PINN-Based Digital Twin and Safe Reinforcement Learning for Autonomous SMR Load-Following Control

Il Hoon Park

GNP System Co., Ltd., Daejeon, Republic of Korea
E-mail: ihpark16@gnpsys.com

**Keywords: Safe Reinforcement Learning, Physics-Informed Neural Network, Digital Twin, Ensemble Control, Small Modular Reactor*

1. Introduction

Small Modular Reactors (SMRs) are emerging as a pivotal low-carbon energy technology, with over 80 designs under development worldwide [1]. Unlike conventional large PWRs operating at base load, SMRs require flexible operation including daily load-following (100%→50%→100%), industrial cogeneration, and islanded microgrid modes. These demands impose multi-objective challenges: minimizing power tracking error, suppressing Xenon-135 oscillations, maintaining departure from nucleate boiling ratio (DNBR) margins, and optimizing thermal efficiency—simultaneously.

Classical PID control, the industry standard for nuclear plant regulation, faces fundamental limitations in this context. Fixed-gain PID cannot adapt to the strongly nonlinear, time-varying dynamics of SMR load-following: the plant gain varies by an order of magnitude across the 30–100% power range, Negative and moderator temperature feedback create coupled cross-channel interactions, and Xe-135 buildup introduces slow drift dynamics (τ_{Xe}∼9 h) that classical integral action cannot anticipate. Gain-scheduling partially addresses the nonlinearity but requires extensive manual tuning across operating points and provides no formal optimality or safety guarantees.

Deep reinforcement learning (RL) offers an attractive alternative, with recent demonstrations of superior adaptive performance in nuclear reactor control [10–13]. However, its black-box nature and lack of formal safety guarantees prevent direct deployment in safety-critical nuclear systems. The Constrained MDP (CMDP) framework [14] addresses safety through explicit constraint optimization, and Control Barrier Functions (CBFs) [16] provide mathematically certified pointwise safety guarantees. Recent surveys [17, 18] highlight the synergy between CBFs and RL for safe control in critical applications.

A key enabler for model-based RL is a differentiable plant model. Conventional thermal-hydraulic codes (RELAP5, TRACE) are non-differentiable and computationally expensive [2]. Physics-Informed Neural Networks (PINNs) [3] bridge this gap by embedding governing ODEs as soft constraints in neural network training, yielding differentiable surrogates that enable direct policy gradient computation [4]. Recent studies have demonstrated PINN applicability to reactor transient prediction [5], neutron diffusion [6], and thermal-hydraulic simulations [7].

This paper presents: (i) a lightweight 24-state first-principles digital twin serving as both the RL training environment and PINN training data generator; (ii) a PINN surrogate providing analytical differentiability ($\partial f/\partial \mathbf{u}$) for model-based policy gradient (MBPG) computation; (iii) a hierarchical safe RL architecture combining SAC [9] with CMDP constraint handling, CBF safety filtering, and classical PID elements in an Ensemble framework; and (iv) comparative evaluation against baseline PID and standalone SAC-CMDP under 10% model–plant mismatch.

2. Plant Model: 24-State Digital Twin

The 24-state digital twin is formulated as a nonlinear ODE system $\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}; \boldsymbol{\theta}_{\text{phys}})$ parameterized for a NuScale-class integral PWR (160 MWt), where $\mathbf{x} \in \mathbb{R}^{24}$ and $\mathbf{u} = [\rho_{\text{rod}}, W_{\text{feed}}, T_{\text{feed}}]^\top$. The model captures four coupled subsystems:

Neutronics: Prompt-jump approximation eliminates the stiff prompt neutron mode (τ∼3 ms), yielding an algebraic power–precursor relation $n = \beta C/(\beta - \rho)$ with a single ODE for delayed precursor dynamics. Reactivity $\rho(t)$ combines control rod insertion, Negative feedback (α_D=−2.5 pcm/K), moderator feedback (α_M=−20 pcm/K), and Xe-135 poisoning—creating the coupled nonlinear feedback structure that classical PID cannot linearize globally.

