}, \quad \Delta t = 1 \text{ day} \tag{5}$$

This value directly reflects the promoting effect of environmental regulation on biomass accumulation.

A survival bonus, $r_{survive,t}$ is defined. If the seedling is considered alive on that day (based on normalised difference vegetation index (NDVI) > 0.3 and no structural damage), then $r_{survive,t}$ = 1; otherwise, it is 0. NDVI is calculated from the RGB image collected at noon every day:

$$NDVI = \frac{I_{NIR} - I_{red}}{I_{NIR} + I_{red}} \tag{6}$$

In formula (6), I_{NIR} and I_{red} represent the reflectance in the near-infrared and red bands, respectively, acquired using a multispectral camera. This metric effectively distinguishes healthy individuals from deceased individuals, avoiding subjective misjudgement. A resource conservation rate reward $r_{eff,t}$ is defined. The system records the total daily energy consumption, E_t (kWh), via a smart meter. It sets a baseline maximum energy consumption $E_{\max}$ = 15 kWh/day (based on the statistical mean of historical high energy consumption threshold control schemes):

$$r_{eff,t} = 1 - \frac{E_t}{E_{\max}} \tag{7}$$

This design's incentive strategy minimises energy consumption while meeting growth requirements. A stress index penalty term, $r_{stress,t}$ is defined. This index consists of two components: wilting indication ($I_{wilting}$) and leaf temperature deviation ($|T_{leaf} - T_{opt}|$). The wilting state is detected using the you only look once (YOLO) v5s model in RGB images. If the confidence level is > 0.7, then $I_{wilting}$ = 1; otherwise, it is 0. Leaf temperature T_{leaf} is measured using a thermal imager, and T_{opt} = 28°C is the optimal leaf temperature for photosynthesis. The weights are set to w_1 = 0.6 and w_2 = 0.4:

$$r_{stress,t} = w_1 \cdot I_{wilting} + w_2 \cdot \frac{|T_{leaf} - 28|}{10} \tag{8}$$

The factor of 10 is used as a multiplier to indicate the degree to which leaf temperature deviates from the optimum temperature of 28°C, determined from long-term field observations and plant measurements. In the historical seedling dataset, leaf temperature deviations exceeding 10°C from the optimum were correlated with visible signs of physiological stress, and photosynthetic efficiency was significantly lower in these plants than in those with proper leaf temperature management. This value effectively maps the typical operational range of observed deviations to the [0, 1] interval for normalisation, ensuring that the temperature deviation component $w_2 \cdot \frac{|T_{leaf} - 28|}{10}$ has a numerical scale comparable to the wilting indicator $w_1 \cdot I_{wilting}$, which is either 0 or 0.6. This balancing allows both components to contribute meaningfully to the composite stress penalty

without one dominating the other due to disparate numerical ranges. Finally, the comprehensive reward function is:

$$R_t = \alpha \cdot r_{growth,t} + \beta \cdot r_{survive,t} + \gamma \cdot r_{eff,t} - \delta \cdot r_{stress,t} \tag{9}$$

The weights in formula (9) are determined through grid search and expert evaluation as $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.2$ and $\delta = 0.1$. The grid search was conducted in the PlantSim simulation environment using a pre-trained base policy to ensure efficiency. The search space for each weight spanned {0.1, 0.2, 0.3, 0.4, 0.5}, constrained by $\alpha + \beta + \gamma + \delta = 1$. Each candidate weight combination was evaluated by running 50 independent simulation episodes for each of the three source tree species, for a total of 150 episodes per combination. The evaluation metric was the mean discounted cumulative return per episode, calculated as $\sum_{t=0}^{T} \gamma^t R_t$. The combination (0.4, 0.3, 0.2, 0.1) yielded the highest mean discounted return. This quantitative result was then reviewed and validated by two forestry domain experts to ensure alignment with practical cultivation priorities, confirming the balance between growth promotion and risk management. Growth is weighted highest; seedling height is the core indicator for seedling placement; survival is weighted second-highest, ensuring a minimum survival rate and avoiding risk-taking in pursuit of growth. Resource efficiency is encouraged while satisfying the first two criteria. Stress is strictly controlled but weighted low because it is already reflected in the penalty term. Excessive avoidance can lead to growth suppression, so a lower weight is assigned to maintain the strategy's exploratory nature. This reflects the forestry seedling cultivation priority of 'growth first, survival guaranteed, energy conservation second, and strict stress control'. The convergence curve for the reward function is shown in Figure 2.

Figure 2 Reward function convergence curve (see online version for colours)

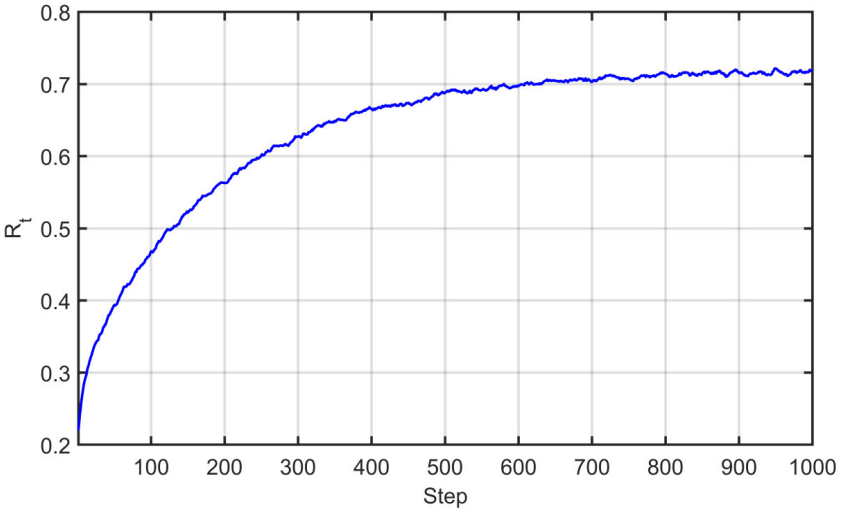


Figure 2 shows the reward function convergence curve. The mean reward value climbs rapidly in the initial phase, stabilises in the mid-term, and then enters a convergence phase with minimal fluctuations, meeting the policy convergence criteria. This curve

demonstrates that the PPO algorithm can effectively optimise the designed multi-objective reward function and quickly learn a high-performance control strategy in a simulation environment. All reward components are calculated in real time by an automated system, eliminating manual intervention. After verifying its gradient differentiability and policy convergence in the PlantSim simulator, the function is deployed in a real-world nursery. The daily reward value provides feedback to the PPO algorithm, driving the policy network to learn an optimal control strategy that balances multiple objectives.

