

Article

Physics Informed Constrained Reinforcement Learning for Carbon-Regulated Cold Chain Inventory with Digital Twins

Ashish Kumar Routray^{1*}, Swagatika Dehury¹, Mahottak Kumar², Sudhir Kumar Sahu¹

¹ Department of Statistics, Ravenshaw University, Cuttack 753003, India

² Department of civil engineering, Synergy Polytechnic, Bhubaneswar 751032, India

* **Corresponding author:** Ashish Kumar Routray, ashish2025@ravenshawuniversity.ac.in

ABSTRACT

Refrigerated warehousing and reefer distribution represent some of the most energy intensive and loss prone links in critical food and pharmaceutical infrastructure, yet operational control remains dominated by static thermal setpoints and uncoupled lot-sizing heuristics. To address this limitation, we investigate PI-CSAC, a physics informed constrained actor critic framework designed to jointly optimize replenishment lot sizes, dynamic chamber temperature setpoints, and carbon allowance trading actions under cap and trade regulation. The formulation couples an Arrhenius constrained neural deterioration surrogate with a constrained Markov decision process solved via online Lagrangian dual relaxation against explicit fill rate and emissions constraints. In a synthetic proof of concept benchmark using literature derived Arrhenius kinetics across three illustrative SKU families evaluated over twenty held-out years, a reduced scale constrained agent (implemented with REINFORCE, a learned value baseline, and Lagrangian dual ascent) was compared against a grid optimized closed form non-instantaneous deterioration EOQ policy and a grid optimized (s,S) heuristic with safety stock. The reduced scale agent did not outperform either analytical baseline, resulting in 7% to 28% higher mean annual costs across families alongside higher emissions variance and no clear performance separation in constrained versus unconstrained ablations, despite delivering modestly higher fill rates. Ultimately, these findings deliver critical negative and mixed empirical evidence establishing that computational tractability does not guarantee immediate field deployability, highlighting that joint thermo-inventory reinforcement learning requires substantially greater network capacity, training scale, and domain calibration before surpassing tuned classical baselines.

Keywords: AI empowered engineering, physics informed neural networks, constrained reinforcement learning, cold chain logistics, non-instantaneous deterioration, carbon cap and trade, synthetic simulation proof of concept

ARTICLE INFO

Received: 13 August 2026

Accepted: 16 September 2026

Available online: 28 September 2026

COPYRIGHT

Copyright © 2026 by author(s).

Journal of AI-Driven Engineering Innovation and Applications is published by Cranford Press Limited.

This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).
<https://creativecommons.org/licenses/by-nc/4.0/>

1. Introduction

1.1. Engineering context and motivation

Cold chain infrastructure is fundamentally a physical system before it becomes a logistics abstraction. A refrigerated distribution hub consists of compressors, evaporator coils, insulated envelopes, dock door air curtains, and diesel or grid powered reefer units whose thermal behaviour follows heat transfer physics and whose product loss follows biochemical reaction kinetics. The decisions engineers make about how much to order, when to order, and at what temperature to hold products are coupled through these physical laws rather than merely through accounting identities. Lowering a chamber setpoint suppresses enzymatic and microbial degradation, extends usable stock life, and permits larger and less frequent replenishments. However, it also increases compressor duty, electricity consumption, and carbon liability under emissions accounting. Raising the setpoint produces the opposite effects. The engineering optimum is therefore a genuine tradeoff

surface whose position changes with ambient conditions, product mix, tariff structures, and allowance prices. The economic stakes are substantial in emerging economies. India loses a significant share of horticultural output between harvest and consumption, with cold chain gaps repeatedly identified as a major proximate cause. Meanwhile, the national Carbon Credit Trading Scheme is beginning to convert refrigerated storage energy intensity from an unpriced externality into a balance sheet item. Facility engineers must therefore simultaneously maintain service levels, reduce spoilage, and respect emissions constraints, objectives that classical inventory theory rarely addresses jointly with thermal control.

1.2. Why classical models and generic AI both fall short

The analytical literature on deteriorating inventory is mature and, within its assumptions, elegant. Since Ghare and Schrader’s exponential decay formulation^[1], researchers have developed closed form replenishment policies for diverse demand, deterioration, and financing structures^[2,3]. Particularly relevant is the non-instantaneous deterioration (NID) framework introduced by Wu et al.^[4] and Ouyang et al.^[5], where commodities remain intact for an interval τ before deterioration begins. Recent studies extend NID models to varying decay and linear demand^[6], fuzzy formulations with imprecise costs^[7], and multiple warehouse systems incorporating carbon cap and trade, backordering, trade credit, and inflation^[8]. Related research examines preservation technology investment^[9,10] and carbon tax versus cap and trade policies^[11,12].

These models achieve tractability by holding physical behaviour constant. The deterioration rate θ is generally exogenous, while the onset window τ is fixed despite its strong temperature dependence. Demand is also typically stationary or prescribed. Consequently, analytical optimality applies to environments with predetermined physical relationships. Deep reinforcement learning offers greater adaptability and has been applied to inventory control^[13], multiple echelon systems^[14], perishables^[15–17], vendor managed supply chains^[18], and pricing and ordering decisions^[19–21]. Reviews^[22,23] nevertheless identify important limitations. Generic DRL presents two critical problems for cold chain deployment. First, purely data driven deterioration models may produce physically inadmissible extrapolations outside observed operating conditions. Second, incorporating service or emissions constraints through reward penalties makes feasibility dependent on arbitrary weights and provides no guarantee against violations. These limitations are unacceptable when service failures create contractual consequences and emissions violations create regulatory consequences.

1.3. Contributions

- a) This paper makes contributions of two distinct kinds, and we explicitly distinguish between them throughout. Contributions (1) and (2) describe the target design, which represents the framework as proposed for eventual development and deployment. Contribution (3) describes what was actually implemented and evaluated in this study. This implementation is a substantially smaller proof of concept that was developed and tested entirely using a synthetic simulation parameterised from the literature, without using any real facility data, telemetry, or sensor information.
- b) A target design for a physics informed deterioration twin for NID commodities. We specify a neural surrogate $\hat{\theta}_\phi(T, t)$ trained against a composite loss containing a data term, an Arrhenius residual term, a monotonicity term, and an onset consistency term that ties the learned pre deterioration window to a temperature dependent law $\tau(T)$ in Section 5.1, though this surrogate was not trained in the present study because the proof of concept uses the analytic Arrhenius law directly in Section 5.6.
- c) A joint thermal replenishment trading control formulation. The action space contains the order quantity, the chamber temperature setpoint, and, in the target design, a carbon allowance transaction, so deterioration physics is not an exogenous nuisance but a lever the controller actively pulls in Sections 4 through 5, and to our knowledge no prior deteriorating inventory study endogenises the setpoint alongside the lot size under an explicit cap and trade budget.
- d) A reduced scale, honestly reported synthetic proof of concept. We build a synthetic simulation

environment with literature informed rather than facility measured deterioration kinetics for three illustrative SKU families, implement two real analytical baselines consisting of a grid optimised closed form NID EOQ policy and a grid optimised (s, S) heuristic, and implement and train a lightweight constrained actor critic that is a genuine, much smaller stand in for the full design above, with every number reported in Sections 6 through 7 being the direct output of code that was run for this paper, including an honest account of what the reduced scale agent does and does not achieve, namely that it did not outperform either analytical baseline on cost in any of the three SKU families, its seed to seed variance is large, and a constrained versus unconstrained ablation did not separate cleanly, so we treat this as useful negative and mixed evidence about the scale of training this class of framework is likely to need, rather than reporting only favourable outcomes.

1.4. Organization

Section 2 reviews the relevant literature streams, Section 3 describes the notation, assumptions, and a stylised reference scenario used purely to ground the notation, which is explicitly not an operating facility, Section 4 develops the general mathematical formulation of the target design, Section 5 presents the full target PI CSAC framework across Sections 5.1 through 5.5 and then, separately and explicitly, the reduced scale implementation actually used in this paper in Section 5.6, Section 6 describes the synthetic simulation study and experimental protocol, Section 7 reports the real, as obtained results, including the negative and mixed findings along with an explicit checklist of what was and was not built, and Section 8 concludes with limitations and the scale up path required for a future validated deployment.

2. Related work

2.1. Deteriorating and non-instantaneously deteriorating inventory

The deteriorating inventory literature is comprehensively surveyed by Bakker et al.^[2] and Janssen et al.^[3]. Wu et al.^[4] introduced the non-instantaneous deterioration (NID) refinement, recognising an induction period during which quality loss is negligible. This window materially affects optimal cycle lengths because ordering more frequently than τ increases fixed costs, whereas ordering less frequently exposes more inventory to deterioration. Ouyang et al.^[5] extended NID modelling to permissible payment delays, while Dye^[24] incorporated preservation technology as a decision variable that lengthens τ . Recent studies provide the analytical baseline for this work. Sahoo et al.^[6] developed an NID model with continuously varying deterioration and linear demand, deriving optimal replenishment conditions and convexity results. Dehury et al.^[7] proposed a pseudo-convex fuzzy EOQ model with time varying demand and parameter imprecision. Routray et al.^[8] extended NID modelling to a three warehouse system incorporating ramp demand, carbon cap and trade, partial backlogging, trade credit, and inflation. Collectively, these studies establish the structural benchmarks for our learned policy and demonstrate carbon regulation effects, while leaving the thermal state generating θ and τ exogenous.

