Energy Engineering

Tech Science Press

http://doi.org/10.32604/ee.2026.087017

ARTICLE

Control-Barrier-Function Shielded Conservative Distributional Reinforcement Learning for Anti-Misoperation AGC of Thermal Power Units

Mujie Zhang[1], Yajun Wu[1], Dongsheng Li[1], Yankai Zhu[2,*] and Zhengyan Zhao[2]

[1]Central China Branch of State Grid Corporation of China, Wuhan, China
[2]State Key Laboratory for Alternate Electrical Power System with Renewable Energy Sources, School of Control and Computer Engineering, North China Electric Power University, Beijing, China
*Corresponding Author: Yankai Zhu. Email: zhuyankai@ncepu.edu.cn

Received: 09 June 2026; Accepted: 28 August 2026; Published: 12 November 2026

ABSTRACT: High shares of variable renewable generation increasingly require thermal units to provide deeper, faster, and more frequent automatic generation control (AGC) while operating close to low-load, steam-pressure, ramp-rate, and actuator limits. Under these conditions, an area-level regulation request that is valid for grid balancing can become unsafe at the plant interface because of nonlinear boiler–turbine dynamics, delayed measurements, sensor bias, and compound disturbances. This study therefore formulates AGC from the plant side: an area-level dispatcher allocates regulation among thermal generation, hydro generation, battery storage, and flexible load, while the proposed controller modifies only the command assigned to a 600 MW reheat thermal unit before it enters the plant distributed control system. A control-barrier-function (CBF) shielded conservative distributional reinforcement learning method is developed to prevent excessive ramping, steam-pressure excursions, valve over-actuation, low-load instability, and errors caused by delayed or biased measurements. Every transition used for training, validation, and testing is generated numerically by the declared behavior policy and benchmark generator; no measured AGC record or plant-historian data enter the reported results. Asynchronous signals are causally aligned on a 4 s decision grid, and a dedicated distributional safety-cost critic estimates the 0.95 conditional value-at-risk (CVaR) of discounted constraint cost. The nonlinear simulator is additionally cross-checked against the published plant-derived Bell–Åström model at seven operating points. Normalized power-transient NRMSE is 0.021–0.079 with correlation $r \geq 0.995$, while pressure-transient NRMSE is 0.084–0.254 with $r = 0.733$–0.970. Ten independent training seeds and 3500 held-out numerical episodes show that the controller reduces cumulative absolute ACE from 218 to 165 MW min and improves the frequency nadir from −0.073 to −0.041 Hz relative to PI AGC. No hard violation occurs in 525,000 feasible test decisions, giving a one-sided 95% binomial upper bound of 0.00057% on the per-decision violation probability. The safety statement remains conditional on the declared plant model, uncertainty set, robust control-invariant subset, and bounded prediction errors. End-to-end CPU execution requires 6.8 ms on average and 9.6 ms at the 95th percentile.

KEYWORDS: Automatic generation control; thermal power unit; safe reinforcement learning; control barrier function; conservative offline learning; conditional value-at-risk; anti-misoperation

1 Safety-Critical Thermal AGC under Renewable-Dominant Operation

The accelerating integration of wind, photovoltaic, storage, hydrogen, and flexible demand is changing the physical role of thermal generation in interconnected power systems. Thermal units were historically scheduled as dispatchable, comparatively smooth resources, and automatic generation control

Copyright © 2027 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(AGC) mainly corrected residual area control error (ACE) around an operating point. In systems with large renewable fluctuations, the same units are increasingly required to provide frequent, deep, and fast regulation while operating near low-load limits. This shift turns the AGC loop into a safety-critical coordination layer rather than a purely economic or tracking mechanism. The conventional separation between dispatch, governor control, boiler-turbine dynamics, actuator protection, and operator intervention is no longer sufficient when the unit must repeatedly move across nonlinear operating regions within short time intervals. Classical frequency-control texts describe the electromechanical basis of frequency response and inter-area control, but the emerging regime adds data-driven uncertainty, cyber-physical delays, and rare-event safety requirements that are only partially represented in traditional deterministic models [1,2].

Misoperation risk in thermal AGC is not a single fault class. It is a compound event that arises when an externally valid regulation command becomes internally unsafe after it interacts with actuator limits, delayed measurements, boiler storage, steam pressure, valve travel, low-load combustion stability, or biased sensor residuals. A command can be acceptable under the network-level balancing objective but still dangerous for the plant because the turbine valve, reheater, mill, or boiler dynamics cannot realize the requested ramp without crossing operational envelopes. Conversely, an overly conservative plant-level limiter can protect equipment but degrade grid frequency regulation. This study treats anti-misoperation AGC as a joint learning and safety-filtering problem: the controller must learn when corrective action improves tracking and risk performance, while an explicit constraint layer must prevent the learned policy from executing commands that violate plant-level safety. This distinction is important because safe operation cannot be reduced to a reward penalty when rare violations are high impact and operationally unacceptable.

The control boundary is plant-side rather than grid-side. The system operator calculates area control error and issues an area-level secondary-regulation request. Participation factors then distribute that request among several regulating resources. The proposed controller receives only the portion assigned to the studied 600 MW thermal unit and returns a feasible plant command; it does not recalculate system-wide ACE, redispatch the remaining resources, or replace the governor, boiler master, turbine master, or protection system. Grid frequency, tie-line exchange, and the responses of hydro, battery, flexible-load, and other thermal resources remain in the simulation environment because they determine the command and disturbance seen by the target unit. This boundary preserves the practical distinction between area-level coordination and unit-level command execution.

Recent multi-energy operation studies show that flexible resources are now coordinated across electricity, heat, transport, and hydrogen vectors rather than dispatched as isolated assets. Temporally decomposed stochastic programming for airport microgrids, for example, emphasizes climatic correlation, flexible loads, and multi-energy coupling across operational horizons [3]. Carbon-oriented pricing in two-layer multi-energy scheduling further demonstrates that operational signals can reshape electricity-hydrogen-transport decisions under uncertainty [4]. These studies motivate a broader system context for thermal AGC: even when the CBF-shielded controller is deployed inside a thermal unit, its disturbances and set points are generated by a coupled grid in which renewable power, flexible load, and market signals create nonstationary operating conditions. The reported controller is nevertheless trained and evaluated only on the fully numerical benchmark defined in Sections 3 and 6; no historical baseload record is used.

Learning-based AGC is attractive because it can map high-dimensional operation histories to corrective commands without requiring an exact analytical model for every boiler-turbine subsystem. Reinforcement learning (RL) is particularly relevant because AGC is sequential: an aggressive action that improves ACE in one interval can produce pressure stress, valve saturation, or degraded ramp capability in later intervals. RL has been widely used for sequential decision-making and has matured through deep

value approximation, actor-critic control, and entropy-regularized optimization [5–7]. Power-system researchers have also surveyed RL for decision and control tasks, including generation dispatch, demand response, voltage regulation, and system restoration [8,9]. However, direct online exploration in a thermal plant is not acceptable. Unsafe exploration can violate hard constraints, while a reward-only policy can learn high average performance at the cost of rare but serious misoperation events.

The safety problem has two layers. The first is statistical: the trained policy avoids actions in state-action regions unsupported by its fixed offline dataset. Offline RL addresses this issue by learning without online exploration, but it can extrapolate optimistic values for actions outside the behavior distribution. Conservative Q-learning reduces this failure mode by lowering the value of out-of-distribution actions and is therefore aligned with a numerically generated behavior dataset in which the engineering limiter admits only a narrow subset of actions [10,11]. The second layer is formal: even a conservative learned policy can produce an unsafe action under rare combinations of low load, delayed communication, abnormal sensor residuals, or rapidly changing AGC commands. Control barrier functions (CBFs) provide a mechanism for enforcing forward invariance of a safe set through a real-time inequality constraint and are commonly implemented through quadratic programs that minimally modify a nominal control input [12]. The framework combines these two layers rather than relying on either alone.