3.4 PPO-based policy network training

To achieve stable and efficient regulation of the cultivation environment for tree seedlings, this paper uses a proximal policy optimisation algorithm to train the policy network (Hu et al., 2023; Alagha et al., 2022; Son et al., 2023). This algorithm ensures the stability of policy updates while applying to high-dimensional state and discrete-action spaces, meeting the control requirements of this study.

The policy network uses an actor-critic architecture, consisting of two independent, three-layer, fully connected multilayer perceptrons. The actor network takes as input a normalised state vector $\mathbf{s}_t \in \mathrm{R}^{12}$, which, after activation through three layers of reluctant units (256-256-256 units), outputs a 243-dimensional action probability distribution $\pi_\theta(a_t|\mathbf{s}_t)$. The critic network has the same structure and outputs a state value estimate $V_\phi(\mathbf{s}_t)$. Network parameters θ and ϕ are updated using policy gradients and temporal difference errors, respectively.

The training adopts the PPO-Clip objective function to limit the strategy update step size and prevent performance collapse:

$$L^{CLIP}(\theta) = E_t\left[\min\left(r_t(\theta)\hat{A}_t,\, clip\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right] \tag{10}$$

In formula (10), $r_t(\theta) = \dfrac{\pi_\theta(a_t|\mathbf{s}_t)}{\pi_{\theta_{old}}(a_t|\mathbf{s}_t)}$ is the action probability ratio under the new and old strategies, and $\hat{A}_t$ is the generalised advantage estimation (GAE), which is defined as:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\mho\lambda)^l \left(r_{t+l} + \mho V_\phi(\mathbf{s}_{t+l+1}) - V_\phi(\mathbf{s}_{t+l})\right) \tag{11}$$

In formula (11), r_{t+l} is the actual reward at time $t + l$; $\mho = 0.99$ is the discount factor; $\lambda = 0.95$ is the GAE smoothing parameter; T is the trajectory length. This advantage estimation effectively reduces the variance of the policy gradient and improves sample efficiency.

The training process consists of two phases. The first phase involves pre-training in the proprietary PlantSim simulator. This simulator is based on five years of historical seedling data (covering three tree species, 12 environmental combinations, and over 2,000 seedling growth trajectories). It uses a long short-term memory (LSTM) dynamic model to simulate state transitions, such as seedling height and stress response. The fidelity of the state transition function $\mathbf{s}_{t+1} = f_{sim}(\mathbf{s}_t, a_t)$ is verified using mean squared error. The LSTM dynamic model employs a two-layer architecture with 128 hidden units per layer and a dropout rate of 0.2 between layers to prevent overfitting. The model takes

the current 12-dimensional state vector $\mathbf{s}_t$ and the encoded discrete action a_t as input at each timestep, and outputs a prediction for the subsequent state vector $\mathbf{s}_{t+1}$, focusing on the dynamic environmental components and the derived seedling height increment. The model was trained on the historical dataset using a mean squared error loss function and the Adam optimiser. The validation error was evaluated on a held-out 20% of the historical training dataset; that portion was not included in the modelling or training process. The overall result was a normalised MSE of 0.018 for seedling height and an average error rate of less than 5% with respect to the supporting environmental variables, compared with subsequent measurements. The error range indicates the extremely high level of accuracy required for pre-training in support of use in RL-based applications. In this environment, 10^5 interaction trajectories are collected, and the actor-critic network is trained until policy convergence. To establish a common approach to policy convergence, a quantitative metric was used to assess the effectiveness of the policies. The assessment process concluded when the moving average of the total reward for training episodes from the last one hundred episodes reached a stable state, meaning that the average reward for those training episodes did not vary by more than 0.5% from its mean value during the last three evaluation checkpoints, which occurred every 5,000 steps during the training process. Determining convergence enables stable policy performance before transitioning to the actual fine-tuning of policies for application in a real-world context.

In the second phase, the pre-trained model is fine-tuned on a real-world smart nursery. To mitigate the simulation-reality gap, an online fine-tuning strategy is employed: a PPO update is performed every 50 decision-step interactions, with a batch size of 64, a learning rate of $\eta = 3 \times 10^{-4}$, and an Adam optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$) An experience replay buffer (capacity = 2,048 steps) is also introduced to break data correlation and improve training stability.

This two-stage design is necessary for three reasons: safety avoids high-risk exploration on real seedlings (such as extreme temperature and humidity combinations); efficiency leverages the simulator's unlimited sampling capabilities to accelerate policy convergence and shorten the actual deployment cycle; generalisation is achieved because the pre-trained policy has already learned the general environment-growth mapping law, and fine-tuning only needs to adapt to specific cabin deviations and tree species characteristics, achieving stable performance in just seven days.

During training, the policy network receives state input $\mathbf{s}_t$ and outputs actions every 10 minutes $a_t \sim \pi_\theta(\cdot|\mathbf{s}_t)$. The actuator mapping table performs control. The system synchronously records the reward R_t and the next state $\mathbf{s}_{t+1}$, forming a transfer tuple ($\mathbf{s}_t$, at, R_t, $\mathbf{s}_{t+1}$) and storing it in the playback buffer.

3.5 Category embedding and transfer learning mechanism

To address physiological differences in the responses of different tree species to the environment during forest seedling cultivation and to reduce the data and time costs of developing new tree species strategies, this paper introduces a learnable species embedding mechanism into the PPO strategy network. It combines it with parameter-efficient fine-tuning (PEFT) to achieve cross-species strategy migration (Wang et al., 2025b).

The specific implementation steps are as follows: a 5-dimensional trainable embedding vector $\mathbf{e}_s \in R^5$ is assigned to each target tree species. Assuming that the tree

species set s = {*Pinus massoniana Lamb.*, *Cunninghamia lanceolata (Lamb.) Hook.*, *Larix gmelinii (Rupr.) Kuzen.*, *Pinus tabuliformis Carrière*, then each row of the embedding matrix corresponds to the embedding vector of a tree species. During the training phase, this matrix is used as part of the network parameters and is jointly optimised with the policy network through backpropagation.

The state input structure is modified. The species part of the original state vector $\mathbf{e}_s \in \mathrm{R}^5$ (including the one-hot encoded species identifier), is replaced with the corresponding embedding vector, that is, if the current tree species belongs to the tree species set, its embedding $\mathbf{e}_s$ is concatenated with the remaining 11-dimensional non-species features to form an enhanced state vector:

$$\mathbf{s}_t = \left[h_0, d_0, SPAD_0, r_{root}, T_a, RH, PAR, C_{CO_2}, \theta_v, \mathbf{e}_s^\top\right]^\top \in \mathrm{R}^{16} \tag{12}$$

This design enables the policy network to implicitly learn species-specific physiological parameters (e.g., optimal temperature range and water sensitivity) via embedding vectors, rather than relying on hard-coded rules. The original state vector formulation in equation (1) used a 3-dimensional one-hot encoding for three source species. In the enhanced design, this 3-dimensional one-hot vector is replaced by a 5-dimensional trainable embedding vector. Consequently, the enhanced state vector's dimensionality becomes: 4 initial traits + 7 environmental variables + 5 species embedding = 16 dimensions. The increase in dimensionality from 12 to 16 reflects a richer, learnable representation of species.