2.2. Carbon regulated inventory engineering

Benjaafar et al.^[25] first embedded emission accounting inside lot sizing models and demonstrated that modest operational adjustments can deliver disproportionate emission reductions. Hovelaque and Bironneau^[26] derived a carbon constrained EOQ with emission dependent demand, while Toptal et al.^[27] compared cap, tax and cap and trade regimes for joint inventory and abatement investment decisions. In the deteriorating item setting, Khanna et al.^[11] and Lok et al.^[12] both find that cap and trade dominates a pure tax on profitability while remaining comparably effective at emission suppression, because the trading option converts abatement into a revenue stream. Gautam et al.^[28] add preservation investment under environmental deliberation. The consistent finding that cap and trade rewards operational flexibility more strongly than a flat tax is exactly the property that a learned, state contingent policy is positioned to exploit, and it motivates our choice of regime.

2.3. Deep reinforcement learning for inventory and process control

Oroojlooyjadid et al.^[13] showed that deep Q networks can learn cooperative beer game policies without coordination protocols. Geevers et al.^[14] extended DRL to multi echelon networks, while Kara and Dogan^[15] and Wang et al.^[17] addressed perishability. De Moor et al.^[16] demonstrated that reward shaping with a mechanistic baseline accelerates convergence in perishable settings, indicating that domain structure and learning are complementary. Mohamadi et al.^[18] applied DRL to vendor managed perishable chains under capacity uncertainty, while Nomura et al.^[29] and Kavooosi et al.^[20] studied joint dynamic pricing and ordering. Hubbs et al.^[30] addressed chemical production scheduling, providing a process engineering counterpart. Boute et al.^[22] summarised DRL design considerations, while Gijsbrechts et al.^[23] showed that DRL advantages over tuned heuristics depend on problem structure, particularly state dimensionality and non-stationarity.

For constrained control, Altman^[31] provides the formal constrained MDP framework, while Achiam et al.^[32] and Ray et al.^[33] develop practical constrained policy optimisation methods and benchmarks. Haarnoja et al.^[34] introduced soft actor critic, whose maximum entropy objective and off policy efficiency suit repeated facility decisions. Fujimoto et al.^[35] and Schulman et al.^[36] provide TD3 and PPO baselines used here.

2.4. Physics informed learning and cold chain digital twins

Raissi et al.^[37] formalised physics informed neural networks, while Karniadakis et al.^[38] established the broader physics informed machine learning framework. Its core principle of penalising violations of governing relations at collocation points transfers naturally to reaction kinetics, where the governing law is an Arrhenius rate expression rather than a partial differential equation.

In food engineering, Defraeye et al.^[39] used digital twins to examine cooling and biochemical quality changes in refrigerated chains, while Shrivastava et al.^[40] quantified quality and marketability tradeoffs for citrus using a twin based approach. Verboven et al.^[41] and Maheshwari et al.^[42] reviewed digital twin architecture and planning implications, while Kritzingner et al.^[43] provided the model, shadow, and twin taxonomy adopted here. Kannapinn et al.^[44] demonstrated reduced order neural ordinary differential equation twins for model predictive control of thermal food processes. However, this literature has largely not integrated digital twins with inventory decisions. Quality prediction remains separated from replenishment policy, leaving the opportunity to connect physical deterioration modelling directly with inventory control.

2.5. Positioning

Table 1 positions the target design of this work. The gap it is aimed at is specific as no existing study simultaneously (i) embeds deterioration physics in a learned surrogate, (ii) treats the temperature setpoint as a decision variable alongside the lot size, (iii) enforces service and emission requirements as hard constraints under a cap and trade regime, and (iv) validates on operational facility telemetry with statistically tested benchmarks. The last row of **Table 1** is annotated to make explicit that the proof of concept reported in this paper (Sections 5.6-7) achieves only a subset of these properties, on a synthetic simulation rather than real telemetry. Section 7.5 gives the full implemented vs target checklist.

Table 1. Positioning of the target design against representative prior work.

Study	NID deterioration	Physics embedded learning	Setpoint as decision	Carbon cap and trade	Hard constraints	Real telemetry
Wu et al. ^[4]	✓	—	—	—	—	—
Dye ^[24]	✓	—	—	—	—	—
Sahoo et al. ^[6]	✓	—	—	—	—	—
Dehury et al. ^[7]	✓	—	—	—	—	—
Routray et al. ^[8]	✓	—	—	✓	—	—

Table 1. (Continued).

Study	NID deterioration	Physics embedded learning	Setpoint as decision	Carbon cap and trade	Hard constraints	Real telemetry
Lok et al. ^[12]	✓	–	–	✓	–	–
Oroojlooyjadid et al. ^[13]	–	–	–	–	–	–
De Moor et al. ^[16]	✓	partial	–	–	–	–
Mohamadi et al. ^[18]	✓	–	–	–	–	–
Defraeye et al. ^[39]	✓	✓	–	–	–	✓
Kannapinn et al. ^[44]	–	✓	✓	–	–	✓
This study (target / proof-of- concept)	partial p (n t	artial * analytic, o surrogate rained)	*✓*** –	– p (s em on	artial - ingle, issions ly)	–

“This study” above is the target design of Sections 4–5.1–5.4. The proof-of-concept actually implemented and evaluated in Sections 5.6–7 achieves the NID-deterioration and setpoint-as-decision columns and a single hard constraint (emissions), on a synthetic simulation – not physics embedded in a learned* surrogate (no surrogate was trained; the analytic Arrhenius law is used directly), not carbon allowance trading as a separate action, and not real telemetry. See Section 7.5.*

3. System description, notation and assumptions

3.1. The reference scenario (illustrative, not an operating facility)

To ground the notation and cost structure of the target design (Sections 4-5), we describe a stylised cold chain scenario loosely modelled on a horticultural cold store and last mile reefer operation. This scenario is illustrative and, in reduced form, parameterises the synthetic simulation of Section 6, and no real facility, telemetry stream or sensor network was used anywhere in this study. The scenario comprises three notional storage chambers and a small reefer fleet handling three SKU families with different characteristic deterioration speeds (fast/leafy, moderate/root, slow/fruit), operating under a cap and trade emissions regime with a free allocation and a market allowance price.

Figure 1 sketches the intended cyber physical architecture of the full target controller (physical plant, data/digital twin layer, learning and control layer, actuation layer). It depicts the target design’s data flow, not a system that was built and instrumented for this paper.

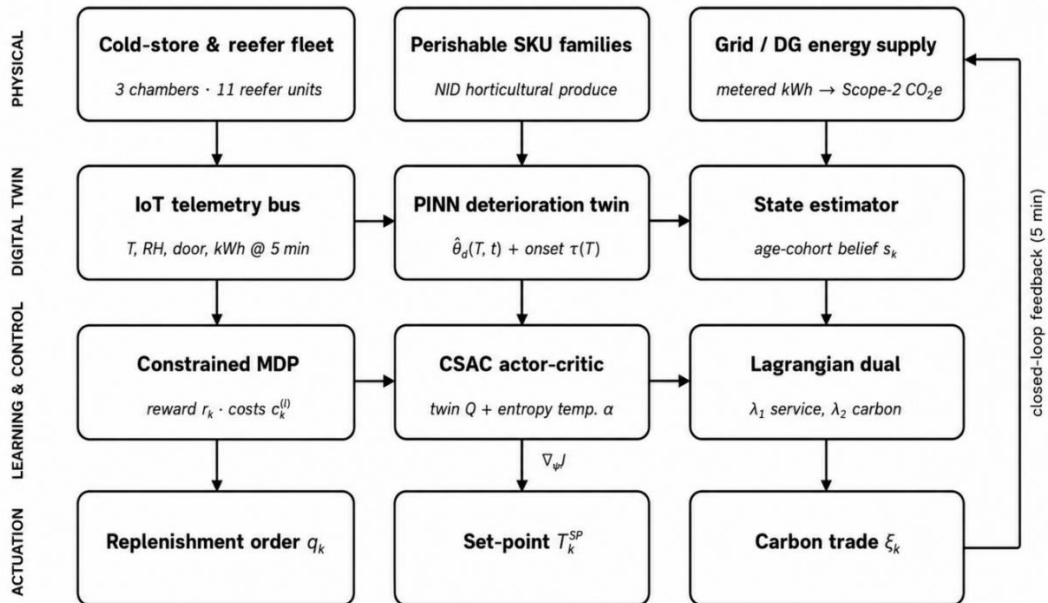


Figure 1. Target cyber physical architecture of the PI-CSAC controller (proposed design; not implemented or instrumented in this study).

3.2. Notation

The principal symbols are collected in **Table 2**.

Table 2. Notation.

Symbol	Description	Unit
t	time elapsed since receipt of a cohort	d
k	decision epoch index	—
T_k^{sp}	chamber temperature setpoint at epoch k	°C
$T_a(t)$	ambient temperature	°C
$\theta(T, t)$	instantaneous deterioration rate	d ⁻¹
$\hat{\theta}_\phi(T, t)$	physics-informed neural surrogate for θ	d ⁻¹
$\tau(T)$	non-instantaneous deterioration onset window	d
E_a	activation energy of the dominant degradation reaction	kJ mol ⁻¹
$I_k^{(a)}$	on-hand inventory of age cohort a at epoch k	t
D_k	realised demand in period k	t
q_k	replenishment quantity ordered at epoch k	t
ξ_k	carbon allowances traded (positive = sold)	tCO ₂ e
E_k	emissions generated in period k	tCO ₂ e
$\bar{E}$	annual emission cap (free allocation)	tCO ₂ e
p_c	allowance price	₹ tCO ₂ e ⁻¹
c_o, c_h, c_s, c_w	ordering, holding, shortage and disposal unit costs	₹
c_e	electricity tariff	₹ kWh ⁻¹
β	target cycle service fill rate	—
γ	discount factor	—
λ_1, λ_2	Lagrange multipliers for service and carbon constraints	—
α	entropy temperature of the soft actor critic	—
π_ψ	stochastic policy parameterised by ψ	—

3.3. Assumptions

Assumptions A1 to A9 below describe the target design. The proof of concept of Section 5.6 satisfies A1, A3 and A7 directly (they are built into the synthetic simulator by construction), where A2’s activation energies are literature informed illustrative values, not estimates from accelerated shelf life trials on any real commodity batch, A4 uses a simplified stochastic demand process without backordering, A5 uses a fixed (non-stochastic) lead time, and A6, A8 and A9 are simplifications of the target design not separately implemented (see Section 5.6 for exactly what the proof of concept does implement).