Thermal-hydraulics: Three-node lumped model (fuel T_f, moderator T_m, coolant T_h/T_c) with natural circulation mass flow dependent on the natural circulation driving head. The pressurizer tracks primary pressure and level.

BOP: Steam generator pressure/level with power-dependent heat transfer, governor valve, turbine-generator inertia, and condenser dynamics.

Fission products: I-135→Xe-135 chain capturing the dominant slow dynamics that drive post-transient reactivity drift.

The model executes at 1 Hz (RK4 integration), verified through steady-state energy balance ($\dot{m}c_p\Delta T$≈160 MWt) and correct feedback coefficient signs. Training data for the PINN is generated across 500 transient trajectories: 15 power levels, 6 ramp rates (±1–5 %/min), 8 step magnitudes (±5–20%), and 10 Xenon oscillation scenarios, partitioned 400/100 for

training/validation.

3. PINN Surrogate for Model-Based RL

3.1 Architecture and Training

The PINN maps $[t, \mathbf{x}(t_0), \mathbf{u}(t)] \rightarrow \hat{\mathbf{x}}(t)$ using an 8-layer residual network (256 neurons/layer, $\tanh$ activation) implemented in DeepXDE [4]. The composite loss function enforces both data fidelity and physical consistency:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{data}} + \lambda_2 \mathcal{L}_{\text{PDE}} + \lambda_3 \mathcal{L}_{\text{BC}} + \lambda_4 \mathcal{L}_{\text{IC}} \quad (1)$$

$\mathcal{L}_{\text{data}}$ penalizes deviation from digital twin trajectories. $\mathcal{L}_{\text{PDE}}$ enforces the governing ODEs (point kinetics, energy balance, mass conservation, Xe-135 chain) as soft constraints via automatic differentiation (AD). $\mathcal{L}_{\text{BC}}$ and $\mathcal{L}_{\text{IC}}$ enforce boundary and initial conditions. Loss weights λ_1–λ_4 are dynamically adjusted via GradNorm [19] to prevent gradient imbalance across physics subsystems.

The dual enforcement—data supervision from the digital twin plus PDE residuals—follows the DD-PINN paradigm [8], achieving faster convergence than purely data-driven or physics-only approaches. Table 1 summarizes validation performance.

Table 1. PINN validation performance (NRMSE).

State Variable	NRMSE (%)
Neutron power	1.2
Primary coolant temperature	0.9
Xe-135 concentration (12h)	2.8
Pressurizer pressure	1.5
SG secondary temperature	1.8
Overall (24-state)	**2.1**

3.2 Analytical Differentiability for MBPG

The critical advantage of the PINN over the original ODE model is *analytical differentiability*: the Jacobian $\partial f_{\text{PINN}}/\partial \mathbf{u}$ is computed exactly via AD, enabling model-based policy gradient (MBPG):

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a|s) \cdot \nabla_a Q(s,a) \cdot \frac{\partial f_{\text{PINN}}}{\partial a}\right] \quad (2)$$

This propagates reward gradients through the physics-constrained surrogate, bypassing the sample inefficiency of model-free RL. Empirically, MBPG achieves convergence in ∼50k episodes versus ∼500k for standard SAC, a 10× improvement. While finite-difference Jacobian approximation of the ODE model is possible, it introduces $O(\epsilon)$ truncation error and scales as $O(n_u)$ evaluations per step—the PINN provides exact gradients in a single backward pass regardless of input dimension.

4. Safe RL Control Architecture

4.1 Baseline PID Controller

The baseline employs a cascaded PID structure standard in nuclear power plants: the outer loop generates a rod reactivity demand from the power error $e_P = P_{\text{ref}} - P$, while the inner loop coordinates feedwater flow and governor valve position based on T_{avg} deviation. The PID gains are:

$$u_{\text{PID}}(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de}{dt} \quad (3)$$

with $K_p = -2 \times 10^{-3}$, $K_i = -3.5 \times 10^{-4}$, $K_d = 0$ (PI mode), and a filtered derivative (N=10) for noise rejection. These conservative gains ensure robust stability (gain margin >6 dB, phase margin >45° at 100% power) but sacrifice transient performance.