Tail risk is another reason that standard expected-reward RL is inadequate. Misoperation probability is often low in the training data, yet its operational impact is severe. Conditional value-at-risk (CVaR) is a coherent risk measure that focuses on the expected magnitude of the worst outcomes beyond a specified quantile, and it is well suited for representing rare high-impact AGC events [13,14]. Distributional RL provides a natural way to estimate return distributions rather than only expected returns; it allows the controller to distinguish an action with stable moderate performance from an action with similar mean but dangerous lower-tail behavior [15,16]. The framework uses a conservative distributional critic to estimate the risk distribution of corrective AGC actions, then embeds CVaR in a constrained objective.

This study makes four contributions. First, it defines a plant-side constrained Markov decision process within a multi-resource AGC environment and specifies the frequency, tie-line, command-allocation, actuator, boiler-turbine, sensing, and delay equations needed to reconstruct the transition model. Second, it trains separate distributional critics for reward return and cumulative safety-cost return; the upper tail of the latter gives an unambiguous 0.95-CVaR safety signal, while conservative regularization suppresses unsupported actions. Third, it derives a robust discrete-time CBF projection with explicit model-error and linearization-remainder margins. The execution problem contains no safety slack, and infeasibility invokes a verified fallback instead of sending a relaxed command. Fourth, it evaluates all controllers with matched architectures, data, update budgets, ten independent seeds, held-out mixed-stress episodes, uncertainty intervals, and traceable denominators for every violation statistic. These elements turn the complete controller, rather than any isolated component, into an auditable unit of analysis.

The adopted perspective is intentionally narrower than broad grid resilience and wider than conventional turbine control. It is narrower because the study does not attempt to co-optimize unit commitment, market clearing, reserve procurement, and transmission security. Those problems determine the operating condition under which AGC is requested, but they are not the execution problem faced by the thermal unit every few seconds. It is wider because the study does not view the AGC command as a scalar tracking reference alone. The same scalar command is interpreted through the current output, the recent ramp history, the thermal storage condition, the available valve travel, the quality of measurements, and the delay state of the communication channel. A safe command at medium load with healthy pressure margin becomes unsafe at deep peak load after several consecutive upward ramps. This context dependence is pre-

cisely why a static limiter is insufficient and why a data-driven policy with a plant-aware safety interface is attractive.

A second motivation is the support mismatch that arises in any fixed offline dataset. In the present study that dataset is entirely numerical: ordinary trajectories are generated by the behavior policy in Section 6, and rare stress trajectories are generated by the same declared plant, disturbance, sensing, and delay equations. The term "historical operation" is used only when discussing the general motivation for offline RL and never denotes the provenance of a reported transition. A naive learner can still assign excessive value to an unobserved correction because its consequence is absent from the fixed numerical dataset. Conservative value estimation therefore reduces optimism outside behavior-policy support, while numerically generated stress cases expose the critic to states near the safety boundary.

A third motivation is the need for transparent intervention diagnostics. Operators are unlikely to trust an opaque AGC controller that occasionally refuses dispatch commands without explaining whether the issue is ramping, pressure, valve motion, sensor residuals, or aggregate risk. The CBF projection creates interpretable shadow information: the active constraints identify which safety boundary caused the modification, the shield deviation quantifies how far the learned policy was from feasible execution, and the slack variables indicate whether the chosen envelope was temporarily infeasible. These diagnostics are important for engineering acceptance because they convert a black-box learning recommendation into a command accompanied by a safety certificate at each decision step.

The evidence reported here concerns the implemented two-area, multi-resource simulation and the stated parameter and disturbance ranges. The nonlinear boiler-turbine simulator used to assess executed trajectories is independent of the reduced affine model used by the CBF, which prevents safety evaluation from being a direct restatement of the shield constraints. Its normalized power and pressure transients are also compared with the independently published, plant-derived Bell–Åström benchmark in a dedicated external cross-model validation. This is an external model check rather than hardware-in-the-loop or experimental validation. The results do not constitute a field-commissioning certificate; plant deployment additionally requires manufacturer-model validation, hardware-in-the-loop tests, cybersecurity review, operator approval, and staged shadow operation. This scope is maintained consistently in the results and conclusions.

2 Research Context and Technical Gap

Classical AGC and load-frequency control rely on structured models that connect generator inertia, governor response, turbine dynamics, tie-line exchange, and area control error. These models remain essential because they encode physical conservation laws and operational conventions. In thermal plants, however, the grid-facing AGC signal is filtered through a complex plant response involving steam storage, boiler pressure, valve rate limits, reheater dynamics, feedwater control, and combustion constraints. Standard generation-operation references describe how dispatch and reserve decisions interface with unit constraints, but the modern requirement for deep and frequent regulation makes the interface more dynamic and less forgiving [17,18]. The literature on robust frequency control has addressed parameter uncertainty and load disturbance rejection, yet it generally assumes controller structures designed before deployment rather than policies learned from high-dimensional operational data.

RL extends this controller design space by optimizing sequential decisions through interaction data. Deep Q-networks showed that nonlinear value functions can learn policies from high-dimensional observations [6], while continuous-control methods such as deterministic policy gradients and soft actor-critic made action-space control more practical for energy systems [7,19]. Demand-response surveys and power-

system RL reviews show that energy applications often require constrained actions, delayed rewards, partial observability, and domain-specific safety layers [8,20]. For AGC, these issues are amplified because an unsafe action can immediately affect frequency response and equipment health. Thus, the literature suggests that RL can improve adaptivity but cannot be deployed without explicit safety design.

Safe RL has developed along three main directions: constrained policy optimization, risk-sensitive objectives, and external safety filters. Constrained policy optimization formalizes expected constraint costs, but it still depends on accurate constraint estimation during learning [21]. General safe RL surveys emphasize that safety can be pursued through constrained exploration, robust policy design, teacher intervention, or post-hoc shielding; each approach trades learning flexibility against formal guarantees [22]. Risk-constrained RL introduces percentile or CVaR criteria so that the policy accounts for distribution tails rather than only the mean [23]. These ideas are conceptually close to thermal AGC because misoperation events are not frequent enough to dominate average rewards but are serious enough to dominate operational acceptability. Nevertheless, safe RL methods designed for generic benchmarks do not automatically capture boiler-turbine feasibility, actuator saturation, or ACE-specific performance metrics.

Recent literature sharpens this implementation boundary. A 2025 review distinguishes power-system safe-RL methods that use explicit safety layers from methods that integrate constraints into policy optimization, and identifies frequency regulation, voltage control, and energy management as the principal application families [24]. Broader reviews of safe learning and safe RL emphasize uncertainty-aware certificates, intervention logic, and verifiable constraint handling rather than reward shaping alone [25,26]. Recent power-system applications accordingly embed operational constraints in distribution-network control [27,28], real-time energy-hub control [29], and decentralized frequency control with stability-aware policy structures [30,31]. Coordinated frequency control and zero-violation constrained policy optimization further show the importance of declaring the constraint denominator and the mechanism used to prevent unsafe actions [32,33].

The same trend appears in physically informed safety layers for multi-energy microgrids [34], safe multi-agent bidding and maintenance scheduling [35], delay-aware decentralized inverter control [36], barrier-based regulation-service control [37], adaptive hard-constraint layers for multi-energy management [38], and hard-barrier constrained optimization in unknown stochastic environments [39]. Stability- and safety-guaranteed transient frequency control [40] and constraint-penalized offline Q-learning [41] extend these ideas to distributed dynamics and fixed datasets. Collectively, these studies establish a current benchmark: learned policies should be paired with explicit feasibility machinery and evaluated under distribution shift. They do not, however, jointly address plant-side thermal AGC, conservative offline support control, a distributional safety-cost critic, CVaR tail optimization, and a robust CBF projection of the executable command.

Distributional RL is particularly relevant for anti-misoperation control. In an expected-value critic, two corrective commands can be indistinguishable even if one has a small probability of driving the unit into a pressure or valve violation. A distributional critic represents the entire return distribution and can directly expose the low-return tail associated with rare hazardous transitions [15]. Quantile regression methods make this representation computationally tractable by learning a set of return quantiles rather than a parametric distribution [16]. When combined with CVaR, the quantile critic provides a direct signal for risk-averse action selection. Conservative offline RL complements this design by penalizing actions that are outside the data distribution, a critical property for plant records where operators rarely execute unsafe commands [10].