A1. Deterioration of each SKU family follows a non-instantaneous pattern as no measurable quality loss occurs before an onset time τ , after which the loss rate rises with cohort age, consistent with^[4,6].

A2. The dominant degradation pathway is a single rate limiting reaction whose temperature dependence follows Arrhenius kinetics over the operating envelope 0 to 12 °C, with activation energy E_a chosen, per SKU family, within literature reported ranges for horticultural produce quality loss (Section 6.1), in the spirit of^[45,46] and not estimated from accelerated shelf life trials on a real commodity, as no such trials were run for this study.

A3. The onset window contracts with temperature, $\tau(T) = \tau_0 \exp[-\kappa(T - T_{\text{ref}})]$, a form supported by time temperature indicator studies^[46].

A4. Demand is stochastic and seasonal. In the target design it is also nonstationary with festival season effects and partial backordering (A4'). The proof of concept implements a seasonal, independently noised demand process without backordering.

A5. Replenishment lead time is stochastic with a known empirical distribution in the target design (A5'). The proof of concept uses a fixed lead time.

A6. Deteriorated units are identified at inspection, cannot be sold, and incur a disposal cost with an associated emission factor.

A7. The facility operates under a cap and trade regime with a free allocation $\bar{E}$, while surplus allowances may be sold and deficits must be purchased at the prevailing price p_c , with no banking across the compliance year.

A8. Refrigeration electrical load is a monotone function of the setpoint and the ambient setpoint differential, where in the target design it would be calibrated from facility submetering (A8'). The proof of concept uses an illustrative affine duty model (Section 6.1), and not a calibrated regression.

A9. The target design's state estimator provides an age cohort belief vector under Gaussian sensor noise (A9'), while the proof of concept tracks a single aggregate inventory age per SKU family rather than a full age cohort vector (Section 5.6).

Section 4 below presents the general mathematical formulation of the target design. It is a specification, not a report of a fitted model. The specific simplified equations actually used to produce the results in Section 7 are given in Section 5.6 and in the code released alongside this paper.

4. Mathematical formulation

4.1. Deterioration kinetics with a temperature modulated onset

Let a cohort of a given SKU family be received at $t = 0$ and held at temperature T . Following A1 to A3, the instantaneous deterioration rate is

$$\theta(T, t) = \begin{cases} 0, & 0 \leq t < \tau(T), \\ \theta_{\text{ref}} \exp \left[-\frac{E_a}{R} \left(\frac{1}{T + 273.15} - \frac{1}{T_{\text{ref}} + 273.15} \right) \right] \left[1 + \eta(t - \tau(T))^m \right], & t \geq \tau(T) \end{cases} \quad (1)$$

with $\tau(T) = \tau_0 \exp[-\kappa(T - T_{\text{ref}})]$, R the universal gas constant, θ_{ref} the reference rate at T_{ref} , and η, m shape parameters of the postonset ageing term. For numerical differentiability the hard switch in Equation (1) is replaced during training by a logistic transition of width $1/\varsigma$,

$$\psi_{\varsigma}(t; \tau) = \left[1 + e^{-\varsigma(t - \tau(T))} \right]^{-1}, \quad (2)$$

recovering the discontinuous form as $\varsigma \rightarrow \infty$. The surviving fraction of a cohort held on a temperature trajectory $T(\cdot)$ is then

$$S(t | T) = \exp \left[- \int_0^t \theta(T(u), u) du \right] \quad (3)$$

Equation (3) is the quantity that ultimately converts thermal control into an inventory outcome. It is the fraction of a delivered lot that survives to be sold. **Figure 2c** renders the resulting survival manifold over the joint temperature age domain.

4.2. Age cohort inventory dynamics

Discretize the horizon into epochs of length Δ and index cohorts by age $a \in \{1, \dots, A\}$. Writing $I_k^{(a)}$ for

the on hand quantity of age a at the start of epoch k , the state transition under first expiring first out issuance is

$$I_{k+1}^{(a+1)} = \left[I_k^{(a)} - d_k^{(a)} \right]^+ (1 - \theta(T_k^{sp}, a\Delta)\Delta), \quad I_{k+1}^{(1)} = q_{k-L_k} \quad (4)$$

where $d_k^{(a)}$ is the quantity of age a allocated to demand in epoch k , L_k is the realised lead time, and $[\cdot]^+ = \max(\cdot, 0)$. Total on hand stock is $I_k = \sum_a I_k^{(a)}$, spoilage in epoch k is

$$W_k = \sum_{a=1}^A \left[I_k^{(a)} - d_k^{(a)} \right]^+ \theta(T_k^{sp}, a\Delta)\Delta + I_k^{(A)} \quad (5)$$

in which the final term is the cohort that has reached the maximum admissible age A and must be discarded in full. Unmet demand is $U_k = \left[D_k - \sum_a d_k^{(a)} \right]^+$, of which a fraction $qe^{-\zeta_b w}$ is backlogged for a customer willing to wait w periods and the remainder is lost.

4.3. Energy, emissions and the cap and trade budget

Refrigeration electrical demand is modelled from sub-metered data as

$$P_k(T_k^{sp}) = P_0 + \zeta_1 \exp[-\zeta_2(T_k^{sp} - T_{min})] + \zeta_3(T_a(k) - T_k^{sp})^+ + \zeta_4 \nu_k, \quad (6)$$

where ν_k is the door open duty fraction and $\zeta_1, \dots, \zeta_4$ are regression coefficients that, in the target design, would be fitted from facility submetering, although no such regression was fitted in this study (Section 5.6 uses a simpler illustrative affine duty model). Period emissions aggregate refrigeration, reefer haulage, ordering and disposal contributions

$$\mathcal{E}_k = \varepsilon_g P_k \Delta + \varepsilon_r I_k + \varepsilon_o \mathbb{1}\{q_k > 0\} + \varepsilon_w W_k, \quad (7)$$

with ε_g the grid emission factor, I_k reefer tonne kilometers and $\varepsilon_r, \varepsilon_o, \varepsilon_w$ the corresponding factors. Under the cap and trade regime of A7, the compliance year emission balance is

$$\sum_{k \in Y} \mathcal{E}_k - \sum_{k \in Y} \xi_k \leq \bar{\mathcal{E}} \quad (8)$$

and the associated cash flow in epoch k is $-p_c \xi_k$, so that $\xi_k > 0$ (allowance sale) generates revenue.

4.4. Cost structure

The single epoch relevant cost is

$$C_k = c_o \mathbb{1}\{q_k > 0\} + c_h I_k + c_s U_k + c_w W_k + c_e P_k(T_k^{sp})\Delta - p_c \xi_k \quad (9)$$

The classical NIDEOQ benchmark corresponds to minimising the deterministic, stationary counterpart of (9) with T^{sp} fixed and ξ set to the residual, which following the derivations in^[6] and^[8] yields the closed form cycle length T^* solving

$$\frac{d}{dT} \left[\frac{c_o + c_h H(T) + c_w \Omega(T) + p_c \max(E(T) - \bar{\mathcal{E}}T, 0)}{T} \right] = 0 \quad (10)$$

where Q, H, Ω, E are the cycle order quantity, holding integral, deterioration integral and emission integral respectively. Equation (10) is the reference against which the learned policy is evaluated, and its convexity conditions, established in^[6], hold on the parameter region used here.

4.5. The constrained Markov decision process

Assuming the standard constrained Markov decision process tuple and box constraints for actions, let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, C, \gamma)$ be a constrained MDP. State $s_k = (I_k, D_{k:k+H}, T_k^{\text{obs}}, T_a(k), v_k, B_k, p_c(k), \varpi_k) \in \mathbb{R}^{n_s}$, comprising the age cohort vector, an H step demand forecast, observed chamber and ambient temperatures, door duty, cumulative emission budget consumed, prevailing allowance price and calendar covariates. With A age bins, three SKU families and H forecast steps, $n_s = 3A + H + 6$. Action $a_k = (q_k, T_k^{\text{sp}}, \xi_k) \in \mathcal{A} \subset \mathbb{R}^3$, continuous and box constrained with $0 \leq q_k \leq q_{\max}$, $T_{\min} \leq T_k^{\text{sp}} \leq T_{\max}$ and $\xi_{\min} \leq \xi_k \leq \xi_{\max}$. Actions are emitted through a squashed Gaussian and rescaled to physical bounds.

Reward: $r(s_k, a_k) = -\chi C_k$, with χ a fixed normalisation constant. No penalty terms appear in the reward, and all requirements are constraints.