A two-stage transfer training process is implemented. In the first stage, the main policy network is trained on the source tree species until convergence. At this point, the backbone parameters of the actor and critic networks are frozen, leaving only the embedding and output layers trainable. In the second stage, the target tree species are fine-tuned: the embedding vector of the target tree species is initialised to the weighted average of the source tree species embedding, and a small amount of interaction data is collected in the real seedling cabin. The average calculations were based on how closely the known optimal ranges for specific targeted environmental variables (i.e., temperature, humidity, and light) of the target species match those of each possible source species. Each potential source species was assigned a 'similarity score' by calculating the inverse of the Euclidean distance between the normalised optimal range for that species and the documented optimal range for the target species. The similarity scores were then converted to weights by summing all scores to obtain a single value of 1 (normalisation). Using this method allowed the use of existing ecological knowledge and the physiological tolerances of respective species, rather than relying solely on the arithmetic mean or phylogenetics, in determining initial weights. Only the embedding layer and output layer parameters are updated, and the backbone network remains frozen. This parameter-efficient transfer learning strategy reduces the number of trainable parameters and helps prevent overfitting. Finally, the policy network dynamically selects the embedding vector based on the input seedling species index during inference, thereby enabling unified regulation across 'one network, multiple tree species'. This design significantly improves the generalisation and deployment efficiency of the strategy without increasing the model's complexity, providing a scalable technical path for intelligent seedling cultivation of multiple tree species.

4 Experimental results

4.1 Experimental setup

To verify the effectiveness and robustness of the proposed reinforcement learning control strategy in forest seedling cultivation, this study conducts a controlled experiment under a controlled environment. Four typical afforestation tree species are selected for the experiment: Masson pine (*Pinus massoniana* Lamb.), Chinese fir (*Cunninghamia lanceolata* (Lamb.) Hook.), larch (*Larix gmelinii* (Rupr.) Kuzen.), and Chinese pine (*Pinus tabuliformis* Carrière). The experimental materials are distributed using a completely randomised block design. The experimental configuration is shown in Table 1.

Table 1 Experimental configuration table

Configuration item	*Details*
Growth racks	Light-emitting diode (LED) lights, ultrasonic humidifiers, heating/cooling modules, CO_2 release units, drip irrigation
Growing substrate	Peat:perlite = 3:1 (v/v); pH 5.8 ± 0.2; initial volumetric water content: 35% ± 2%
Temperature and humidity	SHT35 sensor; sampled every 10 min
Photosynthetically active radiation	LI-190R quantum sensor
CO_2 concentration	Senseair S8 sensor; accuracy: ±30 ppm
Soil volumetric water content	TDR-300 time-domain reflectometer; accuracy: ±2%
Data transmission	LoRa wireless module → Raspberry Pi 4B + Ubuntu Core
Height measurement	LDM71
Multispectral imaging	MicaSense Altum camera
Thermal imaging	FLIR A655sc thermal camera; resolution 640 × 480; accuracy ±1°C; used for leaf temperature monitoring
Edge computing	Raspberry Pi 4B
Storage	Time-series database
Actuator communication	Advantech UNO-2484G

The experiment is divided into two groups: an experimental group and a control group. The experimental group deploys the PPO policy controller described in this paper, generating control actions every 10 minutes based on real-time state input and driving the actuators via Modbus transmission control protocol (TCP). The control group adopts a threshold control strategy commonly used in forestry. Aside from the control strategy, all other management measures remain the same in both groups. The experimental period lasts 90 days, covering the main seedling-raising period in spring.

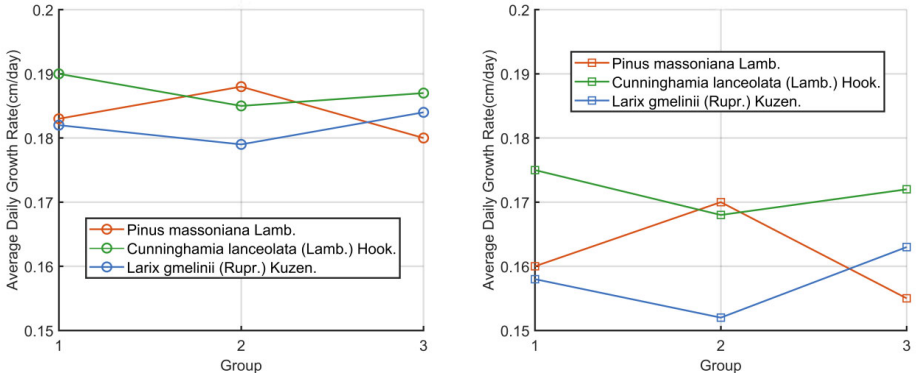
4.2 Growth performance evaluation

This study quantitatively evaluates the average daily growth rate of seedlings, a core indicator of the effectiveness of environmental manipulation strategies in promoting

biomass accumulation. High-precision automated measurement and standardised calculation processes are used. The measurement equipment used is an LDM71 laser ranging sensor, mounted on a track above the nursery chamber and automatically operating at 8:00 AM daily. To ensure consistent measurement, the system performs an automatic calibration before each day's measurement. All height measurement data is verified in real time by an edge computing gateway.

The average daily growth rate is calculated on an individual basis, assuming linear growth, meaning that the total increment is evenly distributed over 90 days. In this experiment, each tree species is divided into three groups within the experimental and control groups. A comparison of the average daily growth rates of forest seedlings of different tree species is shown in Figure 3.

Figure 3 Comparison of average daily growth rates of tree seedlings of different tree species, (a) average daily growth rate of the experimental group (b) average daily growth rate of the control group (see online version for colours)



(a) (b)

Figure 3 shows the average daily growth rates of different tree species. As shown in Figure 3(a), the average daily growth rates of the three tree species in the experimental group are 0.184 cm/day (*Pinus massoniana*), 0.187 cm/day (*Cunninghamia lanceolata*), and 0.182 cm/day (larch). The corresponding values for the control group in Figure 3(b) are lower than those for the experimental group: 0.162 cm/day, 0.172 cm/day, and 0.158 cm/day, respectively. The improvements across the experimental group are relatively balanced, demonstrating the universal efficiency-enhancing capabilities of the reinforcement learning strategy across tree species with diverse physiological characteristics. The observed variation in growth rate improvement among species, however, is attributed to their inherent physiological characteristics and the composition of the pre-training data. Masson pine, with its broader documented tolerance to environmental fluctuations in the historical dataset, enabled the policy to pursue more aggressive growth-promoting actions, resulting in a moderate but consistent gain. The Chinese fir's control group's higher baseline growth demonstrates that this species has a propensity to perform well when environmental conditions are optimal, and that the reinforcement learning approach has helped to further optimise the statement of the Chinese fir's baseline growth through the implementation of RL-based reinforcement learning, resulting in a 17% improvement relative to baseline conditions.