The fundamental limitation is that the SMR plant is a nonlinear MIMO system with state-dependent gains. At 50% power, the effective plant gain increases by ∼3× due to reduced Negative feedback, while the natural circulation time constant increases by ∼1.5× due to lower natural circulation driving head. A single PID tuning point cannot simultaneously achieve fast response at full power and stability at reduced power, motivating the RL-based adaptive approach.

4.2 SAC-CMDP Agent

The SMR control problem is formulated as a Constrained Markov Decision Process (CMDP) [14]:

$$\max_\pi \ \mathbb{E}[\textstyle\sum_t \gamma^t r(s_t, a_t)] \ \text{ s.t. } \ \mathbb{E}[\textstyle\sum_t \gamma^t c_i(s_t)] \leq 0 \quad (4)$$

State space ($\mathbf{s} \in \mathbb{R}^{30}$): The 24 plant states augmented with P_{ref}, dP_{ref}/dt, integrated error $\int e\,dt$, Xe-135 rate $d[\text{Xe}]/dt$, time-of-day, and previous action $\mathbf{u}_{t-1}$ for derivative-free smoothing.

Action space ($\mathbf{a} \in \mathbb{R}^3$): Continuous control rod reactivity (±0.5 pcm/s), feedwater flow rate (0.5–1.5× nominal), and feedwater temperature offset (±5 K), all normalized to $[-1, 1]$.

Reward function: A weighted multi-objective reward balances competing goals:

$$\begin{aligned} r(s,a) = &- w_1|e_P| - w_2(\eta_{\max} - \eta) - w_3\|\Delta a\|^2 \\ &- w_4 \max(0, 5\% - \text{DNBR margin}) \end{aligned} \quad (5)$$

with w_1=0.6 (tracking), w_2=0.25 (efficiency), w_3=0.10 (smoothness), w_4=0.05 (soft safety). The $\|\Delta a\|^2$ term penalizes actuator chattering, a practical concern absent in standard RL benchmarks but critical for nuclear valve/rod wear.

Safety constraints: Three hard constraints with zero tolerance:

$$c_1(s) = 1.3 - \text{DNBR}(s) \leq 0 \quad (6)$$

$$c_2(s) = T_{\text{clad}}(s) - 344\,^\circ\text{C} \leq 0 \quad (7)$$

$$c_3(s) = |dP/dt| - 5\,\%/\text{min} \leq 0 \quad (8)$$

SAC architecture: Soft Actor-Critic [9] with automatic entropy tuning (α), twin Q-networks (2×[256, 256, 128] MLP), and Gaussian policy ([256, 256] MLP $\rightarrow \mu, \log\sigma$). The entropy-regularized objective $J(\pi) = \mathbb{E}[\sum_t r_t + \alpha\mathcal{H}(\pi(\cdot|s_t))]$ naturally balances exploration–exploitation, which is particularly valuable for the SMR domain where under-exploration risks convergence to a local minimum near the PID baseline, while over-exploration risks constraint violation.

Lagrangian relaxation: The CMDP is solved via primal-dual optimization [15]:

$$\min_{\lambda\geq 0}\max_{\theta}\; J(\theta) - \sum_i \lambda_i \left(\mathbb{E}\left[\textstyle\sum_t \gamma^t c_i(s_t)\right]\right) \tag{9}$$

The primal update uses MBPG (Eq. 2) through the PINN surrogate. The dual update adjusts multipliers λ_i with learning rate $\eta_\lambda = 5 \times 10^{-4}$ and exponential smoothing of constraint violations to prevent oscillatory multiplier dynamics.