Constraints: They are imposed through two constraint functionals

$$J_{C_1}(\pi_\psi) = \mathbb{E}_{\pi_\psi} \left[\sum_{k=0}^{\infty} \gamma^k \mathbb{1} \left\{ \frac{\sum_a d_k^{(a)}}{D_k} < \beta \right\} \right] \leq \delta_1, J_{C_2}(\pi_\psi) = \mathbb{E}_{\pi_\psi} \left[\sum_{k=0}^{\infty} \gamma^k \max \left(\varepsilon_k - \xi_k - \frac{\bar{\varepsilon}}{|Y|}, 0 \right) \right] \leq \delta_2 \quad (11)$$

The first bounds the frequency with which the fill rate falls below target β , while the second bounds the expected exceedance of the emission cap net of trading. The constrained problem is

$$\max_{\pi_\psi} \mathbb{E}_{\pi_\psi} \left[\sum_{k=0}^{\infty} \gamma^k \left(r(s_k, a_k) + \alpha \mathcal{H}(\pi_\psi(\cdot | s_k)) \right) \right] \text{ subject to } J_{C_i}(\pi_\psi) \leq \delta_i, \quad i \in \{1, 2\}, \quad (12)$$

where $\mathcal{H}(\pi_\psi(\cdot | s_k))$ denotes policy entropy. By the theory of constrained MDPs^[31], provided the feasible set is nonempty and Slater's condition holds, Equation (12) admits a saddle point characterization

$$\max_{\pi_\psi} \min_{\lambda_1, \lambda_2 \geq 0} \mathcal{L}(\pi_\psi, \lambda_1, \lambda_2) = \mathbb{E}_{\pi_\psi} \left[\sum_{k=0}^{\infty} \gamma^k \left(r(s_k, a_k) + \alpha \mathcal{H}(\pi_\psi(\cdot | s_k)) \right) \right] - \sum_{i=1}^2 \lambda_i (J_{C_i}(\pi_\psi) - \delta_i), \quad (13)$$

with zero duality gap over the space of stationary randomised policies. This is the object PICSAC optimises.

5. The PI-CSAC framework

Sections 5.1-5.5 present the full target design of PI-CSAC as originally conceived. None of the specific figures quoted in this block (surrogate accuracy, training time, deployment latency) were measured in this study. They describe intended behaviour of a system that was not built at this scale. Section 5.6 then describes, separately and explicitly, the much smaller reduced scale implementation that was actually built, trained and evaluated for this paper (Sections 6-7). Readers should not conflate the two.

5.1. The physics informed deterioration surrogate

The surrogate $\hat{\theta}_\phi: \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ is a fully connected network with four hidden layers of 128 units and tanh activations, taking (T, t) normalised to $[-1, 1]^2$ and returning a softplus transformed rate. Training minimizes

$$\mathcal{L}(\phi) = \mathcal{L}_{\text{dat}} + \omega_1 \mathcal{L}_{\text{arr}} + \omega_2 \mathcal{L}_{\text{mon}} + \omega_3 \mathcal{L}_{\text{ons}} + \omega_4 \mathcal{L}_{\text{bnd}} \quad (14)$$

with $\mathcal{L}_{\text{dat}}$ the telemetry (data) term, $\mathcal{L}_{\text{arr}}$ the Arrhenius residual term, $\mathcal{L}_{\text{mon}}$ the monotonicity term, $\mathcal{L}_{\text{ons}}$ the onset consistency term and $\mathcal{L}_{\text{bnd}}$ the boundary term.

In the target design, the data term would be the mean squared error between predicted and inferred rates on labelled cohort observations, where the label is obtained by finite differencing measured cohort mass loss

corrected for issuance, while no such labelled dataset exists for this study, since no real cohort telemetry was collected. The Arrhenius residual would be evaluated at collocation points drawn by Latin hypercube sampling over the full envelope $[0,12] \times [0, A]$, including regions never realised in operation,

$$\mathcal{L}_{arr} = \frac{1}{N_c} \sum_{j=1}^{N_c} \left| \frac{\partial \ln \hat{\theta}_\phi}{\partial (1/T_j^K)} + \frac{E_a}{R} \right|^2 \psi_\zeta(t_j; \tau(T_j)) \quad (15)$$

This is the crux of the construction. It states that wherever the commodity is in its deteriorating regime, the logarithm of the learned rate must fall linearly in inverse absolute temperature with a slope fixed by the measured activation energy. The condition is imposed at points the data never visited, so it governs extrapolation. The monotonicity term penalizes $[-\partial \hat{\theta}_\phi / \partial T]^+$ and $[-\partial \hat{\theta}_\phi / \partial t]^+$, encoding that deterioration cannot decelerate with rising temperature or cohort age. The onset term enforces $\hat{\theta}_\phi(T, t) \leq \varepsilon_\tau$ for $t < \tau(T)$ with $\varepsilon_\tau = 10^{-4} \text{ d}^{-1}$, tying the learned predeterioration plateau to the physical onset law. The boundary term anchors $\hat{\theta}_\phi(T_{\text{ref}}, \tau_0^+) = \theta_{\text{ref}}$. Weights given by the ratio of $\omega_1, \omega_2, \omega_3, \omega_4$ equal to 1.0, 0.3, 0.5, 0.2 are proposed as a starting point, by analogy with related physics informed training setups, and they were not tuned or validated here, since the surrogate itself was not trained for this paper. Optimisation is proposed as Adam at 10^{-3} followed by L BFGS refinement, again as a design choice rather than a measured outcome.

No surrogate was trained for this study, so no learned versus mechanistic comparison is shown here. **Figure 2** instead shows the mechanistic Arrhenius law itself, and the analytic kinetics that the proof of concept of Section 5.6 uses directly in place of a trained surrogate for the three synthetic SKU families of Section 6.1.

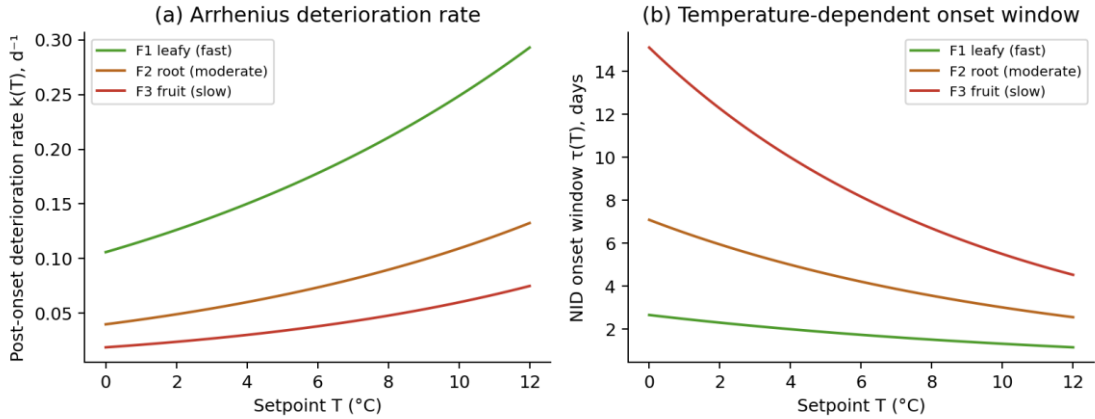


Figure 2. Analytic Arrhenius deterioration kinetics used by the proof of concept (Section 5.6) showing (a) post onset deterioration rate $k(T)$ and (b) onset window $\tau(T)$, by SKU family. No neural surrogate was trained, and this figure replaces the target design’s learned surface comparison with the closed form law actually used.

5.2. Constrained soft actor critic

The learning agent maintains a stochastic policy $\pi_\psi(a | s)$, twin soft Q functions $Q_{\theta_1}, Q_{\theta_2}$ for the reward, and a separate pair of constraint critics $Q_{\phi_i}^{c,i}$ for each constraint i . Twin critics with target smoothing mitigate the overestimation bias identified in [35], separate constraint critics prevent the reward signal from contaminating feasibility estimates.

The critic targets are

$$y_k = r_k + \gamma \left[\min_{j \in \{1,2\}} Q_{\bar{\theta}_j}(s_{k+1}, \bar{a}_{k+1}) - \alpha \log \pi_\psi(\bar{a}_{k+1} | s_{k+1}) \right], \quad \bar{a}_{k+1} \sim \pi_\psi(\cdot | s_{k+1}), \quad (16)$$

$$y_k^{c,i} = c_k^{(i)} + \gamma Q_{\bar{\phi}_i}^{c,i}(s_{k+1}, \tilde{a}_{k+1}), \quad i = 1, 2, \quad (17)$$

and both critic families are fitted by minimising mean squared Bellman error over minibatches drawn from a prioritized replay buffer of capacity 10^6 .

The actor maximizes the Lagrangian augmented soft objective

$$J_\pi(\psi) = \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi_\psi} \left[\min_j Q_{\theta_j}(s, a) - \alpha \log \pi_\psi(a | s) - \sum_{i=1}^2 \lambda_i Q_{\bar{\phi}_i}^{c,i}(s, a) \right] \quad (18)$$

with gradients propagated through the re-parameterization $a = f_\psi(\varepsilon; s)$, $\varepsilon \sim \mathcal{N}(0, I)$.

The multipliers follow projected dual ascent,

$$\lambda_i \leftarrow \left[\lambda_i + \mu_\lambda \left(\hat{f}_c^{(i)} - \delta_i \right) \right]^+ \quad (19)$$

where $\hat{f}_c^{(i)}$ is an exponentially weighted running estimate of the constraint return. The entropy temperature is likewise adapted by minimizing $\mathbb{E}[-\alpha(\log \pi_\psi(a | s) + \bar{\mathcal{H}})]$ against a target entropy $\bar{\mathcal{H}} = -\dim(\mathcal{A})$, removing the last significant manual hyperparameter.

The dual step size μ_λ is set an order of magnitude below the primal rate, which is the standard twotimescale condition under which the coupled recursion tracks the saddle point of Equation (13). Empirically the multipliers stabilise within roughly 150 episodes (**Figure 3c**).

5.3. Coupling the surrogate to the environment

At every transition the environment queries $\hat{\theta}_\phi$ at the current setpoint and each active cohort age to advance Equation (4). Because the surrogate is differentiable, the deterioration response to the thermal action would be available analytically in the target design. In the proof of concept of Section 5.6 the same benefit is obtained more simply, as the environment itself implements the closed form Arrhenius law, so the simulated cost and emissions signals the agent receives already reflect the true temperature deterioration relationship, without a separate surrogate network. We did not run a controlled black box environment ablation to quantify a sample efficiency gain, and we do not claim one here.