In contrast, the larch's optimisation landscape is much more restricted because of its sensitivity to water stress and the temperature ranges experienced during early growth stages; therefore, larch's survival and deferring stress were prioritised in larch's use in the RL strategy, resulting in slower growth rate improvements for larch, although the improvements were consistent and stable. Overall, this validated the differential species-based response achieved through an embedding mechanism, which allowed the RL strategy to build on each species' underlying physiological characteristics rather than implementing a uniform dynamic RL strategy. The growth advantage reflected in Figure 3 is essentially the result of the policy network dynamically generating discrete actions based on the state space.

The final survival rate of seedlings is a key indicator for evaluating the stress tolerance and long-term adaptability of forest seedlings to environmental manipulation strategies. To ensure objectivity and reproducibility of statistical results, a dual-judgement mechanism of 'automatic initial screening + manual review' is used to determine the survival status of each seedling at the end of the experiment. Automatic survival determination is based on multispectral image analysis. At 12:00 noon daily, a multispectral camera captures full coverage of each nursery rack in each chamber at a fixed height. The system calculates the normalised vegetation index in real time to determine seedling mortality. To eliminate misjudgements in images, all automatic determinations are manually reviewed. Two forestry professionals independently observe the aboveground parts of the seedlings to confirm their survival status. Only seedlings that are unanimously determined to be dead are counted in the mortality statistics. In case of disagreement, a third senior technician arbitrates the results. The review process is blinded; the observers are unaware of the seedling's treatment group to avoid subjective bias. A comparison of the final survival rates of different tree species is shown in Figure 4.

Figure 4 Comparison of the final survival rates of different tree species, (a) final survival rate of the experimental group (b) final survival rate of the control group (see online version for colours)

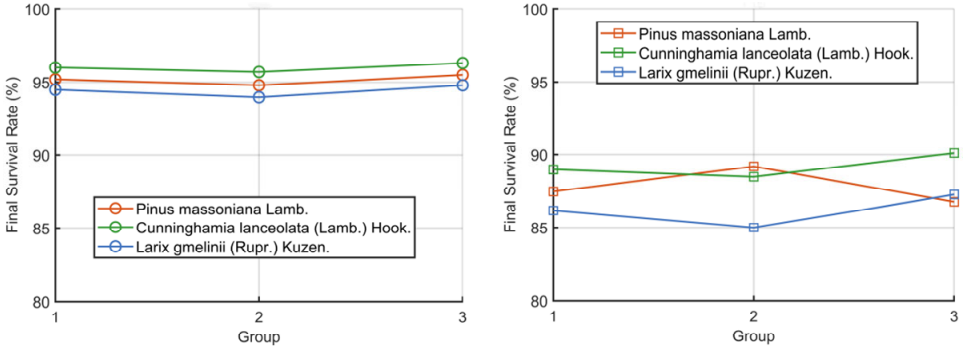


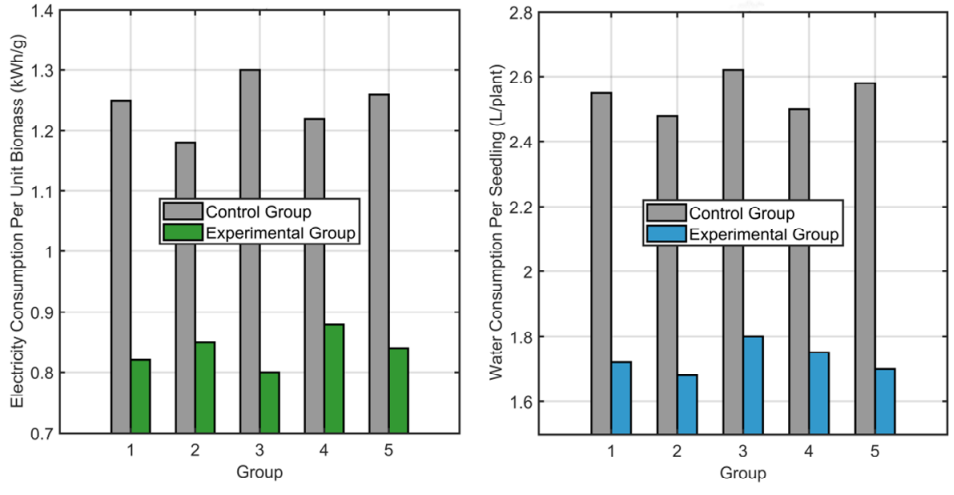
Figure 4 shows the final survival rates of different tree species. Figure 4(a) shows that the final survival rates of the three tree species in the experimental group are 95.2% (*Pinus massoniana*), 96% (*Cunninghamia lanceolata*), and 94.4% (larch). In contrast, the corresponding values for the control group [Figure 4(b)] are lower at 87.8%, 89.2%, and

86.2%, respectively. The mean survival rate for all samples in the experimental group reaches 95.2%, while that in the control group is 87.7%, representing a 7.5% improvement. This improvement in survival rate stems from the emphasis on 'survival guarantee' in the multi-objective reward function and the policy network's ability to dynamically respond to individual states.

4.3 Resource utilisation and stress

To quantify the effectiveness of the reinforcement learning control strategy in optimising resource utilisation, environmental resource consumption efficiency is evaluated from two core dimensions: electricity and irrigation water. Five groups are assigned to each experimental and control group. Electricity consumption per unit biomass and water consumption per unit seedling are calculated. All data are automatically collected using high-precision sensing equipment. A comparison of environmental resource consumption efficiency is shown in Figure 5.

Figure 5 Comparison of environmental resource consumption efficiency, (a) comparison of power consumption per unit biomass (b) comparison of water consumption per unit of seedling (see online version for colours)



(a) (b)

Figure 5 shows the efficiency of environmental resource utilisation. As shown in Figure 5(a), the power consumption per unit biomass of the experimental group is lower than that of the control group. The values of the five groups are 0.82 kWh/g, 0.85 kWh/g, 0.8 kWh/g, 0.88 kWh/g, and 0.84 kWh/g, respectively, while those of the control group are 1.25 kWh/g, 1.18 kWh/g, 1.3 kWh/g, 1.22 kWh/g, and 1.26 kWh/g, with an average reduction of about 0.4 kWh/g in power consumption per unit biomass. In Figure 5(b), the water consumption per unit seedling of the experimental group is 1.72 L/plant, 1.68 L/plant, 1.8 L/plant, 1.75 L/plant, and 1.7 L/plant, respectively. In comparison, those of the control group are 2.55 L/plant, 2.48 L/plant, 2.62 L/plant, 2.5 L/plant, and 2.58 L/plant, reducing water consumption by an average of approximately 0.82 L/plant. Overall, the experimental group consumes fewer resources. The standard deviations of

electricity consumption per unit biomass and water consumption per unit seedling are shown in Table 2.