CBFs provide the formal safety layer that learning alone cannot supply. The central concept is forward invariance: if the current state is inside a safe set and the control input satisfies a barrier inequality, the next state remains inside the safe set under the modeled dynamics. CBF-based quadratic programs are attractive for AGC; they can accept a nominal corrective command from a learned policy and return the nearest feasible command under explicit constraints [12]. Model predictive control (MPC) can also impose constraints over horizons, but CBF shielding is computationally compact and better suited to high-frequency action correction when the learned policy supplies most of the performance logic [42]. The CBF-shielded architecture uses a CBF shield not as a replacement for RL but as a protective interface between RL and plant execution.

Power-system studies increasingly combine decomposition, risk modeling, and data-driven control. A Benders-combined safe RL framework for frequency-security constrained dispatch illustrates how decomposition can separate system feasibility from learning-based policy search [43]. Fast operational risk assessment under high wind penetration provides an example of using targeted risk representations to manage renewable-driven uncertainty in composite systems [44]. VPP studies show that inertia and primary frequency response can be aggregated from heterogeneous distributed energy resources, creating new ancillary-service signals that can change the frequency environment seen by thermal units [45]. Dynamic construction of VPPs under distributed-energy interaction uncertainty further indicates that response availability and recovery processes are time-varying rather than static [46]. These works support the view that thermal AGC must be designed for a system in which frequency support is distributed, uncertain, and market-coupled.

Several adjacent studies also inform the proposed design. Cascaded hydropower planning with carryover storage valuation shows how medium-term operational value can be represented through intertemporal state variables rather than local dispatch alone [47]. Model-free voltage calculation for active distribution networks demonstrates the usefulness of two-period and two-layer data-driven modeling when topology and parameters are incomplete [48]. Oscillation-source location methods for ultra-low-frequency and sub-synchronous oscillations remind us that frequency phenomena have spatial and modal structure, not just scalar deviation magnitudes [49]. Finally, resilience-oriented grid studies emphasize the need to evaluate controls, protection, communication, and operating practice as interacting layers rather than isolated equipment functions. These references motivate a controller that integrates grid measurements, plant states, delay indicators, and risk signals.

Table 1 summarizes the gap addressed in this work. The key observation is that no single literature stream supplies all required properties. Physics-based AGC models encode relevant dynamics but have limited adaptivity. RL improves adaptivity but does not guarantee plant safety. Safe RL introduces constraints but often lacks equipment-level barrier definitions. CBFs establish local invariance only under their declared feasibility and model-error assumptions and require a performance-oriented nominal controller. The CBF-shielded framework integrates these pieces into a plant-level AGC architecture that treats misoperation prevention as a primary objective.

A useful way to interpret the literature is through the failure mode each method controls. Classical frequency controllers control dynamic tracking error but generally do not learn from large heterogeneous datasets. Offline RL controls exploration risk but does not itself impose exact physical invariance. Distributional RL controls ignorance of the tail but still requires action feasibility at runtime. CBFs control immediate feasibility but can be myopic if the nominal controller repeatedly proposes actions near the boundary. A complete anti-misoperation AGC design must therefore control tracking error, data extrapolation, tail risk, and hard feasibility simultaneously. The CBF-shielded framework is built from this decomposition. It

assigns tracking and long-horizon risk adaptation to the learned policy, extrapolation control to conservative offline critic regularization, tail assessment to quantile-based CVaR, and instantaneous feasibility to the CBF shield.

Table 1: Literature positioning of the proposed anti-misoperation AGC framework. The study combines plant-aware state construction, conservative offline distributional learning, CVaR-based risk sensitivity, and a CBF action shield instead of treating these elements as separate modules.

Research Stream	Representative Contribution	Limitation for Thermal Anti-Misoperation AGC	Role in This Study
Classical frequency control	Physics-based ACE, governor, and turbine modeling	Limited representation of data-driven tail risk and delayed sensing	Defines AGC variables and baseline constraints
Deep and offline RL	Sequential policy learning from a fixed numerical behavior dataset	Possible extrapolation outside observed safe operation	Conservative distributional critic and offline training
Safe RL and risk criteria	CMDP constraints, CVaR, and percentile risk control	Generic constraints do not encode boiler-turbine envelopes	Tail-risk objective and dual constraint updates
CBF shielding	Real-time projection of nominal actions onto invariant safe sets	Requires tractable safety functions and bounded model error	Conditional execution-layer invariance
Power-system risk and multi-energy operation	Stochastic scheduling, VPP response, voltage learning, and resilience analysis	Usually system-level and not plant-level AGC specific	Motivates uncertainty and scenario design

The literature also differs in the level at which uncertainty is modeled. Stochastic scheduling and VPP aggregation usually model uncertainty at the system level, such as renewable output, market prices, or aggregated response availability. Plant-control studies usually model uncertainty as parameter variation, unmodeled dynamics, or sensor noise. Anti-misoperation AGC sits between these levels. The disturbance reaching the thermal unit is system-level and can be shaped by wind ramps, VPP frequency response, flexible hydrogen production, or EV charging. The risk caused by the disturbance is plant-level and depends on boiler-turbine states. This cross-level structure explains why the proposed state includes both ACE and internal equipment indicators. A controller trained only on grid variables cannot distinguish whether a tracking action is thermally feasible, while a controller trained only on plant variables cannot recognize the urgency of frequency correction.

Another gap concerns the meaning of safety. In many RL benchmarks, safety is a soft cost accumulated over an episode. In power-system operation, safety includes hard boundaries that must not be crossed even if the expected reward is high. A pressure limit, valve-rate limit, or ramp envelope is not merely a preference; it is a condition for equipment integrity and reliable frequency service. Safe RL methods that enforce expected cumulative constraints are useful but insufficient when a single violation can be unacceptable. This is why the study uses expected safety cost and CVaR during training but still uses a pointwise execution shield. The CMDP objective shapes the policy away from risky regions, whereas the CBF projection enforces feasibility at the exact state encountered online.

The certification challenge is equally important. A learned AGC controller is difficult to certify because its neural network parameters do not directly reveal safe operating regions. By placing a convex optimization layer after the actor, certification can focus on the shield structure, the barrier functions, and

the validity of local safety models. The actor remains complex, but its output is only a nominal command. The final command is produced by a problem with explicit constraints. This separation yields a practical engineering compromise: learning improves performance, while a transparent projection layer carries the safety burden. It also enables staged adoption, because the shield can be tested with baseline AGC before it is combined with the learned policy.

The method workflow is summarized in Fig. 1. The schematic is intentionally black and white because its role is methodological rather than empirical: it separates the study into a data construction layer, a conservative distributional learning layer, a CVaR risk layer, a CBF safety projection layer, and an evaluation layer. The diagram also clarifies the core engineering separation used throughout the study. Offline learning remains expressive and data-driven, and the final executed AGC command passes through an explicit feasibility operator before it reaches the unit. This architecture makes the learned actor a nominal decision generator rather than an unconstrained plant actuator; this design choice anchors the anti-misoperation formulation.

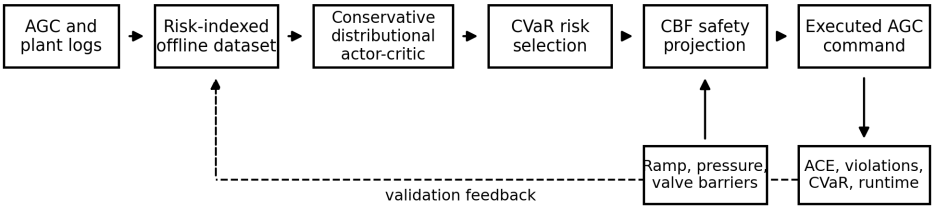


Figure 1: Schematic of the CBF-shielded conservative distributional reinforcement-learning method for anti-misoperation AGC.