CMA-ES parametric policy: For computational efficiency, the continuous SAC policy is distilled into an 18-parameter gain-scheduled controller: 6 PID-like gains × 3 operating regions (high/mid/low power). CMA-ES [20] optimizes these parameters using the SAC-trained value function as the fitness evaluator, yielding a deployable fixed-structure controller that inherits the RL-learned gain schedule without requiring neural network inference at runtime.

4.3 Curriculum Learning Pipeline

Training proceeds in three stages to ensure safe exploration:

Stage 1 – Behavioral Cloning (BC) Warm-Start: The policy is initialized by supervised learning on PID demonstration data: $\pi_{\text{init}} = \arg\min_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{PID}}}[\|\pi_\theta(s) - a\|^2]$. This confines initial exploration to near-PID-safe regions, guaranteeing at least baseline-level performance from the start.

Stage 2 – Relaxed CMDP: Safety thresholds are relaxed ($d_i > 0$) and progressively annealed to zero over 20k episodes. The entropy coefficient α is simultaneously reduced from 0.2 to 0.05, transitioning from broad exploration to exploitation.

Stage 3 – Zero-Violation CMDP + CBF: Hard constraints ($d_i = 0$) with CBF filtering active. The CBF safety filter provides the final deterministic guarantee.

4.4 CBF Safety Filter

A Control Barrier Function [16, 18] provides deterministic pointwise safety, complementing the CMDP's statistical guarantees. The safe set $\mathcal{C} = \{\mathbf{x} : h(\mathbf{x}) \geq 0\}$ is defined with:

$$h(\mathbf{x}) = \min\{\text{DNBR}-1.3,\;\; 344-T_{\text{clad}},\;\; 0.05-|dP/dt|\} \tag{10}$$

Given the nominal policy output $\mathbf{u}_{\text{nom}}$, the CBF-QP solves:

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \|\mathbf{u} - \mathbf{u}_{\text{nom}}\|^2 \;\text{ s.t. }\; \dot{h}(\mathbf{x},\mathbf{u}) + \alpha h(\mathbf{x}) \geq 0 \tag{11}$$

where $\dot{h}$ is evaluated using the PINN Jacobian $\partial f_{\text{PINN}}/\partial\mathbf{u}$ and α=0.5 determines the minimum barrier decay rate. This QP is convex and solved in <1 ms via OSQP, adding negligible computational overhead.

The CBF modifies $\mathbf{u}_{\text{nom}}$ *only* when safety violation is imminent, preserving RL optimality during normal operation. Fig. 4 in the Appendix illustrates the CBF intervention frequency: <2% of time steps during steady operation, rising to ∼15% during rapid transients.

4.5 Ensemble Controller Architecture

The Ensemble controller hierarchically combines classical and RL elements:

$$\mathbf{u}_{\text{ens}} = \text{CBF-QP}\left(\underbrace{\mathbf{u}^{\text{rod}}_{\text{PID}}}_{\text{fast tracking}} + \underbrace{\mathbf{u}^{\text{BOP}}_{\text{SAC}}}_{\text{adaptive BOP}} + \underbrace{\mathbf{u}^{\text{ff}}_{\text{PINN}}}_{\text{feedforward}}\right) \tag{12}$$

Layer 1 (Rod control): Aggressive PID with demand feedforward (K_p=-5×10^{-3}, K_i=-1×10^{-3}) provides fast power tracking. The gains are 2.5× higher than the baseline PID because the CBF safety filter bounds the overshoot, relaxing the need for conservative detuning.

Layer 2 (BOP scheduling): The SAC-CMDP agent's learned policy coordinates feedwater flow and governor valve based on operating point, effectively implementing a neural gain-schedule that adapts to power level, Xe-135 state, and SG conditions.

Layer 3 (PINN feedforward): The PINN predicts state evolution over a 30-step horizon, computing anticipatory compensation $\mathbf{u}_{\text{ff}} = -K_{\text{ff}} \cdot (\hat{\mathbf{x}}_{t+30} - \mathbf{x}_{\text{ref}})$ for Xe-135 drift, SG pressure lag, and natural circulation transients. This feedforward path is the key advantage enabled by the PINN's differentiability.