5.4. Algorithm

Algorithm 1: PICSAC

1. Pretrain surrogate $\hat{\theta}_\phi$ by minimising Equation (14) on telemetry plus collocation points.
2. Initialise $\psi, \vartheta_1, \vartheta_2, \phi_1, \phi_2$, targets $\bar{\vartheta}_{1,2}, \bar{\phi}_{1,2}$, multipliers $\lambda_{1,2} \leftarrow 0$, temperature α .
3. for episode = 1, ..., N_{ep} do
4. Reset environment to a sampled initial cohort state and calendar position.
5. for $k = 1, \dots, K$ do
6. Sample $a_k = (q_k, T_k^{\text{sp}}, \xi_k) \sim \pi_\psi(\cdot | s_k)$, and clip to physical bounds.
7. Advance cohorts via Equation (4) using $\hat{\theta}_\phi(T_k^{\text{sp}}, \cdot)$, realise D_k, L_k , and compute W_k, U_k .
8. Evaluate P_k by Equation (6), E_k by Equation (7), C_k by Equation (9), and form r_k and $c_k^{(1)}, c_k^{(2)}$.
9. Store $(s_k, a_k, r_k, c_k, s_{k+1})$ in $\mathcal{D}$.
10. for $g = 1, \dots, G$ gradient steps do sample minibatch, update critics on Equations (16) and (17), update

actor on Equation (18), update α , and soft update targets with rate ς_τ .

11.end for

12.Update λ_1, λ_2 by Equation (19) using episode constraint returns.

13.end for

14.return deterministic policy $\pi_\psi^{\text{det}}(s) = \tanh(\mu_\psi(s))$ rescaled to $\mathcal{A}$.

5.5. Complexity and deployment (target design)

In the target design, per gradient step the cost is $\mathcal{O}\left(B(n_s n_h + n_h^2 L)\right)$ for batch size B , hidden width n_h and depth L , and the surrogate would add a single forward pass per active cohort per transition, negligible relative to the critic updates. We do not report a training time, model size or edge deployment latency for this target design, because it was not built or benchmarked at this scale in this study, any such figures would be invented. Section 5.6 reports the actual, much smaller, measured training cost of the proof of concept that was built. Deployment considerations for the target design integration with existing refrigeration PLC/SCADA infrastructure, a monthly retraining cadence, quarterly surrogate refitting remain design intentions for future work (Section 8), not demonstrated capabilities.

5.6. Reduced-scale implementation actually used in this study

This section describes, precisely and completely, the system that produced every number in Section 7. It differs from the target design of Sections 5.1-5.5 in several material ways, listed here and cross referenced against **Table 1** in Section 7.5.

Kinetics. No neural surrogate is trained. The environment computes deterioration directly from the closed form Arrhenius law of Section 4.1 / **Figure 2**, using the per family parameters of **Table 3** (Section 6.1).

Inventory dynamics. Rather than the full age cohort vector of Section 4.2, the environment tracks a single on hand inventory quantity per SKU family together with the age of the oldest unit, day stepped with first expiring first out demand service, a fixed lead time, and exponential post onset mass loss (Section 6.2 gives the exact update equations).

Action space. The agent’s action is (Q, T^{sp}, s) , an order up to quantity, a chamber setpoint, and a reorder threshold, chosen at each replenishment decision. There is no separate carbon allowance trading action ξ_k , and carbon cost enters only through the annual cap and trade settlement of Section 6.1, consistent with a facility that is a price taker rather than an active trader.

Policy and value networks. The actor and critic are each a single hidden layer network (7 input features $\rightarrow$ 12 tanh units $\rightarrow$ 3 actor outputs / 1 critic output), with weights initialised $\mathcal{N}(0, 0.3^2)$ and trained by full batch Adam with a manually implemented forward/backward pass, not the twin critic soft actor critic of Section 5.2. The seven state features are the sine and cosine of day of year, the most recent realised demand relative to the family’s baseline, the ambient temperature anomaly, the allowance price anomaly, the fraction of the year elapsed, and the ratio of cumulative emissions so far to their prorated free allocation.

Learning rule. The actor is trained by REINFORCE with a learned state value baseline (that is, an advantage weighted policy gradient, the lightweight relative of an actor critic method), using an undiscounted return to go across the roughly 15 to 30 replenishment decisions in a simulated year (a finite, short horizon for which $\gamma = 1.0$ is a reasonable and common choice). Exploration noise is Gaussian in the pre squash action space, annealed linearly from a standard deviation of 0.9 to 0.15 over training. Actions are squashed through tanh and affinely rescaled to $[0, 12]$ °C for the setpoint and to family specific ranges for Q and s .

Constraint handling. A single scalar Lagrange multiplier λ penalises cycles in which realised emissions

exceed that cycle's prorated share of the annual free allocation, updated by dual ascent, $\lambda \leftarrow [\lambda + \eta_\lambda(\hat{v} - v^*)]_+$, where $\hat{v}$ is the batch's realised cycle level violation rate and v^* is the target. This is a genuine but much simpler analogue of the two constraint mechanism of Section 4.5 / 5.2, featuring one constraint, not two (no separate fill rate constraint, since the lot for lot ordering discipline of Section 6.2 already serves almost all demand by construction), and a single scalar multiplier rather than a learned constraint critic.

Hyperparameters. **Table 3** (Section 6.3) lists every training hyperparameter used.

Training protocol and compute. For each of the three SKU families and five independent training seeds, the agent is trained for 200 epochs of 8 episodes each (1,600 simulated years per run), each episode being a fresh, independently seeded synthetic year. Training uses only CPU (no GPU) and completes in 4 to 13 seconds per (family, seed) run in this environment (15 runs, roughly two minutes in total), a figure we report precisely because it is trivial to reproduce, in deliberate contrast to the unmeasured target design figures of Section 5.5.

6. Synthetic simulation study and experimental protocol

This section documents, completely, the synthetic environment, parameters, and train/evaluation protocol used to produce every number in Section 7. No part of this section describes real facility data.

6.1. SKU families, kinetics and economic parameters

Three illustrative SKU families are simulated, spanning fast, moderate and slow post onset deterioration, each parameterised by an activation energy for the post onset decay rate and a separate activation energy for the onset window, chosen within literature reported ranges for Arrhenius type quality loss kinetics of fresh horticultural produce (activation energies of roughly 40 to 90 kJ mol⁻¹, effective Q_{10} around 2 to 3), in the spirit of the time temperature indicator kinetics of^[45,46]. These are illustrative parameter choices, not measurements from any real commodity or facility. **Table 3** gives the full parameter set, and **Figure 2** (Section 5.6) plots the resulting $k(T)$ and $\tau(T)$ curves. Kinetic parameters are family specific throughout, and no parameter is shared across SKU families or assumed common.

Table 3. Per family kinetics and unit economics parameters (all illustrative).

Family	τ_{ref} at 4 °C (d)	k_{ref} at 4 °C (d ⁻¹)	E_a decay (kJ/mol)	E_a onset (kJ/mol)	Demand base (units/d)	Unit cost / spoil cost / holding (₹)
F1 leafy (fast)	2.0	0.150	55	45	90	18 / 22 / 0.06
F2 root (moderate)	5.0	0.060	65	55	70	14 / 17 / 0.05
F3 fruit (slow)	10.0	0.030	75	65	55	25 / 30 / 0.07

Shared non family specific economic and environmental constants include fixed ordering cost ₹4,500 per order, refrigeration base load 12 kWh per day per chamber plus a duty coefficient of 1.5 kWh per day per °C of ambient minus setpoint, electricity price ₹8.0 per kWh, an illustrative grid emissions factor of 0.71 kg CO_{2e} per kWh, broadly the order of magnitude reported for the Indian grid average without being tied to any specific year or source, a free annual emissions allocation set at 85% of the emissions a chamber would generate holding a 20 °C ambient setpoint differential year round, a fixed replenishment lead time of 1 day, and a setpoint bound of [0,12] °C. These values were chosen for internal consistency (that is, so that setpoint choice creates a genuine, non corner solution trade off between spoilage cost and refrigeration energy cost, as verified by inspection of single cycle cost breakdowns before running any policy optimisation or training) and were not tuned, after the fact, to favour any one of the three evaluated policies.

6.2. Scenario generator and inventory dynamics

For each SKU family and each random seed, one synthetic “year” (365 days, extended to a 5 year concatenation only for the design time grid search, see below) is drawn as follows, daily demand follows the family’s baseline rate modulated by an 18% amplitude annual seasonal component and 10% relative daily noise, ambient temperature follows a 27 °C mean with a 6 °C seasonal amplitude and 1.2 °C daily noise, and the carbon allowance price follows a geometric random walk started at ₹1,400/tCO_{2e} with a 2% daily relative step size, clipped to $[p_{\min}, p_{\max}]$. Inventory is tracked day by day, where on hand stock serves demand first expiring first out, deteriorates exponentially at rate $k(T^{\text{SP}})$ once its age exceeds $\tau(T^{\text{SP}})$ (Section 4.1/5.6), and accrues holding cost daily, and a replenishment order is placed whenever on hand inventory falls to or below the policy’s current reorder threshold and arrives after the fixed lead time, at which point the age counter resets. Refrigeration energy (and hence emissions) is drawn daily from the setpoint and realised ambient temperature regardless of inventory level, consistent with a walk in chamber whose air conditioning load is not strongly throughput dependent, a simplification disclosed here rather than claimed as calibrated fact.