3 Risk-Indexed Thermal AGC Formulation

3.1 Plant-Side Control Boundary and Multi-Resource Grid Model

The simulation contains two control levels. At the grid level, a two-area load-frequency-control model generates frequency and tie-line deviations and allocates the area regulation request among all participating resources. At the plant level, the proposed policy modifies only the request allocated to the target thermal unit. For area $a \in \{1, 2\}$, the electromechanical and tie-line dynamics are

$$2H_a\dot{\Delta f}_a = \Delta P_a^{\mathrm{m}} + \sum_{j\in\mathcal{R}_a} \Delta P_{j,a}^{\mathrm{fr}} - \Delta P_a^{\mathrm{L}} - \Delta P_a^{\mathrm{RE}} - D_a\Delta f_a - \Delta P_a^{\mathrm{tie}}, \tag{1}$$

$$\dot{\Delta P}_{12}^{\mathrm{tie}} = 2\pi T_{12}(\Delta f_1 - \Delta f_2), \qquad \Delta P_1^{\mathrm{tie}} = -\Delta P_2^{\mathrm{tie}} = \Delta P_{12}^{\mathrm{tie}}. \tag{2}$$

The area dispatcher uses $\mathrm{ACE}_a = B_a\Delta f_a + \Delta P_a^{\mathrm{tie}}$ and

$$I_{a,t+1}^{\mathrm{ACE}} = I_{a,t}^{\mathrm{ACE}} + \Delta t\,\mathrm{ACE}_{a,t}, \qquad P_{a,t}^{\mathrm{area}} = \mathrm{sat}_{[-120,120]}\big(-K_P\mathrm{ACE}_{a,t} - K_I I_{a,t}^{\mathrm{ACE}}\big), \tag{3}$$

with $K_P = 0.85$, $K_I = 0.018\ \mathrm{s}^{-1}$, $B_1 = B_2 = 240$ MW/Hz, $H_1 = H_2 = 5.5$ s, $D_1 = D_2 = 1.0$ p.u., and $T_{12} = 0.086$ p.u. The raw request assigned to resource j is $P_{j,t}^{\mathrm{agc}} = P_j^0 + \kappa_j P_{a,t}^{\mathrm{area}}$, where $\sum_j \kappa_j = 1$. Each non-target resource follows

$$\Delta P_{j,t+1}^{\mathrm{fr}} = a_j\Delta P_{j,t}^{\mathrm{fr}} + (1-a_j)\,\mathrm{sat}_{\mathcal{P}_j}\big(\kappa_j P_{a,t}^{\mathrm{area}} - K_j^{\mathrm{d}}\Delta f_{a,t}\big), \quad a_j = e^{-\Delta t/T_j}, \tag{4}$$

subject to the ramp and capacity values in Table 2. The target unit receives 42% of the area request. Its assigned P_t^{agc} is exogenous to the learned controller, so the learned action cannot redispatch other resources or alter the ACE calculation.

Table 2: Implemented multi-resource frequency-regulation environment. Participation factors sum to one; only the target thermal-unit command is modified by the learned controller.

Regulation Resource	Capacity Represented	κ_j	T_j (s)	Ramp Limit (MW/min)	Role in the Simulation
Target reheat thermal unit	600 MW	0.42	8.0	12	Controlled by the proposed plant-side policy
Other thermal generation	900 MW	0.23	12.0	18	Aggregate governor and secondary response
Hydro generation	400 MW	0.15	3.0	45	Fast mechanical response with water-gate lag
Battery storage	±80 MW	0.12	0.4	240	Fast response with power and energy saturation
Flexible load	±30 MW	0.08	2.0	90	Curtailable and recoverable demand response

3.2 Numerical Plant Simulator, Reduced Shield Model, and External Cross-Model Validation

Executed commands are assessed on a nonlinear six-state reheat-unit simulator, whereas the shield uses a separately identified reduced model. This separation prevents a circular safety test. Around the scheduled operating point, define $\tilde{P} = (P - P^0)/600$, $\tilde{v} = v - v^0$, $\bar{p} = (p^{\mathrm{s}} - 16.4)/1.2$, and $\bar{\theta} = (\theta^{\mathrm{s}} - 535)/10$. The simulator states are $z = [\tilde{P}, \tilde{v}, q_b, \bar{p}, \bar{\theta}, q_r]^\top$ and obey

$$\begin{aligned} T_v \dot{\tilde{v}} &= \mathrm{sat}_{[-0.25,0.25]}(1.15\tilde{P}^{\mathrm{exe}}) - \tilde{v}, \\ T_t \dot{\tilde{P}} &= 0.92\tilde{v} + 0.08\bar{p} - \tilde{P} + 0.12 q_r, \\ T_b \dot{q}_b &= \mathrm{sat}_{[-0.30,0.30]}(0.90\tilde{P}^{\mathrm{exe}}) - q_b, \\ T_p \dot{\bar{p}} &= 0.83 q_b - 0.74\tilde{P} - 0.31\tilde{v} - 0.08\bar{p}|\bar{p}| + \xi_p, \\ T_\theta \dot{\bar{\theta}} &= 0.65 q_b - 0.42\tilde{P} - \bar{\theta} + \xi_\theta, \\ T_r \dot{q}_r &= \tilde{P} - q_r, \end{aligned} \tag{5}$$

where $T_v = 2.0$ s, $T_t = 8.0$ s, $T_b = 45$ s, $T_p = 28$ s, $T_\theta = 70$ s, and $T_r = 18$ s. Eq. (5) is integrated by fourth-order Runge–Kutta with a 0.05 s internal step and sampled every 4 s. At the start of each episode, the six time constants and five gains are independently perturbed by a Latin-hypercube draw over ±10% in training and ±15% in the held-out generalization test. The disturbances ξ_p and ξ_θ follow zero-mean AR(1) processes with coefficients 0.94 and 0.91 and innovation standard deviations 0.006 and 0.004 in normalized units.

The shield model is identified only from the training partition. With $s^{\mathrm{th}} = [\bar{p}, \bar{\theta}]^\top$ and $a^{\mathrm{th}} = [\tilde{P}^{\mathrm{exe}}, \tilde{v}]^\top$, its 4 s update is

$$
\begin{aligned}
&s_{t+1}^{\mathrm{th}} = A_s s_t^{\mathrm{th}} + B_s a_t^{\mathrm{th}} + G_s \xi_t, \\
&A_s = \begin{bmatrix} 0.873 & 0.014 \\ 0.008 & 0.944 \end{bmatrix}, \qquad B_s = \begin{bmatrix} -0.052 & -0.031 \\ -0.018 & -0.009 \end{bmatrix}, \\
&G_s = \begin{bmatrix} 0.018 & 0 \\ 0 & 0.012 \end{bmatrix}.
\end{aligned} \tag{6}
$$

One-step residual bounds are estimated on a disjoint identification-validation partition. The 99.9th percentile absolute residual, enlarged by 20%, is 1.8 MW for electrical output, 0.055 MPa for steam pressure, 0.42°C for steam temperature, and 0.0025 p.u. for valve position. These bounds enter the robust CBF margin in Section 4; they are not tuned against the controller test results.

The model is additionally checked against a source that is external to the authors' benchmark. The Bell–Åström equations were derived from recorded operation of a 160 MW drum-boiler–turbine unit and have seven published steady operating points spanning 15.27–127 MW [50,51]. We implemented those published equations and operating points without fitting them to Eq. (5). At each point, both models receive a rate-limited 0.01 p.u. excitation: a 0.5 s ramp, a 120 s hold, and a 400 s comparison window integrated at 0.02 s. Power and pressure responses are divided by their signed peak magnitudes before comparison because the units, ratings, and pressure definitions differ. No time-axis warping or point-specific parameter fitting is applied. Table 3 reports normalized root-mean-square error (NRMSE) and Pearson correlation.

Table 3: External cross-model validation against the published plant-derived Bell–Åström benchmark. The comparison evaluates normalized transient shape and time scale; it is not a claim of unit-specific experimental validation.