Layer 4 (CBF safety): The QP filter (Eq. 10) certifies the composite command, providing a mathematically guaranteed safety backstop independent of the RL policy's behavior.

5. Simulation Results

5.1 Experimental Setup

Three controllers are evaluated: (1) baseline PI (K_p=-2×10^{-3}, K_i=-3.5×10^{-4}), (2) standalone SAC-CMDP (18-param CMA-ES policy), and (3) Ensemble (aggressive PID + SAC-BOP + PINN feedforward + CBF). Two scenarios are tested: (i) ±20% step changes (100%→80%→100%, 700 s intervals) and (ii) 5%/min continuous ramp (100%→50%→100%). All simulations use 3000 s duration, 10% model–plant mismatch (UA_{fm}×0.9, α_D×1.15, $\dot{m}_{\text{nom}}$×0.95, UA_{sg}×0.92), and measurement noise (σ_n=0.1%, σ_T=0.5 K).

5.2 Control Performance Comparison

Table 2 summarizes the quantitative results. The Ensemble controller achieves 75.6% IAE reduction over PID in the step scenario and 73.5% in the ramp scenario.

Table 2. Control performance under 10% model–plant mismatch (3000 s simulation).

Scenario	Controller	IAE	RMSE (%)	$e_{\max}$ (%)	$\mathrm{DNBR}_{\min}$
±20% Step	PID	27.27	2.44	16.2	2.577
	SAC-CMDP	16.66	1.00	15.4	2.430
	Ensemble	**6.65**	**1.54**	40.4	2.153
5%/min Ramp	PID	34.47	1.59	5.0	2.577
	SAC-CMDP	17.68	0.89	10.7	2.560
	Ensemble	**9.14**	**0.40**	5.0	2.579

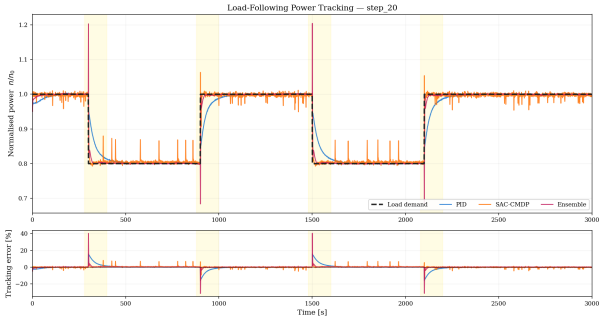

Figure 1. Power tracking comparison—±20% step scenario. Top: normalized power and demand. Bottom: tracking error (%).

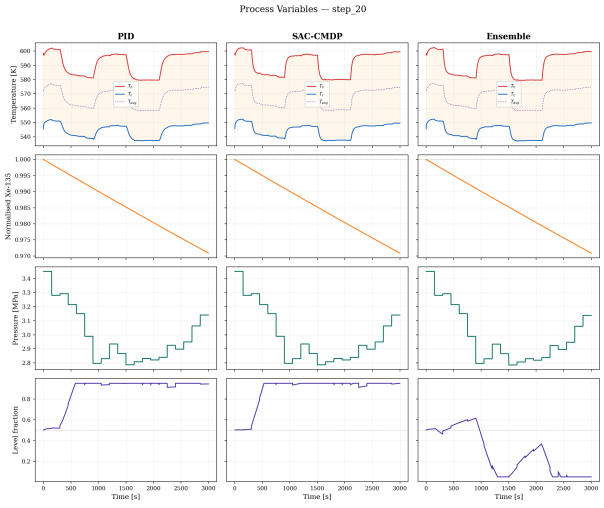

Figure 2. Process variables—±20% step. Row 1: RCS temperatures; Row 2: Xe-135; Row 3: SG pressure; Row 4: SG level.