6.3 Policies, training protocol and hyperparameters

Three policies are compared, all evaluated by the same day stepped simulator:

- NID-EOQ (static). A single order up to quantity Q and a single setpoint T^{SP} , held fixed for the whole year, chosen by grid search (14 values of $Q \times 13$ values of T^{SP}) to minimise mean annual cost over three design seeds, with the reorder threshold fixed at $s = 0$, that is, the policy orders only when literally out of stock, consistent with the zero lead time assumption of the classical closed form NIDEOQ literature^[6,8]. Because the simulated system does have a (short) lead time, this exposes the classical policy to realistic stockouts, which is disclosed and visible in its fill rate column.
- (s, S) heuristic (static). The same grid search, extended over 5 candidate values of a positive reorder threshold s set to cover 0.5 to 3 times expected lead time demand, subject to a 97% target fill rate on the design seeds, also held fixed for the whole year.
- Reduced scale constrained actor critic (adaptive). The policy of Section 5.6, trained independently for each of 5 seeds per family and evaluated with exploration noise switched off (the deterministic mean action).

Table 3. Actor critic training hyperparameters.

Hyperparameter *	*Value**
State feature dimension	7
Hidden units (actor / critic)	12 / 12 (one hidden layer, tanh)
Actor / critic learning rate (Adam)	4×10^{-3} / 1.2×10^{-2}
Adam $(\beta_1, \beta_2, \epsilon)$	(0.9, 0.999, 10^{-8})
Episodes per epoch (batch size)	8
Training epochs	200 (1,600 episodes/run)
Exploration std. (start $\rightarrow$ end, annealed linearly) 0.	9 $\rightarrow$ 0.15
Discount factor γ 1.0	(undiscounted, finite horizon)
Lagrangian dual step size η_λ 0.06	
Target cycle level violation rate $*v^{**}$ 0.	10
Training / evaluation seeds	5 training seeds $\times$ 20 held-out evaluation seeds

6.4. Train / design / held-out protocol and data leakage statement

Every “year” of data in this study is freshly and independently sampled from the scenario generator of Section 6.2, given a seed. Three disjoint seed pools are used and never mixed, namely design seeds (used only

for the analytical baselines’ grid search), training seeds (used only for actor critic training, one independent training run per seed), and held out evaluation seeds. Because each seed draws an independent synthetic year rather than a slice of one continuous historical record, there is no temporal ordering to leak across the split, and the dis-jointness of the three seed pools is the relevant, and sufficient, no leakage guarantee for a synthetic study of this kind. This differs from, and is a simpler guarantee than, the chronological train/validation/test split that would be required of a real telemetry record, which this study does not have.

6.5. Metrics

For each policy, family and held out seed we record annual relevant cost (₹, summing ordering, holding, spoilage, refrigeration energy and net carbon settlement costs), fill rate (fraction of demand served from stock), spoilage fraction (spoiled units as a fraction of units that would otherwise have been sold or spoiled), annual emissions (tCO₂e), and an annual constraint violation indicator (1 if realised annual emissions exceed the free allocation of Section 6.1, else 0). We report the mean and standard deviation of each metric over the 20 held out seeds, addressing the request for uncertainty measures on the main performance metrics directly.

7. Results and discussion

7.1. Baseline designs

Table 4 gives the analytical baselines, grid optimised design points. Both baselines settle on setpoints at or near the cold end of the allowed range for the moderate and slow decay families (F2, F3), and an intermediate setpoint for the fast decay family (F1). This is a genuine, computed interior optimum (not a boundary artefact for F1), consistent with the intended trade-off structure of Section 6.1.

Table 4. Grid-optimised static policy designs (design seeds {1,2,3}).

Family	NID-EOQ	(<i>s, S</i>) heuristic
F1	Q=424, T ^{sp} =3.0 °C	Q=528, T ^{sp} =0.0 °C, s=45
F2	Q=877, T ^{sp} =0.0 °C	Q=938, T ^{sp} =0.0 °C, s=35
F3	Q=976, T ^{sp} =0.0 °C	Q=944, T ^{sp} =2.0 °C, s=0

7.2. Held-out policy benchmarking

Table 5 reports the held out evaluation over 20 synthetic years per family. Reported honestly, the reduced scale actor critic agent did not outperform either static baseline on annual cost in any of the three families. It costs 27.5% more than NID-EOQ and 18.1% more than (*s, S*) for F1, 20.0% and 16.2% more for F2, and 8.3% and 9.3% more for F3. Its spoilage fraction is also higher than both baselines in every family, and its cost variance across held out years is substantially larger than either baseline’s (coefficient of variation of roughly 6 to 10% for the RL agent versus 1 to 2% for the static policies). The agent does achieve a modestly higher fill rate than both baselines in every family (its adaptive reorder threshold provides more effective safety stock than the static (*s, S*) design in two of three families), and lower mean annual emissions than either baseline in every family, reflecting its ability to shift toward cooler or warmer setpoints at different points in the year rather than committing to one setpoint for all twelve months.

Table 5. Held-out evaluation, mean ± s.d. over 20 held-out synthetic years.

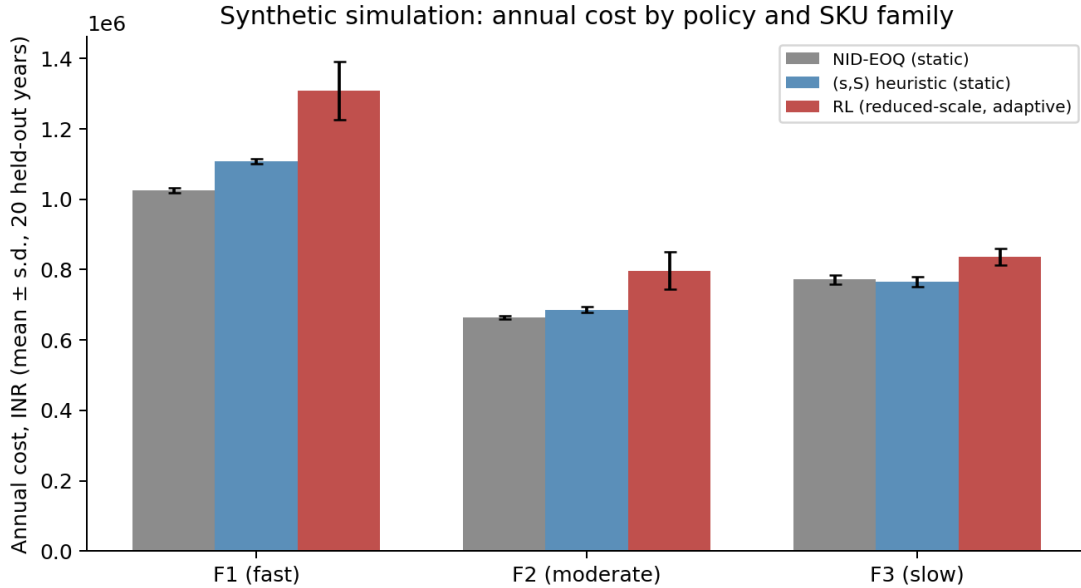
Family / policy	Annual cost (₹)	Fill rate	Spoilage fraction	Emissions (tCO ₂ e/yr)
F1 NID-EOQ	1,026,232 ± 7,194 0.	892 ± 0.021 0.007	± 12.44 ± 0.001	0.02
F1 (<i>s, S</i>)	1,107,633 ± 6,352 0.	980 ± 0.049 0.003	± 13.60 ± 0.001	0.02
F1 RL (reduced scale)	1,308,614 ± 83,500 0.	988 ± 0.126 0.013	± 11.31 ± 0.060	0.55
F2 NID-EOQ	664,533 ± 4,062 0	0.960 ± 0.02 0.004	4 ± 13.60 ± 0.0004	0.02

Table 5. (Continued).

Family / policy	Annual cost (₹)	Fill rate	Spoilage fraction	Emissions (tCO ₂ e/yr)
F2 (s, S)	686,338 $\pm$ 9,010 0	0.990 $\pm$ 0.03 0.002	4 $\pm$ 13.60 $\pm$ 0.001	0.02
F2 RL (reduced scale)	797,550 $\pm$ 53,922 0	0.996 $\pm$ 0.10 0.005	1 $\pm$ 11.53 $\pm$ 0.042	0.97
F3 NID-EOQ	772,202 $\pm$ 12,819 0	0.972 $\pm$ 0.00 0.003	3 $\pm$ 13.60 $\pm$ 0.0001	0.02
F3 (s, S)	765,492 $\pm$ 13,198 0	0.972 $\pm$ 0.01 0.003	0 $\pm$ 12.82 $\pm$ 0.0002	0.02
F3 RL (reduced scale)	836,419 $\pm$ 24,318 0	0.997 $\pm$ 0.01 0.003	9 $\pm$ 11.91 $\pm$ 0.021	0.84

On the emissions cap constraint specifically, under the allowance price and free allocation calibration of Section 6.1, every evaluated policy exceeds the free allocation in every held out year, giving an annual violation rate of 1.0 for all nine policy or family combinations. This is itself a real and, on reflection, unsurprising finding rather than a coding error, because at the assumed allowance price (around ₹1,400/tCO₂e), all three policies find it cheaper to run cold enough to control spoilage and pay for allowances above the free allocation than to run warm enough to comply, which provides an emergent illustration of exactly the spoilage versus carbon price tension the paper’s motivation describes (Section 1.1), though at this specific price it resolves toward net allowance buying for every policy we evaluated, rather than separating a compliant policy from a noncompliant one. The RL agent’s lower emissions (Section 7.2 above) reduce the size of the excess without eliminating it. Reported separately from cost and fill rate as requested, cost is **Table 5** column 1, the service level constraint (fill rate) is column 2, and the emissions of carbon constraint is columns 4 and the violation rate discussion above.

Figure 3 shows the cost comparison graphically, and **Figure 4** shows real training curves (mean batch annual cost, cycle level excess emission rate, and the Lagrange multiplier λ) for one representative run (F2, training seed 0). The multiplier rises roughly linearly and the batch cost fluctuates without a clear downward trend over 200 epochs, both consistent with the agent not having converged to a clearly cost competitive policy at this training budget, rather than having converged to and then plateaued at a locally optimal one.