Table 2 Standard deviation of electricity consumption per unit biomass and water consumption per unit seedling

Metric	*Group 1*	*Group 2*	*Group 3*	*Group 4*	*Group 5*
Experimental group – electricity consumption per unit biomass (kWh/g)	0.03	0.02	0.02	0.03	0.02
Control group – electricity consumption per unit biomass (kWh/g)	0.04	0.05	0.03	0.04	0.03
Experimental group – water consumption per seedling (L/plant)	0.05	0.04	0.04	0.03	0.04
Control group – water consumption per seedling (L/plant)	0.06	0.07	0.05	0.06	0.05

Table 2 shows that the standard deviations of electricity consumption per unit biomass in the experimental groups are 0.03 kWh/g, 0.02 kWh/g, 0.02 kWh/g, 0.03 kWh/g, and 0.02 kWh/g, respectively, and in the control groups, 0.04 kWh/g, 0.05 kWh/g, 0.03 kWh/g, 0.04 kWh/g, and 0.03 kWh/g, respectively. The standard deviations of water consumption per unit seedling in the experimental and control groups are 0.05/0.06 L/plant, 0.04/0.07 L/plant, 0.04/0.05 L/plant, 0.03/0.06 L/plant, and 0.04/0.05 L/plant, respectively. The standard deviations of all groups in the experimental group are lower than those in the control group, indicating less fluctuation in the experimental group. The reduced variance in resource consumption is a direct consequence of the reinforcement learning strategy's stability and adaptability. The threshold-based control in the control group reacts uniformly to fixed environmental setpoints, leading to periodic, high-magnitude actuator adjustments that cause oscillatory consumption patterns.

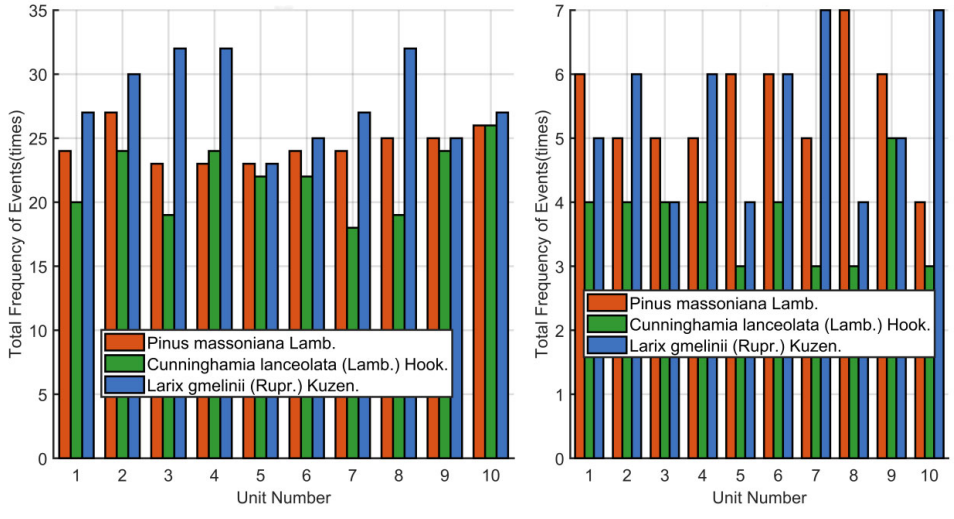
In contrast, the PPO policy generates smoother control actions by continuously optimising for the multi-objective reward. It anticipates state transitions and avoids abrupt, energy-intensive corrections. The policy uses an integrated representation of species attributes and real-time responses from each species to deliver resources incrementally and precisely at the appropriate levels to match seedling demand, thereby minimising waste associated with oversupply and undersupply. Because of this, there is reduced overall mean resource consumption and much lower variance when using this data-based technique or approach to control seedling resource delivery and create sustainable planting environments.

To evaluate the ability of different environmental control strategies to mitigate physiological stress in tree seedlings, the experimental and control groups are each divided into 10 units. Figure 6 compares the frequency of stress events across different treatment units.

Figure 6 shows the frequency of stress events in different treatment units. Figure 6(a) shows that the average frequency of stress events for the three tree species in the control group is 24.4, 21.8, and 28, respectively, with an average of 24.73 for all samples. In contrast, Figure 6(b) shows significantly lower frequencies in the experimental group at 5.5, 3.7, and 5.4, respectively, with an average of 4.87 across all samples, a decrease of 19.86 events compared to the control group. This significant reduction in stress frequency

stems from the explicit penalty mechanism for stress signals in the multi-objective reward function and the policy network's ability to perceive physiological states in real time.

Figure 6 Comparison of the frequency of stress events in different treatment units, (a) frequency of stress events in different treatment units of the control group (b) frequency of stress events in different treatment units of the experimental group (see online version for colours)



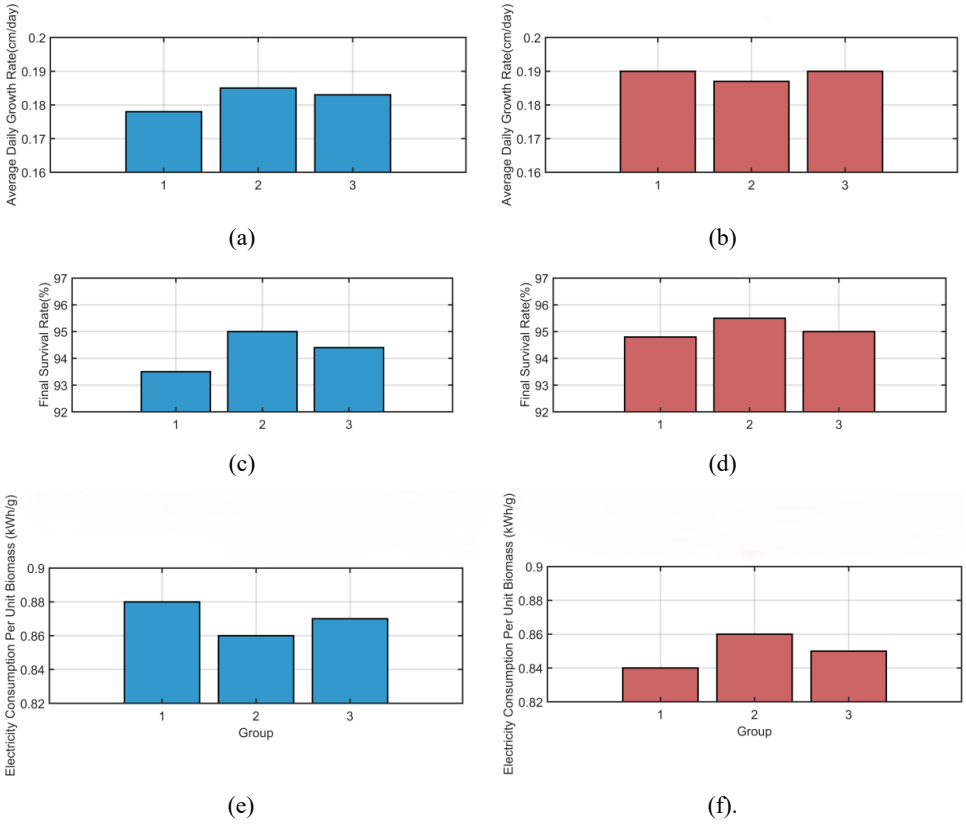
4.4 Strategy generalisation verification