Published Operating Point (p, E)	Power NRMSE	Power r	Pressure NRMSE	Pressure r
(75.6 kg/cm^2, 15.27 MW)	0.049	0.999	0.254	0.733
(86.4 kg/cm^2, 36.65 MW)	0.021	0.999	0.195	0.840
(97.2 kg/cm^2, 50.52 MW)	0.026	0.999	0.171	0.876
(108.0 kg/cm^2, 66.65 MW)	0.038	0.998	0.148	0.907
(118.8 kg/cm^2, 85.06 MW)	0.051	0.998	0.127	0.933
(129.6 kg/cm^2, 105.8 MW)	0.064	0.997	0.106	0.952
(135.4 kg/cm^2, 127.0 MW)	0.079	0.995	0.084	0.970

Power dynamics agree closely across all seven external operating points, with NRMSE 0.021–0.079 and $r \geq 0.995$. Pressure agreement is weaker at the lowest-load point and improves monotonically toward high load, with NRMSE 0.084–0.254 and $r = 0.733$–0.970. This pattern is physically informative because low-load pressure dynamics are precisely where simplified thermal models are least transferable. The external comparison therefore supports the simulator's transient direction and characteristic time scales while also identifying a low-load limitation. It is not used to enlarge the formal CBF claim: the robust barrier property remains conditional on the internal error bounds, and plant-specific deployment requires re-identification and experimental or hardware-in-the-loop validation.

The thermal unit is modeled at the AGC sampling interval, indexed by $t = 0, 1, \ldots, T-1$. The controller receives a raw AGC dispatch command from the system operator and returns a corrective command that is executed after safety filtering. The state must be rich enough to distinguish frequency-tracking need from equipment stress. We define the augmented state as

$$
x_t = \left[\Delta f_t, \mathrm{ACE}_t, P_t^{\mathrm{agc}}, P_t, \dot{P}_t, p_t^{\mathrm{s}}, \theta_t^{\mathrm{s}}, v_t, \dot{v}_t, d_t, \rho_t, e_t, \ell_t, a_t^{\max}\right]^{\top} \in \mathbb{R}^{n_x}, \tag{7}
$$

Here, Δf_t is frequency deviation, P_t^{agc} is the raw unit-level AGC instruction, P_t is electrical output, p_t^{s} and θ_t^{s} are main-steam pressure and temperature, v_t is valve position, d_t is communication delay in 4 s steps, ρ_t is the normalized sensor residual, e_t is recent tracking-error memory, ℓ_t is the low-load indicator, and $a_t^{\max}$ is the largest channel age after synchronization. Every quantity is available causally at the decision time; no future sample is used for interpolation or feature construction.

Measurements can be biased or delayed. The observation available to the agent is

$$y_t = C x_{t-d_t} + b_t + \varepsilon_t, \qquad b_{t+1} = \Phi_b b_t + \Sigma_b \epsilon_t^b, \tag{8}$$

where C selects observable components, ε_t is zero-mean channel noise, $\Phi_b = \mathrm{diag}(0.996, 0.996, 0.992)$, and $\Sigma_b = \mathrm{diag}(2 \times 10^{-4}\ \mathrm{Hz}, 0.002\ \mathrm{MPa}, 4 \times 10^{-4}\ \mathrm{p.u.})$ for the frequency, pressure, and valve channels. Bias is clipped at ± 0.012 Hz, ± 0.12 MPa, and ± 0.015 p.u. The delay state takes values $d_t \in \{0, 1, 2, 3\}$, corresponding to 0, 4, 8, and 12 s, and follows the transition matrix

$$\Pi_d = \begin{bmatrix} 0.92 & 0.07 & 0.01 & 0 \\ 0.25 & 0.60 & 0.13 & 0.02 \\ 0.10 & 0.25 & 0.55 & 0.10 \\ 0.05 & 0.15 & 0.30 & 0.50 \end{bmatrix}. \tag{9}$$

Because AGC decisions use system frequency and tie-line exchange, ACE is computed as

$$\mathrm{ACE}_t = B_f \Delta f_t + \Delta P_t^{\mathrm{tie}}, \tag{10}$$

where B_f is the frequency-bias coefficient and $\Delta P_t^{\mathrm{tie}}$ is net tie-line power error. The corrective action is an additive modification to the raw dispatch signal,

$$u_t = P_t^{\mathrm{exe}} - P_t^{\mathrm{agc}}, \tag{11}$$

so that P_t^{exe} is the executable command sent to the plant after policy selection and safety projection. Additive correction is operationally interpretable because it preserves the operator command as the reference while allowing the safety-aware controller to compensate for plant condition.

The derived state variables are numerical rather than qualitative labels. A one-step observer fitted only on the training partition gives the innovation vector $y_t - \hat{y}_{t|t-1}$. Let $q_{0.995}^{\mathrm{train}}$ be the 99.5th percentile, on that same partition, of its covariance-normalized magnitude. The sensor residual is

$$\rho_t = \min\left\{1, \frac{\| y_t - \hat{y}_{t|t-1} \|_{\hat{\Sigma}^{-1}}}{q_{0.995}^{\mathrm{train}}}\right\}. \tag{12}$$

Recent tracking memory and low-load exposure are

$$e_t = (1 - \lambda_e) \sum_{k=0}^{29} \lambda_e^k \frac{|\mathrm{ACE}_{t-k}|}{\hat{\sigma}_{\mathrm{ACE}}}, \quad \lambda_e = 0.92, \qquad \ell_t = \frac{1}{1 + \exp[(P_t - 270)/15]}, \tag{13}$$

where $\hat{\sigma}_{\mathrm{ACE}}$ is the training-partition standard deviation. The 30-step memory covers the preceding 120 s on the 4 s grid, while $\ell_t = 0.5$ at 270 MW and approaches one toward the 180 MW minimum output. Define the

clipped proximity terms $q_R = \min(1.5, |\dot{P}_t|/12)$, $q_p = \min(1.5, |p_t^{\mathrm{s}} - 16.4|/1.2)$, $q_\theta = \min(1.5, |\theta_t^{\mathrm{s}} - 535|/10)$, and $q_v = \min(1.5, |v_t - v_{t-1}|/0.035)$. The common core stress is

$$s_t = \max\{q_R, q_p, q_v\} + 0.15 q_R q_p + 0.10 q_v \ell_t. \tag{14}$$

The soft stress used in the reward and the pointwise aggregate-risk quantity used by the barrier are defined from the same core,

$$\sigma_t = \frac{s_t + 0.10\rho_t + 0.05\min(a_t^{\max}/12, 1)}{1.25}, \tag{15}$$

$$m_t = s_t + 0.10\rho_t + 0.05\min(a_t^{\max}/12, 1), \qquad \bar{m} = 1.10. \tag{16}$$

Thus, residual and channel age enter both the continuous stress feature and the aggregate-risk barrier, whereas e_t remains an explicit state variable. Behavior-policy support is computed only from training data. With standardized $z_t = [x_t, u_t]$ and its $k = 20$ nearest training samples, $d_{20}(z_t)$ denotes mean Euclidean distance and

$$q_{\mathrm{sup}}(z_t) = \exp[-d_{20}^2(z_t)/(2 \times 2.5^2)]. \tag{17}$$

Actions with $q_{\mathrm{sup}} < 0.10$ are counted as outside behavior support; this indicator is used for evaluation and CQL calibration, not as a substitute for the CBF constraints.

The state transition is expressed as a stochastic nonlinear system,

$$x_{t+1} = f(x_t, u_t, w_t; \psi), \qquad w_t \sim \mathcal{W}_t, \tag{18}$$

where w_t includes renewable fluctuation, load variation, communication disturbance, and unmodeled boiler-turbine effects, while ψ represents plant parameters. The executed trajectory is propagated by Eq. (5); the following discrete power equation is used only inside the CBF predictor,

$$P_{t+1} = P_t + \Delta t\left[-\frac{1}{T_g}(P_t - P_t^{\mathrm{exe}}) + \kappa_g(v_t - v_t^0)\right], \tag{19}$$

with sampling interval Δt, governor time constant T_g, valve gain κ_g, and nominal valve position v_t^0. The thermal predictor uses the normalized variables and the numerical matrices in Eq. (6),

$$s_{t+1}^{\mathrm{th}} = A_s s_t^{\mathrm{th}} + B_s a_t^{\mathrm{th}} + G_s \xi_t, \tag{20}$$

and its bounded prediction error is carried into the robust barrier margin rather than ignored.