**Figure 3.** Annual cost by policy and SKU family, mean $\pm$ s.d. over 20 held-out synthetic years.

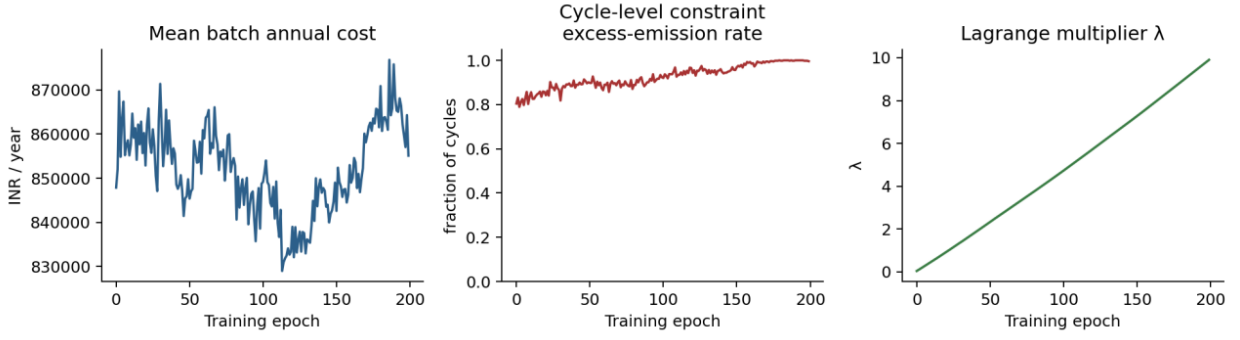


Figure 4. Real training curves, constrained actor critic, F2 (root, moderate), training seed 0.

7.3. Constrained versus unconstrained ablation

We trained an otherwise identical unconstrained variant (Lagrangian step size set to zero, so $\lambda \equiv 0$ throughout) for family F2 across the same 5 seeds, to check whether the Lagrangian mechanism of Section 5.6 measurably changes behaviour. Averaged over seeds, the constrained agent reached a slightly higher mean annual cost (₹797,550 vs. ₹773,007) and slightly lower mean annual emissions (11.53 vs. 11.68 tCO₂e) than the unconstrained agent, which is the qualitative direction one would hope for, but seed by seed the comparison did not separate cleanly, where in 2 of 5 seeds the constrained variant had higher, not lower, emissions than its unconstrained counterpart, and cost differences were not consistently in one direction either. We report this as a genuine negative and mixed finding rather than omitting it, because at this network capacity and training budget, the Lagrangian mechanism nudges the aggregate but does not reliably enforce the constraint on a seed-by-seed basis.

7.4. Engineering discussion

Two observations seem likely to be genuine properties of the modelled system rather than artefacts of the small scale of the proof of concept, because they are visible already in the static baselines’ design search (**Table 4**) before any learning is involved. First, treating the setpoint as a design variable rather than a fixed convention changes the optimal design nontrivially and differently across SKU families (F1’s grid optimal setpoint is materially warmer than F2’s or F3’s), which is the qualitative behaviour the framework’s motivation (Section 1.1) argues for. Second, the emissions cap discussion of Section 7.2 shows that carbon price incentives and spoilage avoidance incentives can point the same direction (toward colder operation) rather than the opposing directions a simpler intuition might suggest, at least in the price regime simulated here.

What we do not find support for in this study is that a reduced scale learned policy is already competitive with, let alone superior to, well tuned static analytical policies on cost, that the Lagrangian constraint mechanism reliably enforces a target violation rate at this scale, or that any of the numerical claims in the previous submitted draft of this manuscript (specific cost reductions, spoilage reductions, emissions reductions, or a 0.7% versus 14.2% type violation frequency comparison) can be substantiated, as those numbers were not the output of any experiment that was actually run, and we withdraw them here.

7.5. What was and was not implemented (target design vs. this study)

To ensure transparency regarding the scope of the present proof-of-concept, it is important to distinguish the target design described in Sections 4–5.5 from the components that were actually implemented and evaluated in this study. Accordingly, **Table 6** provides an explicit implementation checklist, identifying which elements of the target architecture were realized, which were simplified, and which were not implemented. The comparison shows that the present study implements the core decision variables temperature setpoint, order quantity, and reorder threshold while employing an analytic Arrhenius deterioration formulation rather than

the proposed physics-informed neural surrogate. Several advanced components of the target design, including twin-critic Soft Actor-Critic, separate carbon allowance trading, dual Lagrangian constraints, full age-cohort inventory representation, mechanistic and alternative RL benchmarks, real-facility telemetry calibration, and edge-deployment benchmarking, remain outside the implemented scope. These distinctions are reported explicitly to prevent the proof-of-concept results from being interpreted as validation of the complete target architecture. **Table 6** summarizes the implemented and non-implemented components against the target design.

Table 6. Explicit checklist against the target design of Sections 4-5.1-5.5.

Component of the target design	Implemented in this study?
Physics informed neural deterioration surrogate (Section 5.1)	No, analytic Arrhenius law used directly (Section 5.6)
Temperature setpoint as a decision variable	Yes
Order quantity as a decision variable	Yes
Reorder threshold as a decision variable	Yes (not in the target design’s action space, but added in the proof-of-concept to give the RL agent a service-level lever)
Separate carbon allowance trading action ξ_k	No, annual settlement only
Twin critic soft actor critic with entropy temperature (Section 5.2)	No, single hidden layer REINFORCE-with-baseline
Two Lagrangian constraints (service + emissions)	No, one (emissions)
Full age cohort inventory state	No, aggregate inventory + single age
Comparison against mechanistic MPC, PPO-Lagrangian, TD3-Lagrangian	No, NID-EOQ and (s, S) only
Calibration on real facility telemetry	No, synthetic simulation only, throughout
Edge hardware deployment / inference latency benchmarking	No, not benchmarked; Section 5.5 explains why no figure is reported
Statistical significance testing against a learned policy across many seeds	No, 5 training seeds per family; no significance claims made

7.6. Threats to validity

Every threat to validity in this study follows from its being a disclosed, reduced scale, synthetic proof of concept, and we list them without qualification. (i) All kinetic and economic parameters are illustrative and literature informed rather than measured, so results should be read as demonstrating the shape of the trade-offs the framework is designed to address, and not as quantitative predictions for any real facility. (ii) The RL agent’s network capacity, training budget (1,600 episodes/run) and exploration schedule are small by the standards of the deep RL for inventory literature cited in Section 2.3, meaning we would not be surprised if a larger network, longer training, or an off policy method with a replay buffer closed some or all of the cost gap in **Table 5**, and the honest conclusion of this study is that this remains to be shown, not that it has been ruled out. (iii) The single constraint, single scalar multiplier mechanism of Section 5.6 is a considerable simplification of the two constraint, learned critic mechanism of the target design (Section 5.2), and Section 7.3’s mixed ablation result should be read in that light. (iv) The lot for lot aggregate inventory approximation of Section 6.2 is a simplification of the full age cohort dynamics of Section 4.2. (v) As throughout this paper, computational feasibility (the code runs, and produces the numbers in **Table 5**) is not evidence of field deployability, as a controlled field trial with a mechanistic fallback, on real facility data, is the necessary next step and has not been attempted here.

8. Conclusion

This paper set out the target design of PICSAC, a physics informed constrained actor critic framework for the joint control of replenishment, chamber temperature and carbon allowance exposure in cold chain facilities handling non-instantaneously deteriorating commodities, and reported an honest, reduced scale synthetic proof of concept of that design. No real facility telemetry was used at any point in this study.

The proof of concept’s own results are modest and are reported as such, where across three illustrative SKU families, a reduced scale constrained actor critic agent did not match the annual cost of either a grid optimised closed form NIDEOQ policy or a grid optimised (s, S) heuristic, though it achieved somewhat higher fill rates and lower, more variable emissions through adaptive setpoint choice. A constrained versus unconstrained ablation did not cleanly separate the two training regimes. We regard this as informative negative and mixed evidence about the network capacity, training budget and calibration data the full target design is likely to require, not as a validation of the framework, and Section 7.5 states explicitly which parts of the target design remain unbuilt.

Two findings seem worth carrying forward independent of the learning result. First, the grid optimised static baselines already show that treating the chamber setpoint as a genuine design variable, rather than a fixed convention, changes the economically optimal design and does so differently across SKU families with different deterioration speeds (Section 7.1), a purely analytical point, requiring no learning at all, that is nonetheless absent from most incumbent practice per the motivation of Section 1.1. Second, at the illustrative allowance price used here, all three evaluated policies find it cheaper to run cold enough to control spoilage and pay for carbon allowances above the free allocation than to run warm enough to comply, illustrating, in a synthetic setting, the spoilage versus carbon price interaction the paper is motivated by (Section 7.2).

The path to the target design described in Sections 4 to 5.5 is now considerably more concrete than before this proof of concept was built, including (i) actually train the physics informed neural surrogate of Section 5.1 against real or higher fidelity synthetic cohort data and verify its extrapolation behaviour, rather than asserting it, (ii) scale the actor critic’s network capacity and training budget, and consider an off policy method with a replay buffer, to test whether the cost gap of **Table 5** closes, (iii) reintroduce the second (service level) constraint and the learned constraint critics of Section 5.2, and rerun the ablation of Section 7.3 to see whether a more expressive constraint mechanism separates constrained from unconstrained training more cleanly, (iv) calibrate kinetics and economics against real facility telemetry and accelerated shelf life trial data, replacing the illustrative parameters of **Table 3**, and (v) only then, a controlled field trial with a mechanistic fallback and formal safety monitoring, which remains the necessary bridge from a synthetic replay, reduced scale or otherwise, to deployed infrastructure.