To validate the proposed reinforcement learning strategy's adaptability to unseen tree species, this study uses *Pinus tabuliformis* Carrière as a test case to assess its cross-species generalisation. *Pinus tabuliformis* is a zero-shot transfer task that serves as a test case for the effectiveness of the species-embedding mechanism.

A transfer unit is established: healthy *Pinus tabulaeformis* seedlings are randomly divided into three groups and trained for 7 days of fine-tuning. All seedlings are treated with the same substrate and management as the experimental group. As a control, a retraining unit is established: three groups are randomly divided into groups and trained from scratch using the PPO strategy. Aside from the strategy source, all other conditions are identical between the two groups. The cross-species generalisation ability test for *Pinus tabulaeformis* is shown in Figure 7.

Figure 7 shows three core performance metrics. Figures 7(a) and 7(b) show that the average daily growth rates of the migration and retraining units are 0.182 cm/day and 0.189 cm/day, respectively, with minimal difference. Figures 7(c) and 7(d) show that the average final survival rates are 94.3% and 95.1%, respectively. Figures 7(e) and 7(f) show that the average power consumption per unit biomass is 0.87 kWh/g and 0.85 kWh/g, respectively. The performance of these three metrics is relatively similar, indicating that the migration strategy's performance closely approaches the baseline level achieved after 90 days of training from scratch, achieving near-optimal control with only 7 days of fine-tuning.

Figure 7 Testing of the cross-species generalisation ability of *Pinus tabulaeformis*, (a) average daily growth rate of migration units (b) average daily growth rate of retraining units (c) final survival rate of migration units (d) final survival rate of retraining units (e) energy consumption per unit biomass of migration units (f) energy consumption per unit biomass of retraining units (see online version for colours)

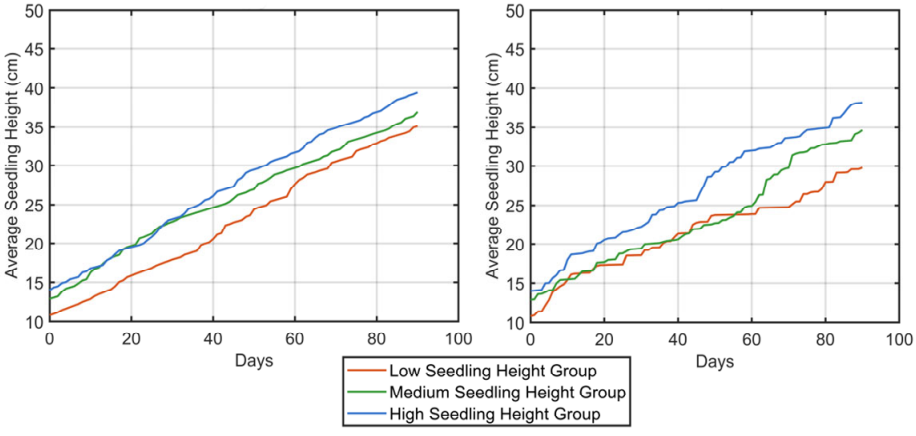


4.5 Personalised regulation adaptability

To verify the proposed reinforcement learning strategy's ability to adaptively regulate individual seedling heterogeneity, this study groups experimental seedlings by initial height. It analyses the differences in the strategy's regulatory effects under different initial growth states. This analysis aims to test whether the strategy can dynamically compensate for the growth disadvantages of vulnerable individuals, demonstrating its personalised regulatory advantages.

Experimental data are obtained from forest seedlings. Based on seedling height measurements taken with a laser altimeter on the first day of the experiment, seedlings within each species are divided into three groups using the tertile method: low seedling height group (<11.5 cm), medium seedling height group (11.5–13.2 cm), and high seedling height group (>13.2 cm). For each seedling group, seedling height is recorded daily, and the average daily growth rate within the group is calculated. A 90-day growth trajectory curve is plotted. The 90-day growth trajectories for different initial seedling height groups are shown in Figure 8.

Figure 8 Comparison of 90-day growth trajectories of different initial seedling height groups, (a) growth trajectory of the experimental group (b) growth trajectory of the control group (see online version for colours)



(a) (b)

Figure 8 shows the growth trajectories of the groups with different initial seedling heights. As shown in Figure 8(a), in the experimental group, the low seedling height group starts at approximately 10.8 cm and reaches 35.1 cm after 90 days; the medium seedling height group increases from 12.9 cm to 37 cm; the high seedling height group increases from 14 cm to 39.4 cm. The final differences among the three groups are relatively small. In the control group of Figure 8(b), the three groups have similar starting seedling heights. Still, the final differences are greater than those in the experimental group: the low group reaches 29.8 cm, the medium group reaches 34.7 cm, and the high group reaches 38.2 cm. This regulatory effect in the experimental group stems from the reinforcement learning strategy's dynamic response to the individual's initial state and from the coordinated guidance of the multi-objective reward function. The model explicitly includes initial traits such as seedling height and ground diameter in the state input, and identifies the physiological characteristics of tree species through species embedding, so that the strategy can actively increase nighttime temperature or morning light intensity for vulnerable individuals to stimulate their photosynthetic and metabolic activity; for already strong tall seedling individuals, the strategy moderately reduces resource input to avoid the risk of high leaf temperature or water stress, sacrificing a small amount of growth in exchange for higher survival rate and lower energy consumption.