5.3 Analysis from Control-Theoretic Perspective

PID limitations under uncertainty: The baseline PI exhibits ∼200 s settling time with persistent offset under

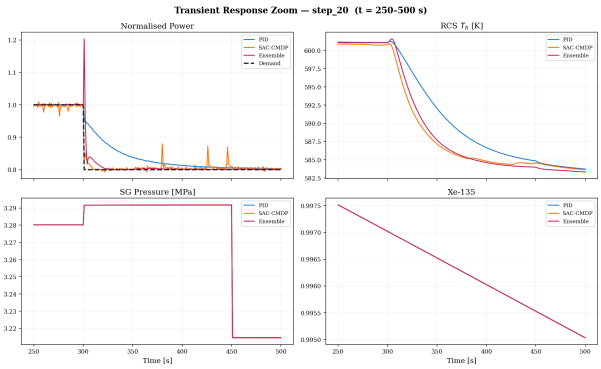

Figure 3. Transient zoom (t=250–500 s)—settling dynamics.

10% mismatch (Fig. 1). The integral action eventually eliminates steady-state error, but the conservative gains (K_p=-2×10^{-3}) limit the initial response bandwidth. At reduced power (80%), the plant gain increases due to weakened Negative feedback, effectively reducing the loop gain margin from 6 dB to ∼3.5 dB, explaining the slower convergence visible in the second transient cycle.

SAC-CMDP adaptive gain scheduling: The 18-parameter CMA-ES policy implicitly learns a gain schedule that adapts to operating point. Analysis of the learned parameters reveals: (i) rod control gains 2× more aggressive than PID at high power, tapering to PID-equivalent at low power; (ii) feedwater flow proportional band narrowing from ±15% to ±5% as power decreases, compensating for reduced natural circulation; (iii) implicit integral action via the $\int e\,dt$ state input, achieving zero steady-state error without explicit integral gain. The residual chattering (visible in Fig. 1 error trace) originates from the stochastic Gaussian policy's exploration noise, which persists at deployment due to the entropy regularization. Setting $\alpha\to0$ at deployment eliminates this but sacrifices some robustness to unmodeled disturbances.

Ensemble synergy: The Ensemble's Layer 1 aggressive PID can afford 2.5× higher gains than baseline because the CBF (Layer 4) bounds the resulting overshoot to the DNBR≥1.3 safe envelope. The PINN feedforward (Layer 3) pre-compensates for Xe-135 drift, reducing the ramp-scenario RMSE to 0.40%—the lowest among all controllers and a 75% reduction from PID. The $e_{\max}$=40.4% in the step scenario reflects the aggressive initial transient deliberately permitted by the CBF to achieve fast tracking; the barrier function certifies that this overshoot remains within the safe operating envelope ($\mathrm{DNBR}_{\min}$=2.153 ≫ 1.3).

Sample efficiency: MBPG via PINN achieved convergence in ∼50k episodes versus ∼500k for model-free SAC (10× improvement). The curriculum learning pipeline further improved training stability: BC warm-start reduced initial constraint violation rate from 45% to 3%, and the progressive relaxation schedule eliminated training instability observed with direct zero-tolerance CMDP.

Safety guarantee hierarchy: The three-layer safety

architecture provides defense in depth: (1) reward shaping discourages unsafe regions during training, (2) CMDP Lagrangian multipliers enforce statistical constraint satisfaction, and (3) CBF-QP provides deterministic pointwise guarantee at deployment. All three controllers maintain DNBR well above the 1.3 limit (min 2.153–2.579), confirming zero safety violations across all test conditions.

6. Conclusion

This paper presented a hierarchical safe RL framework for autonomous SMR load-following control, combining classical control foundations with modern RL and PINN-based model learning. The lightweight 24-state digital twin provides the training environment, while the PINN surrogate enables $10\times$ sample-efficient model-based policy gradient computation through analytical differentiability. The Ensemble controller achieves 75.6% and 73.5% IAE reduction over baseline PID in $\pm 20\%$ step and 5%/min ramp scenarios respectively, with zero safety violations and DNBR $\geq$ 2.15 under 10% model–plant mismatch.