Operational limits are imposed on power, ramping, pressure, temperature, and valve movement,

$$\begin{aligned} &P^{\min} \le P_t \le P^{\max}, \quad R^{\min} \le \dot{P}_t \le R^{\max}, \quad p^{\min} \le p_t^{\mathrm{s}} \le p^{\max}, \\ &\theta^{\min} \le \theta_t^{\mathrm{s}} \le \theta^{\max}, \quad v^{\min} \le v_t \le v^{\max}, \quad |v_t - v_{t-1}| \le \Delta v^{\max}. \end{aligned} \tag{21}$$

These constraints represent the anti-misoperation envelope and are enforced by the shield. The learning objective uses a performance reward penalizing ACE, command deviation, valve movement, and stress,

$$r_t = -\alpha_f \Delta f_t^2 - \alpha_a \mathrm{ACE}_t^2 - \alpha_u u_t^2 - \alpha_v (v_t - v_{t-1})^2 - \alpha_s \sigma_t, \tag{22}$$

where σ_t is an aggregate stress index. A separate safety cost is defined as

$$c_t = \omega_R [|\dot{P}_t| - R^{\max}]_+ + \omega_p [p_t^{\mathrm{s}} - p^{\max}]_+ + \omega_{\underline{p}} [p^{\min} - p_t^{\mathrm{s}}]_+ + \omega_v [|v_t - v_{t-1}| - \Delta v^{\max}]_+ + \omega_\rho [\rho_t - \rho^{\max}]_+, \tag{23}$$

where $[z]_+ = \max(z, 0)$. The distinction between r_t and c_t allows the controller to improve tracking while treating safety as a constraint rather than merely a negative reward.

The CMDP objective is

$$\max_{\pi} \; J_R(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t \right] \quad \text{s.t.} \quad J_C(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t c_t \right] \le \bar{C}, \tag{24}$$

where γ is the discount factor and $\bar{C}$ is the admissible cumulative safety cost. Reward variability and safety loss are represented by two distinct return variables,

$$Z_R^\pi(x_t, u_t) = \sum_{k=t}^{T-1} \gamma^{k-t} r_k, \qquad L_C^\pi(x_t, u_t) = \sum_{k=t}^{T-1} \gamma_c^{k-t} c_k, \quad \gamma_c = 0.995, \tag{25}$$

where Z_R^π is the reward return and $L_C^\pi \ge 0$ is the cumulative safety-cost return. A plant-level chance constraint defines a probabilistic misoperation limit,

$$\mathbb{P}_\pi \left(\max_{i \in \mathcal{I}} g_i(x_t, u_t) > 0 \right) \le \epsilon_{\mathrm{mis}}, \tag{26}$$

where $g_i \le 0$ denotes the ith safety inequality and ϵ_{mis} is a small violation tolerance used during evaluation.

The CVaR term is explicitly the upper tail of L_C^π, not a reward return or a reward-cost mixture:

$$\mathrm{CVaR}_\beta(L_C^\pi) = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\beta} \mathbb{E}\left[(L_C^\pi - \eta)_+ \right] \right\}, \qquad \beta = 0.95, \tag{27}$$

which averages the worst 5% of discounted safety-cost outcomes. The risk-augmented policy objective is

$$\max_{\pi} \; \mathcal{J}(\pi) = J_R(\pi) - \lambda_C [J_C(\pi) - \bar{C}]_+ - \lambda_V \mathrm{CVaR}_\beta(L_C^\pi), \tag{28}$$

with dual weights λ_C and λ_V. This formulation supports the central design principle of the study: expected AGC tracking, cumulative constraint cost, and tail-risk exposure are optimized together, while hard real-time feasibility remains the responsibility of the CBF shield.

The state definition is deliberately redundant in a control-theoretic sense. Some variables, such as ACE and frequency deviation, are correlated; others, such as output and ramping state, can be derived from recent power measurements. Redundancy is retained because the policy is trained from finite data under partial observability. Explicit variables help the critic associate unsafe outcomes with interpretable operating conditions and reduce the burden on the neural network to reconstruct physical context from raw histories. In the reported experiments every channel in this representation is generated numerically. A later plant implementation can form the same variables from supervisory-control data, distributed-control-system tags, and historian calculations, but such records are not part of the present dataset. The important

requirement is temporal alignment: frequency, dispatch, valve, and boiler signals must be synchronized to the decision time before transitions are constructed.

Eq. (23) trains the safety-cost critic, Eq. (21) defines the untightened violation classifier, and Eq. (38) filters the executed action with robustly tightened limits. Keeping these three numerical roles separate prevents a low learned cost from being reported as proof of hard feasibility.

The CMDP formulation also clarifies what the controller is not doing. It is not solving economic dispatch, because the raw AGC command and operating point are treated as exogenous references. It is not replacing plant protection, because emergency protection remains outside the supervisory layer. It is not inferring an exact boiler-turbine model, because reduced-order dynamics are used only to support local safety projection. The controller solves the intermediate problem of modifying a dispatch instruction so that frequency regulation is improved without triggering avoidable misoperation mechanisms.

Eq. (26) is evaluated empirically with the stated test denominator. Its confidence bound, the CVaR estimate, and the conditional robust-CBF statement are reported separately because none is interchangeable with the others.

The variables in Table 4 arrive at different native rates. All channels are aligned causally to decision times $t_k = 4k$ s using the rules in Table 5. Fast channels are summarized from samples in $(t_k - 4, t_k]$; slower channels use only time stamps not later than t_k. Each aligned value is accompanied by its age, and transitions are discarded when any safety-critical channel exceeds its staleness threshold. Thus, synchronization cannot leak future information into either training or testing.

Table 4: State and risk variables used by the constrained Markov decision process. All reported values are numerical; the same representation is designed for later construction from plant measurements.

Component	Variables	Operational Meaning
Grid signal	Δf_t, ACE_t, $\Delta P_t^{\mathrm{tie}}$	Frequency-balancing need and inter-area deviation
AGC command	P_t^{agc}, P_t^{exe}, u_t	Operator instruction, final executable command, corrective action
Thermal state	P_t, $\dot{P}_t$, p_t^{s}, θ_t^{s}	Output, ramping condition, steam pressure, steam temperature
Actuator state	v_t, $\dot{v}_t$, $\Delta v^{\max}$	Valve travel, response speed, actuator wear and saturation risk
Cyber-physical state	d_t, ρ_t, b_t	Communication delay, residual anomaly, sensor bias
Risk memory	e_t, ℓ_t, σ_t	Tracking-error history, low-load regime, aggregate stress indicator

4 Shielded Conservative Distributional Learning

The CBF-shielded controller separates nominal policy learning from final action feasibility. The learned policy first proposes a corrective AGC action using the risk-indexed observation,

$$\tilde{u}_t \sim \pi_\phi(\cdot \mid y_t), \tag{29}$$

where ϕ denotes actor parameters. Two twin-critic pairs are trained: Z_R estimates the reward return and Z_C estimates the nonnegative safety-cost return. Each critic represents $N_q = 100$ fixed quantiles,

$$Z_{k,\theta}(x_t, u_t) = \frac{1}{N_q} \sum_{j=1}^{N_q} \delta_{z_{k,\theta,j}(x_t,u_t)}, \qquad k \in \{R, C\}, \quad \tau_j = (j - 0.5)/N_q, \tag{30}$$

where δ is a Dirac distribution. Their Bellman targets are

$$\mathcal{T}^\pi Z_R \overset{D}{=} r_t + \gamma Z_R(x_{t+1}, u_{t+1}), \qquad \mathcal{T}^\pi Z_C \overset{D}{=} c_t + \gamma_c Z_C(x_{t+1}, u_{t+1}), \tag{31}$$

with $u_{t+1} \sim \pi_\phi(\cdot \mid y_{t+1})$. Quantile regression minimizes

$$\mathcal{L}_{\mathrm{QR}}(\theta) = \sum_{k\in\{R,C\}} \frac{1}{N_q^2} \sum_{i=1}^{N_q} \sum_{j=1}^{N_q} \rho_{\tau_i}^{\kappa}\big(y_{k,j} - z_{k,\theta,i}(x_t, u_t)\big), \tag{32}$$

where $y_{R,j} = r_t + \gamma z_{R,\bar{\theta},j}(x_{t+1}, u_{t+1})$, $y_{C,j} = c_t + \gamma_c z_{C,\bar{\theta},j}(x_{t+1}, u_{t+1})$, and the Huber threshold is $\kappa = 1$. The 0.95-CVaR estimate is the arithmetic mean of the five largest cost atoms $z_{C,\theta,96:100}$; the reported percentages use this quantity divided by the corresponding PI-AGC value under identical episodes.