An analysis of action frequency distributions extracted from the policy's execution logs substantiates these differential behaviours. For the low initial height group, the policy selected '+2°C temperature increase' actions during night periods 23% more frequently compared to the high height group. Morning '+200 μmol/m^2/s light increase' actions occurred 31% more often for the low group. Conversely, for the high initial height group, the policy exhibited a 40% higher frequency of '0' actions (maintenance) for irrigation and a 19% higher frequency of '0' actions for CO_2 supplementation during midday periods, indicating restrained resource input. These statistically distinct action patterns demonstrate the policy's capacity to decode the initial height state into concrete, differentiated control sequences that actively compensate for growth disadvantages or

mitigate risks for advantaged individuals. The differences in quantified action frequency were directly correlated with the reduction in final height difference observed. For the low starting height group, there were higher frequencies of nighttime +2°C heating actions at a 23% increased frequency than the high starting height group, as well as a 31% increase in morning +200 μmol/m²/s light boost actions. The allocation of actions increased the rate of photosynthesis/metabolism during key times, thereby accelerating growth during the early and mid phases observed in Figure 8(a) for the low starting height group. Significantly higher rates of both irrigation maintenance and midday CO_2 supply were observed across the high initial height group compared with the low initial height group, with increases of 40% and 19%, respectively. These types of restrained inputs further reduced the likelihood of excessive waterlogging and luxury consumption, both of which can affect key aspects of growth, as illustrated in Figure 8(b). As a direct result of the differences in action patterns noted, the height differential between the low initial height group and the high initial height group of 8.4 cm in the control treatment and 4.3 cm in the experimental treatment (after personalised control) helps to explain approximately 4.1 cm of the 4.1 cm difference in final height called for by the policy.

5 Conclusions

This paper addresses the dynamic regulation of environmental variables in the intelligent cultivation of forest tree seedlings. It proposes a reinforcement learning method that integrates species embedding and proximal policy optimisation. This method constructs a 12-dimensional state space containing the initial characteristics of the seedlings, multidimensional environmental parameters, and tree species encoding, designs a 243-dimensional discrete action space and a multi-objective reward function, and introduces species embedding and efficient parameter transfer mechanisms to achieve precise adaptive regulation of different tree species and individuals. Experiments show that this strategy improves the average daily growth rate, ultimately achieving a survival rate of 95.2%, while reducing energy consumption per unit biomass and the frequency of stress events. Transfer learning enables new tree species to achieve performance levels comparable to 90 days of training from scratch after only 7 days of fine-tuning, verifying the method's efficiency, robustness, and generalisation. The method's robustness is ensured by edge-side noise filtering and imputation for brief data gaps. The learned policy shows resilience to minor state variations, with prolonged sensor failure triggering predefined safety thresholds. Future work will formally test robustness via noise injection.

Declarations

The author declares that he has no conflicts of interest regarding the publication of this article.

References

Ajagekar, A., Mattson, N.S. and You, F. (2023) 'Energy-efficient AI-based control of semi-closed greenhouses leveraging robust optimization in deep reinforcement learning', *Advances in Applied Energy*, Vol. 9, No. 1, p.100119.

Akbar, J.U.M., Kamarulzaman, S.F., Muzahid, A.J.M. et al. (2024) 'A comprehensive review on deep learning assisted computer vision techniques for smart greenhouse agriculture', *IEEE Access*, Vol. 12, No. 1, pp.4485–4522.

Alagha, A., Singh, S., Mizouni, R. et al. (2022) 'Target localization using multi-agent deep reinforcement learning with proximal policy optimization', *Future Generation Computer Systems*, Vol. 136, No. 1, pp.342–357.

Alotaibi, K.D., Alharbi, H.A., Yaish, M.W. et al. (2023) 'Date palm cultivation: a review of soil and environmental conditions and future challenges', *Land Degradation & Development*, Vol. 34, No. 9, pp.2431–2444.

Augusto, L., Borelle, R., Boca, A. et al. (2025) 'Widespread slow growth of acquisitive tree species', *Nature*, Vol. 640, No. 8058, pp.395–401.

Binlajdam, R., Meedeniya, D., Kosala, C. et al. (2025) 'Review on sustainable forestry with artificial intelligence', *ACM Journal on Computing and Sustainable Societies*, Vol. 3, No. 4, pp.1–48.

Caliskan, S., Kartal, E., Balekoglu, S. et al. (2024) 'Using machine learning algorithms to cluster and classify stone pine (*Pinus pinea* L.) populations based on seed and seedling characteristics', *European Journal of Forest Research*, Vol. 143, No. 5, pp.1575–1591.

Castro, J., Alcaraz-Segura, D., Baltzer, J.L. et al. (2024) 'Automated precise seeding with drones and artificial intelligence: a workflow', *Restoration Ecology*, Vol. 32, No. 5, p.e14164.

Ecke, S., Stehr, F., Frey, J. et al. (2024) 'Towards operational UAV-based forest health monitoring: species-specific damage detection', *Computers and Electronics in Agriculture*, Vol. 219, p.108785.

Ehrlich-Sommer, F., Hoenigsberger, F., Gollob, C. et al. (2024) 'Sensors for digital transformation in smart forestry', *Sensors*, Vol. 24, No. 3, p.798.

Fan, X., Zhang, H., Zhou, L. et al. (2024) 'Evaluating drought stress response of poplar seedlings using a proximal sensing platform via multi-parameter phenotyping and two-stage machine learning', *Computers and Electronics in Agriculture*, Vol. 225, p.109261, DOI: 10.1016/j.compag.2024.109261.

Han, S., Miao, X., Zhang, Y. et al. (2025) 'Effects of mixed addition of *Fraxinus mandshurica* Rupr. and *Larix gmelinii* (Rupr.) Kuzen. litter on nitrogen mineralization in dark brown soil of Northeast China', *Forests*, Vol. 16, No. 5, p.842.

Hoppen, M., Chen, J., Kemmerer, J. et al. (2024) 'Smart forestry – a Forestry 4.0 approach to intelligent and fully integrated timber harvesting', *International Journal of Forest Engineering*, Vol. 35, No. 2, pp.137–152.

Hu, B., Zhang, K., Li, N. et al. (2023) 'Toward a theoretical foundation of policy optimization for learning control policies', *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 6, No. 1, pp.123–158.

Huang, X., Guo, W., Yang, L. et al. (2023) 'Effects of drought stress on non-structural carbohydrates in different organs of Cunninghamia lanceolata', *Plants*, Vol. 12, No. 13, p.2477.

Hui, T., Ming, W., Qiushi, Y.U. et al. (2023) 'Root image segmentation method based on improved UNet and transfer learning', *Smart Agriculture*, Vol. 5, No. 3, p.96.

Icarte, R.T., Klassen, T.Q., Valenzano, R. et al. (2022) 'Reward machines: exploiting reward function structure in reinforcement learning', *Journal of Artificial Intelligence Research*, Vol. 73, No. 1, pp.173–208.