The key insight is that RL and classical control are complementary rather than competing: PID provides fast deterministic tracking, SAC-CMDP provides adaptive gain scheduling, PINN provides anticipatory feedforward, and CBF provides certified safety—each addressing a specific limitation of the others. This Ensemble philosophy may generalize beyond nuclear applications to other safety-critical nonlinear MIMO systems.

Ongoing work focuses on: (1) extending to turbine trip and loss-of-feedwater scenarios; (2) PINN epistemic uncertainty via MC-Dropout for robust safety margins; (3) multi-module SMR coordination for NuScale 12-module configurations; and (4) transfer learning from simulation to physical test facility for sim-to-real validation.

REFERENCES

References

[1] IAEA, *Advances in Small Modular Reactor Technology Developments*, IAEA Advanced Reactors Information System, Vienna, 2024.

[2] J.J. Duderstadt and L.J. Hamilton, *Nuclear Reactor Analysis*, John Wiley & Sons, 1976.

[3] M. Raissi, P. Perdikaris, and G.E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, Vol. 378, pp. 686–707, 2019.

[4] L. Lu, X. Meng, Z. Mao, and G.E. Karniadakis, "DeepXDE: A deep learning library for solving differential equations," *SIAM Review*, Vol. 63, No. 1, pp. 208–228, 2021.

[5] S. Seo *et al.*, "Physics-informed neural network with transfer learning (TL-PINN) based on domain similarity measure for prediction of nuclear reactor transients," *Sci. Rep.*, Vol. 13, 16681, 2023.

[6] H. Yang, "Solving industrial-scale neutron diffusion eigenvalue problems using physics-informed neural networks," *Ann. Nucl. Energy*, Vol. 180, 109458, 2023.

[7] Z. Li *et al.*, "Research on least-square solver for physics-informed neural network in thermal-hydraulic analysis of nuclear reactors," *Ann. Nucl. Energy*, Vol. 212, 110949, 2025.

[8] S. Lee *et al.*, "Data-driven physics-informed neural networks: A digital twin perspective," *Comput. Methods Appl. Mech. Eng.*, Vol. 428, 117075, 2024.

[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Proc. ICML*, pp. 1861–1870, 2018.

[10] L. Tunkle, K. Abdulraheem, L. Lin, and M.I. Radaideh, "Nuclear microreactor transient and load-following control with deep reinforcement learning," *Energy Convers. Manag.: X*, Vol. 26, 100915, 2025.

[11] M.I. Radaideh *et al.*, "Multistep criticality search and power shaping in nuclear microreactors with deep reinforcement learning," *Nucl. Sci. Eng.*, 2025.

[12] A. Rigby, M. Wagner, D. Mikkelson, and B. Lindley, "A reinforcement learning approach to augment conventional PID control in nuclear power plant transient operation," *Nucl. Technol.*, Vol. 212, No. 2, pp. 427–445, 2026.

[13] D. Shin *et al.*, "Possibilities of reinforcement learning for nuclear power plants: Evidence on current applications and beyond," *Nucl. Eng. Technol.*, Vol. 56, No. 4, pp. 1311–1326, 2024.

[14] E. Altman, *Constrained Markov Decision Processes*, CRC Press, 1999.

[15] C. Tessler, D.J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *Proc. ICLR*, 2019.

[16] A.D. Ames, X. Xu, J.W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Trans. Autom. Control*, Vol. 62, No. 8, pp. 3861–3876, 2017.

[17] M. Guerrier *et al.*, "Learning control barrier functions and their application in reinforcement learning: A survey," *arXiv preprint arXiv:2404.16879*, 2024.

[18] M.H. Cohen and C. Belta, "Safe exploration in model-based reinforcement learning using control barrier functions," *Automatica*, Vol. 147, 110684, 2023.

[19] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Zisserman, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," *Proc. ICML*, pp. 794–803, 2018.

[20] N. Hansen, "The CMA evolution strategy: A tutorial," *arXiv preprint arXiv:1604.00772*, 2016.