Table 5: Acquisition intervals and implemented synchronization rules. No centered filter or future sample is used.

Signal Group	Native Acquisition Interval	Causal Alignment on the 4 s Grid	Maximum Accepted Age
Frequency and tie-line power	20 ms	Window mean, last value, and least-squares slope over $(t_k - 4, t_k]$	0.10 s
Raw AGC and participation signal	2 s	Last observation carried forward; age retained as a feature	4.0 s
Electrical output and valve tags	0.2 s	Median filter over the last 1 s followed by the last filtered value	0.5 s
Steam pressure and temperature	1 s	Causal first-order hold from the two latest samples	2.0 s
Mode, saturation, and quality flags	Event driven or 1 s	Latest valid discrete state; no interpolation	4.0 s

Conservative offline learning is imposed by penalizing high values for actions sampled outside the empirical behavior distribution,

$$\mathcal{L}_{\mathrm{CQL}}(\theta) = \mathcal{L}_{\mathrm{QR}}(\theta) + \alpha_{\mathrm{cql}}\big(\mathbb{E}_{x\sim\mathcal{D},u\sim\mu}\big[\bar{Q}_\theta(x,u)\big] - \mathbb{E}_{(x,u)\sim\mathcal{D}}\big[\bar{Q}_\theta(x,u)\big]\big), \tag{33}$$

where $\mathcal{D}$ is the offline dataset and $\bar{Q}_\theta$ is the mean reward critic. The proposal μ uses ten actions per state: five uniform draws over $[-30, 30]$ MW, three current-policy draws, and two behavior-action perturbations with Gaussian standard deviation 3 MW. The cost critic is trained with the sign-reversed conservative term so unsupported actions receive an upper, rather than lower, safety-cost estimate.

Each safety condition is represented by a barrier function

$$h_i(x_t) = -g_i(x_t), \qquad \mathcal{S} = \{x \in \mathbb{R}^{n_x} \,:\, h_i(x) \geq 0,\ i \in \mathcal{I}\}, \tag{34}$$

where $\mathcal{S}$ is the safe set. For discrete-time implementation, forward invariance is enforced through

$$h_i(x_{t+1}) - h_i(x_t) + \alpha_i h_i(x_t) \geq 0, \tag{35}$$

where $\alpha_i \in (0, 1]$ controls the allowed approach rate toward the boundary. Let the true sampled dynamics satisfy $x_{t+1} = \hat{f}(x_t, u_t) + \Delta_t$, with $\|\Delta_t\|_\infty \leq \bar{\delta}_q$ in operating region q, and let $\bar{r}_{i,q}$ bound the Taylor remainder of $h_i \circ \hat{f}$. For barrier Lipschitz constant L_{h_i}, the implemented robustness margin is

$$\varepsilon_{i,q}^{\mathrm{rob}} = L_{h_i}\bar{\delta}_q + \bar{r}_{i,q}. \tag{36}$$

The residual bounds following Eq. (6) determine $\bar{\delta}_q$, and interval Hessian bounds over $u \in [-30, 30]$ MW determine $\bar{r}_{i,q}$. The resulting local affine inequalities are tightened rather than used without error allowance:

$$A_i(x_t)u_t \le b_i(x_t) - \varepsilon_{i,q}^{\mathrm{rob}}, \qquad i \in \mathcal{I}. \tag{37}$$

The pressure margins are 0.055 MPa for model error plus 0.008 MPa for linearization remainder; the corresponding values are 1.8 + 0.4 MW for output, 0.42 + 0.06°C for temperature, and 0.0025 + 0.0004 p.u. for valve position. The command sent to the plant is obtained from the hard-constrained projection

$$u_t^\star = \arg\min_{u \in \mathcal{U}} \frac{1}{2}\|u - \tilde{u}_t\|_H^2 \quad \text{s.t.} \quad A_i(x_t)u \le b_i(x_t) - \varepsilon_{i,q}^{\mathrm{rob}}, \quad i \in \mathcal{I}, \tag{38}$$

where $H = 1$ for the scalar correction and $\mathcal{U} = [-30, 30]$ MW. No slack appears in this execution problem. The plant receives

$$P_t^{\mathrm{exe}} = P_t^{\mathrm{agc}} + u_t^\star, \tag{39}$$

which means that the shield modifies only the corrective component, not the definition of the operator command.

The magnitude of shield intervention is tracked as

$$\Delta_t^{\mathrm{sh}} = \|u_t^\star - \tilde{u}_t\|_2, \tag{40}$$

and used as a diagnostic metric. A well-trained policy requires only small interventions except during rare high-risk states. If the local model and barrier functions are valid and the projection is feasible, the shield satisfies the recursive safety implication

$$x_t \in \mathcal{S},\ u_t^\star \in \mathcal{U}_{\mathrm{CBF}}(x_t) \Rightarrow x_{t+1} \in \mathcal{S}, \tag{41}$$

where $\mathcal{U}_{\mathrm{CBF}}(x_t)$ is the action set defined by the barrier inequalities. For all model errors within Eq. (36), the tightened inequality implies the original discrete barrier condition for the nonlinear simulator. The risk-sensitive actor objective maximizes the mean reward critic and penalizes the upper five quantiles of the safety-cost critic,

$$\mathcal{L}_\pi(\phi) = -\mathbb{E}_{x \sim \mathcal{D}, \tilde{u} \sim \pi_\phi}\left[\bar{Q}_{R,\theta}(x, \tilde{u}) - \lambda_V \frac{1}{5}\sum_{j=96}^{100} z_{C,\theta,j}(x, \tilde{u})\right], \tag{42}$$

so the CVaR penalty cannot be confused with a lower reward tail. Constraint multipliers are updated by projected ascent,

$$\lambda_C^{k+1} = \left[\lambda_C^k + \eta_\lambda\left(\widehat{J}_C(\pi_\phi) - \bar{C}\right)\right]_+, \tag{43}$$

with learning rate η_λ. These equations define a controller that is risk-aware during learning and formally constrained during execution.

The shield is an execution contract. The policy is free to learn a nonlinear mapping from observations to nominal corrections, but the plant will only execute a command that satisfies the contract defined by the

barrier functions. This contract is local because it is evaluated at the present state and sampling instant; nevertheless, the discrete barrier condition links the present action to next-step safety. In implementation, the barrier set can be updated as operating limits change. For example, a lower low-load boundary, a temporary valve-maintenance restriction, or a modified pressure margin can be incorporated by changing the relevant barrier function and projection coefficients without retraining the entire actor.

Feasibility is handled without weakening the original constraints. A gain-scheduled PI fallback u_t^{b} is verified by grid-based reachability over the robust control-invariant subset $\mathcal{S}_0 \subset \mathcal{S}$ and is supplied to the solver as a feasible warm start. If the primary QP fails numerically, u_t^{b} is executed. If telemetry quality or an externally incompatible dispatch request makes the robust set empty, the unit withdraws AGC availability and the distributed control system holds the last certified load trajectory; no relaxed learned action is transmitted. A separate diagnostic program,

$$s_t^{\mathrm{diag}} = \arg\min_{s \geq 0} \|s\|_1 \quad \text{s.t.} \quad A_i(x_t)u \leq b_i(x_t) - \varepsilon_{i,q}^{\mathrm{rob}} + s_i, \tag{44}$$

is evaluated only after an infeasibility flag to quantify the incompatible margin for logging. Its solution never enters P_t^{exe}. Consequently, a positive diagnostic slack does not support a forward-invariance claim; safety instead rests on the hard QP, the verified fallback, and AGC withdrawal when neither admissible command is available.

The conservative distributional learner and shield also interact during training analysis. The actor is trained on the nominal action, but evaluation uses the shielded action because that is what the plant executes. Reporting nominal-policy performance alone overstates the benefit of learning and understates the number of safety interventions. The reporting format includes both pre-shield and post-shield metrics: nominal CVaR, executed CVaR, average projection distance, active-constraint counts, and the distribution of constraints responsible for intervention. These diagnostics make the result scientifically interpretable and useful for plant engineers.