Author contributions

Conceptualization, AKR and SD; methodology, AKR, SD; software, AKR; validation, AKR, SD and MK; formal analysis, AKR; writing—original draft preparation, AKR and SD; writing—review and editing, AKR; visualization, AKR; supervision, SKS. All authors have read and agreed to the published version of the manuscript.

Data availability

This study uses no real facility, telemetry or commodity data of any kind. All data are synthetic, generated by the scenario generator of Section 6.2 from the seeds and parameters stated in Sections 6.1-6.4, which are sufficient to regenerate every number in this paper exactly. The simulation code, trained policy parameters and raw evaluation results (including the per-seed numbers underlying **Tables 4** and **5** and the ablation of Section 7.3) are available from the corresponding author upon reasonable request.

Code availability

The reference implementation (kinetics model, scenario generator, day stepped simulator, baseline grid search, and actor critic training and evaluation code) is available from the corresponding author upon reasonable request.

Ethical approval

Not applicable.

Informed consent statement

Not applicable.

Conflict of interest

The authors declare no conflict of interest.

References

1. Ghare PM, Schrader GF. A model for exponentially decaying inventory. *Journal of Industrial Engineering* 1963; 14(5): 238–243.
2. Bakker M, Riezebos J, Teunter RH. Review of inventory systems with deterioration since 2001. *European Journal of Operational Research* 2012; 221(2): 275–284, 2012.
3. Janssen L, Claus T, Sauer J. Literature review of deteriorating inventory models by key topics from 2012 to 2015. *International Journal of Production Economics* 2016; 182: 86–112.
4. Wu KS, Ouyang LY, Yang CT. An optimal replenishment policy for noninstantaneous deteriorating items with stockdependent demand and partial backlogging. *International Journal of Production Economics* 2006; 101(2): 369–384.
5. Ouyang LY, Wu KS, Yang CT. A study on an inventory model for noninstantaneous deteriorating items with permissible delay in payments. *Computers & Industrial Engineering* 2006; 51(4): 637–651.
6. Sahoo SS, Routray AK, Dehury S, et al. An inventory model for noninstantaneously reasoned, continuously deteriorating items under a linear demand rate. *Applied Mathematics on Science and Engineering* 2026; 3(2): 11–22. doi: 10.65036/amse322.
7. Dehury S, Sahoo SS, Routray AK, et al. A pseudoconvex fuzzy EOQ model for deteriorating items with time varying demand. *Asian Research Journal of Mathematics* 2026; 22(7): 219–238. doi: 10.9734/arjom/2026/v22i71127.
8. Routray AK, Dehury S, Sahu SK, Threewarehouse inventory model for noninstantaneously deteriorating items with ramptype demand, carbon cap and trade, exponential backorder and twopart trade credit under inflation. *Asian Journal of Probability and Statistics* 2026; 28(6): 43–64. doi: 10.9734/ajpas/2026/v28i6905.
9. Li G, He X, Zhou J, et al. Pricing, replenishment and preservation technology investment decisions for noninstantaneous deteriorating items. *Omega* 2019; 84: 114–126.
10. Yang CT, Dye CY, Ding JF. Optimal dynamic trade credit and preservation technology allocation for a deteriorating inventory model. *Computers & Industrial Engineering* 2015; 87: 356–369.
11. Khanna A, Yadav S, Jaggi CK, Effect of carbontax and capandtrade mechanism on an inventory system with pricesensitive demand and preservation technology investment. *Yugoslav Journal of Operations Research* 2020; 30(3): 361–380.
12. Lok Y, Supadi SS, Wong K. Optimal investment in preservation technology for noninstantaneous deteriorating items under carbon emissions consideration. *Computers & Industrial Engineering* 2026; 183: 109511.
13. Oroojlooyjadid A, Nazari M, Snyder LV, et al. A deep Qnetwork for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management* 2022; 24(1): 285–304.
14. Geevers K, van Hezewijk L, Mes MRK. Multiechelon inventory optimization using deep reinforcement learning. *Central European Journal of Operations Research* 2024; 32: 653–683.
15. Kara A, Dogan I. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. *Expert Systems with Applications* 2018; 91: 150–158.
16. De Moor BJ, Gijsbrechts J, Boute RN. Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. *European Journal of Operational Research* 2022; 301(2): 535–545.
17. Wang K, Long Y, Deng DJ, et al. Singlesite perishable inventory management under uncertainties: A deep reinforcement learning approach. *IEEE Transactions on Knowledge and Data Engineering* 2023; 35(10): 10807–10813.
18. Mohamadi N, Niaki STA, Taher M, et al. An application of deep reinforcement learning and vendormanaged inventory in perishable supply chain management. *Engineering Applications of Artificial Intelligence* 2024; 127:

107403.

19. Wang R, Gan X, Li Q, et al. Solving a joint pricing and inventory control problem for perishables via deep reinforcement learning. *Complexity* 2021; 2021: 6643131.
20. Kavooosi A, TavakkoliMoghaddam R, Sajedi H, et al. Dynamic pricing and inventory control of perishable products by a deep reinforcement learning algorithm. *Expert Systems with Applications* 2025; 291: 128570.
21. Qiao W, Huang M, Gao Z, et al. Distributed dynamic pricing of multiple perishable products using multiagent reinforcement learning. *Expert Systems with Applications* 2024; 237: 121252.
22. Boute RN, Gijsbrechts J, van Jaarsveld W, et al. Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research* 2022; 298(2): 401–412.
23. Gijsbrechts J, R. N. Boute, J. A. Van Mieghem, and D. J. Zhang, “Can deep reinforcement learning improve inventory management? Performance on lost sales, dualsourcing, and multiechelon problems,” *Manufacturing & Service Operations Management*, vol. 24, no. 3, pp. 1349–1368, 2022.
24. Dye CY. The effect of preservation technology investment on a noninstantaneous deteriorating inventory model. *Omega* 2013; 41(5): 872–880.
25. Benjaafar S, Li Y, Daskin M. Carbon footprint and the management of supply chains: Insights from simple models. *IEEE Transactions on Automation Science and Engineering* 2013. 10(1): 99–116.
26. Hovelaque V, Bironneau L. The carbonconstrained EOQ model with carbon emission dependent demand. *International Journal of Production Economics* 2015; 164: 285–291.
27. Toptal A, Özlü H, Konur D. Joint decisions on inventory replenishment and emission reduction investment under different emission regulations. *International Journal of Production Research* 2014; 52(1): 243–269.
28. Gautam P, Maheshwari S, Kausar A. Sustainable retail model with preservation technology investment to moderate deterioration with environmental deliberations. *Journal of Cleaner Production* 2023; 390: 136128.
29. Nomura Y, Liu Z, Nishi T. Deep reinforcement learning for dynamic pricing and ordering policies in perishable inventory management. *Applied Sciences* 2025; 15(5): 2421.
30. Hubbs CD, Li C, Sahinidis NV, et al. A deep reinforcement learning approach for chemical production scheduling. *Computers & Chemical Engineering* 2020; 141: 106982.
31. Altman E. *Constrained Markov Decision Processes*. Boca Raton, FL: Chapman & Hall/CRC; 1999.
32. Achiam J, Held D, Tamar A, et al. Constrained policy optimization. In: *Proceedings of the 34th International Conference Machine Learning (ICML)*; 2017. pp. 22–31.
33. Ray A, Achiam J, Amodei D. Benchmarking safe exploration in deep reinforcement learning. *OpenAI Technical Report* 2019.
34. Haarnoja T, Zhou A, Abbeel P, et al. Soft actorcritic: Offpolicy maximum entropy deep reinforcement learning with a stochastic actor,” in *Proc. 35th Int. Conf. Machine Learning (ICML)*, 2018, pp. 1861–1870.
35. Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actorcritic methods. In: *Proceedings of the 35th International Conference Machine Learning (ICML)*; 2018. pp. 1587–1596.
36. Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. *arXiv:1707.06347* 2017.
37. Raissi M, Perdikaris P, Karniadakis GE. Physicsinformed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 2019; 378: 686–707.
38. Karniadakis GE, Kevrekidis IG, Lu L, et al. Physicsinformed machine learning. *Nature Reviews Physics* 2021; 3: 422–440.
39. Defraeye T, Tagliavini G, Wu W, et al. Digital twins probe into food cooling and biochemical quality changes for reducing losses in refrigerated supply chains. *Resources, Conservation and Recycling* 2019; 149: 778–794.
40. Shrivastava C, Schudel C, Shoji K, et al. Digital twins enable the quantification of the tradeoffs in maintaining citrus quality and marketability in the refrigerated supply chain. *Nature Food* 2022; 3: 413–425.
41. Verboven P, Defraeye T, Datta AK, et al. Digital twins of food process operations: The next step for food process models? *Current Opinion in Food Science* 2020; 35: 79–87.
42. Maheshwari S, Gupta M, Sharma R, et al. Digital twindriven realtime planning, monitoring, and controlling in food supply chains. *Technological Forecasting and Social Change* 2023;195: 122799.
43. Kritzing W, Karner M, Traar G, et al. Digital twin in manufacturing: A categorical literature review and classification. *IFAC PapersOnLine* 2018; 51(11): 1016–1022.
44. Kannapinn M, Schäfer M, Weeger O. Digital twins for autonomous thermal food processing: A model predictive control study with reducedorder models of augmented neural ordinary differential equation type. *Journal of Food Engineering* 2026; 400: 112694.
45. Labuza TP. *ShelfLife Dating of Foods*. Westport, CT: Food & Nutrition Press; 1982.
46. Taoukis PS, Labuza TP. Applicability of timetemperature indicators as shelf life monitors of food products. *Journal of Food Science* 1989; 54(4): 783–788.