Iwasaki, M., Penfield, S. and Lopez-Molina, L. (2022) 'Parental and environmental control of seed dormancy in *Arabidopsis thaliana*', *Annual Review of Plant Biology*, Vol. 73, No. 1, pp.355–378.

Javaid, M.M., Mahmood, A., Alshaya, D.S. et al. (2022) 'Influence of environmental factors on seed germination and seedling characteristics of perennial ryegrass (*Lolium perenne* L.)', *Scientific Reports*, Vol. 12, No. 1, p.9522.

Jin, L., Song, X., Shi, Y. et al. (2024) 'Photosynthetic acclimation of larch to the coupled effects of light intensity and water deficit in regions with changing water availability', *Plants*, Vol. 13, No. 14, p.1891.

Lee, M.H., Yao, M.H., Kow, P.Y. et al. (2024) 'An artificial intelligence-powered environmental control system for resilient and efficient greenhouse farming', *Sustainability*, Vol. 16, No. 24, p.10958.

Li, J., Sa, R.L., Zhang, M.T. et al. (2024a) 'Analysis of the dominant factors that influence growth of *Larix gmelinii* forests', *Applied Ecology and Environmental Research*, Vol. 22, No. 4, pp.3453–3473.

Li, S., Yang, L., Huang, X. et al. (2023a) 'Mineral nutrient uptake, accumulation, and distribution in *Cunninghamia lanceolata* in response to drought stress', *Plants*, Vol. 12, No. 11, p.2140.

Li, Y., Huang, Y., Wang, M. et al. (2023b) 'An improved U-Net-based in situ root system phenotype segmentation method for plants', *Frontiers in Plant Science*, Vol. 14, p.1115713, DOI: 10.3389/fpls.2023.1115713.

Li, Y., Xia, H., Liu, Y. et al. (2024b) 'Research on the detection method of phenotypic information of *Pinus massoniana* Lamb. seedling root system', *Signal, Image and Video Processing*, Vol. 18, No. 10, pp.6961–6972.

Liu, Z., Zhong, Y., Chen, Z. et al. (2024) 'Landscape dynamics, succession, and forecasts of *Cunninghamia lanceolata* in the central producing regions of China', *Forests*, Vol. 15, No. 10, p.1817.

Mahmood, T., Asif, M. and Raza, Z.H. (2024) 'Smart forestry: the role of AI and bioengineering in revolutionizing timber production and biodiversity protection', *Revista de Inteligencia Artificial en Medicina*, Vol. 15, No. 1, pp.1176–1202.

Maraveas, C., Karavas, C.S., Loukatos, D. et al. (2023) 'Agricultural greenhouses: resource management technologies and perspectives for zero greenhouse gas emissions', *Agriculture*, Vol. 13, No. 7, p.1464.

Platero-Horcajadas, M., Pardo-Pina, S., Camara-Zapata, J.M. et al. (2024) 'Enhancing greenhouse efficiency: integrating IoT and reinforcement learning for optimized climate control', *Sensors*, Vol. 24, No. 24, p.8109.

Shi, D., Yuan, P., Liang, L. et al. (2024) 'Integration of deep learning and sparrow search algorithms to optimize greenhouse microclimate prediction for seedling environment suitability', *Agronomy*, Vol. 14, No. 2, p.254.

Socha, J., Orzl, S., Ochal, W. et al. (2022) 'Effect of seedling production method on the growth of *Pinus sylvestris* L. on reclaimed post-industrial sites in Poland', *Dendrobiology*, Vol. 88, No. 1, pp.124–137.

Son, S., Zheng, L., Sullivan, R. et al. (2023) 'Gradient-informed proximal policy optimization', *Advances in Neural Information Processing Systems*, Vol. 36, No. 1, pp.8788–8814.

Sun, W. and Chang, F.J. (2023) 'Empowering greenhouse cultivation: dynamic factors and machine learning unite for advanced microclimate prediction', *Water*, Vol. 15, No. 20, p.3548.

Wang, G.G., Lu, D., Gao, T. et al. (2024a) 'Climate-smart forestry: an AI-enabled sustainable forest management solution for climate change adaptation and mitigation', *Journal of Forestry Research*, Vol. 36, No. 1, p.7.

Wang, H., Wu, F., Li, M. et al. (2022a) 'Structure and chlorophyll fluorescence of heteroblastic foliage affect first-year growth in *Pinus massoniana* Lamb. seedlings', *Plant Physiology and Biochemistry*, Vol. 170, No. 1, pp.206–217.

Wang, L., Chen, S., Jiang, L. et al. (2025a) 'Parameter-efficient fine-tuning in large language models: a survey of methodologies', *Artificial Intelligence Review*, Vol. 58, No. 8, p.227.

Wang, L., Zhang, H., Bian, L. et al. (2024b) 'Poplar seedling varieties and drought stress classification based on multi-source, time-series data and deep learning', *Industrial Crops and Products*, Vol. 218, p.118905, DOI: 10.1016/j.indcrop.2024.118905.

Wang, T., Xu, Q., Zhang, B. et al. (2024c) 'Effects of thinning and understory removal on water use efficiency of Pinus massoniana: evidence from photosynthetic capacity and stable carbon isotope analyses', *Journal of Forestry Research*, Vol. 35, No. 1, p.41.

Wang, T., Zuo, Y., Manda, T. et al. (2025b) 'Harnessing artificial intelligence, machine learning, and deep learning for sustainable forestry management and conservation: transformative potential and future perspectives', *Plants*, Vol. 14, No. 7, p.998.

Wang, Z., Lu, J., Chen, C. et al. (2022b) 'Investigating the multi-objective optimization of quality and efficiency using deep reinforcement learning', *Applied Intelligence*, Vol. 52, No. 11, pp.12873–12887.

Williams, K.R., Schlossman, R., Whitten, D. et al. (2022) 'Trajectory planning with deep reinforcement learning in high-level action spaces', *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 59, No. 3, pp.2513–2529.

Xu, X., Qiu, J., Zhang, W. et al. (2022) 'Soybean seedling root segmentation using an improved U-net network', *Sensors*, Vol. 22, No. 22, p.8904.

Xu, Z. and Jiang, D. (2025) 'AI-powered plant science: transforming forestry monitoring, disease prediction, and climate adaptation', *Plants*, Vol. 14, No. 11, p.1626.

Xuan, S., Wang, J., Yin, J. et al. (2024) 'Multi-agent reinforcement learning for collaborative optimization of stand structure in *Pinus yunnanensis* secondary forests', *Forests*, Vol. 15, No. 7, p.1143.

Zhao, J., Wang, J., Yin, J. et al. (2024) 'Optimization of the stand structure in secondary forests of *Pinus yunnanensis* based on deep reinforcement learning', *Forests*, Vol. 15, No. 12, p.2181.