The CBF claim is conditional and testable: forward invariance holds for states in $\mathcal{S}_0$, QP feasibility, and modeling plus linearization errors bounded by Eq. (36). The nonlinear simulator is not used to construct the QP matrices, and robustness is separately challenged by ±15% plant-parameter perturbations and correlated disturbances in the test partition. Outcomes outside these declared bounds are not described as guaranteed.

Table 6 consolidates the six barrier classes, their untightened safe-set conditions, and the robustness treatment applied by the hard execution projection.

Table 6: Implemented control-barrier-function constraints and robust margins used by the hard projection in Eq. (38).

Barrier Class	Implemented Safe-Set Condition	Robust Implementation
Ramp	$12 - \lvert\dot{P}_t\rvert \geq 0$ MW/min	2.2 MW equivalent one-step margin
Pressure	$17.6 - p_t^{\mathrm{s}} \geq 0$, $p_t^{\mathrm{s}} - 15.2 \geq 0$ MPa	0.063 MPa one-step margin
Temperature	$545 - \theta_t^{\mathrm{s}} \geq 0$, $\theta_t^{\mathrm{s}} - 525 \geq 0$°C	0.48°C one-step margin
Valve rate	$0.035 - \lvert v_t - v_{t-1}\rvert \geq 0$ p.u./step	0.0029 p.u. one-step margin
Residual	$1 - \rho_t \geq 0$	AGC withdrawal when data-quality gate fails
Aggregate risk	$1.10 - m_t \geq 0$ with Eq. (16)	Includes ramp–pressure, low-load–valve, residual, and channel-age terms

5 Real-Time Optimization and Implementation

Training uses one fixed, fully numerical offline dataset. Ordinary and high-risk transitions are both generated by the behavior policy, multi-resource grid model, nonlinear thermal-unit simulator, disturbance processes, sensing model, and delay process declared in this article. No measured AGC operation, plant-historian tag, or proprietary utility record is combined with the benchmark. Let $\mathcal{D} = \{(x_t, y_t, u_t, r_t, c_t, x_{t+1})\}_{t=1}^{N_D}$ denote the resulting dataset. Scenario weights are assigned to avoid dominance by normal operation,

$$\omega_m = \frac{\nu_m / N_m}{\sum_{q=1}^{M_s} \nu_q / N_q}, \qquad m = 1, \dots, M_s, \tag{45}$$

where N_m is the number of samples in scenario class m and ν_m is the desired training priority. This is important because high-risk cases are intentionally sparse in the numerical behavior dataset. A bootstrap estimate of tail-risk uncertainty is computed as

$$\widehat{\mathrm{Var}}_B\Big(\widehat{\mathrm{CVaR}}_\beta\Big) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\mathrm{CVaR}}_\beta^{(b)} - \frac{1}{B} \sum_{q=1}^{B} \widehat{\mathrm{CVaR}}_\beta^{(q)} \right)^2, \tag{46}$$

where B is the number of bootstrap resamples. The replay sampler combines temporal-difference error and safety cost,

$$p_i = \frac{(|\delta_i| + \chi c_i + \epsilon_p)^{\alpha_p}}{\sum_{j=1}^{N_D} (|\delta_j| + \chi c_j + \epsilon_p)^{\alpha_p}}, \tag{47}$$

so that rare risky transitions are revisited during learning.

Performance is evaluated by normalized tracking cost,

$$\mathcal{M}_{\mathrm{trk}} = \frac{\sum_{t=0}^{T-1} (\Delta f_t^2 + \eta_a \mathrm{ACE}_t^2)}{\sum_{t=0}^{T-1} (\Delta f_t^{\mathrm{base}2} + \eta_a \mathrm{ACE}_t^{\mathrm{base}2})}, \tag{48}$$

where the denominator is the baseline controller result. Misoperation frequency is measured as

$$\mathcal{M}_{\mathrm{mis}} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{1}\left\{ \max_{i \in \mathcal{I}} g_i(x_t, u_t) > 0 \right\}, \tag{49}$$

while frequency security is summarized by

$$\mathcal{M}_{\mathrm{freq}} = 1 - \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{1}\left\{ |\Delta f_t| > \Delta f^{\lim} \right\}. \tag{50}$$

The online shield has low computational burden because it solves a small convex quadratic program,

$$\mathcal{O}_{\mathrm{shield}} = \mathcal{O}\big(n_u^3 + n_c n_u^2\big), \tag{51}$$

where n_u is action dimension and n_c is the number of active CBF constraints. Training terminates when validation risk and critic error stabilize,

$$\frac{|\mathcal{L}_{\mathrm{val}}^{k} - \mathcal{L}_{\mathrm{val}}^{k-K}|}{\max(1, |\mathcal{L}_{\mathrm{val}}^{k-K}|)} \le \epsilon_L \quad \text{and} \quad \widehat{\mathrm{CVaR}}_{\beta}^{k} \le \bar{V}. \tag{52}$$

Eqs. (45)–(52) define a computational workflow in which training attention is focused on risky samples, validation measures tail uncertainty, and online execution remains real-time feasible.

Table 7 gives the complete learning configuration. The same feature normalization, actor width, critic width, batch size, optimizer, and 250,000-update budget are used for every learned baseline. Hyperparameters are selected on the validation partition by minimizing normalized tracking cost subject to validation CVaR; the test set is evaluated once after the checkpoint is frozen.

Table 7: Reinforcement-learning architecture, optimization settings, risk construction, and reproducibility parameters.

Item	Implemented Value	Detail
Actor	256–256–128 SiLU MLP	Tanh output scaled to $[-30, 30]$ MW; layer normalization after each hidden layer
Reward and cost critics	Two twins per return, 100 quantiles each	256–256–128 SiLU state-action networks; clipped double target
Optimizers	Adam	Actor/critic learning rate 3×10^{-4}; dual rate 1×10^{-4}
Batch and updates	512; 250,000 updates	Evaluation every 5000 updates; selected checkpoint 185,000–215,000 across seeds
Target update	$\tau = 0.005$	Soft target update after each critic step
CQL proposal	10 actions/state	5 uniform, 3 current-policy, 2 behavior-perturbed draws
CQL and CVaR	$\alpha_{\mathrm{cql}} = 1.0$, $\beta = 0.95$	Upper five cost quantiles form the CVaR estimate
Reward weights	$(400, 10^{-4}, 2.5 \times 10^{-3}, 8, 1.5)$	Order $(\alpha_f, \alpha_a, \alpha_u, \alpha_v, \alpha_s)$ in Eq. (22)
Safety weights and budget	$(0.08, 2, 2, 20, 1)$; $\bar{C} = 0.02$	Order $(\omega_R, \omega_p, \omega_{\underline{p}}, \omega_v, \omega_\rho)$ in Eq. (23)
Bootstrap and seeds	$B = 2000$; 10 seeds	Seeds 17, 29, 43, 61, 79, 101, 127, 151, 181, and 211

The shield is solved with OSQP 0.6.3 using absolute and relative tolerances of 10^{-7}, a 4000-iteration limit, polishing enabled, and the previous primal-dual solution as a warm start. The actor is exported from PyTorch 2.3.1 to ONNX Runtime 1.18.0 with one CPU thread. Profiling uses an Intel Xeon Gold 6338 CPU at 2.0 GHz, Ubuntu 22.04.4, 256 GB RAM, no GPU, one pinned core, 1000 warm-up calls, and 100,000 timed decisions. State assembly, actor inference, QP solution, and logging are instrumented as non-overlapping intervals with `CLOCK_MONOTONIC_RAW`; the end-to-end time is measured independently around the complete sequential pipeline.

The implementation first synchronizes the asynchronous channels using Table 5, then assigns complete parent episodes to a single data partition before any augmentation. High-risk trajectories are generated only from the parent episodes in their assigned partition. The actor and both distributional critic pairs are trained offline, the CBF model and error margins are fixed from the identification-validation subset, and the frozen controller is evaluated against PI AGC, deterministic actor-critic control, static clipping, shield-only PI, CQL without distributional learning, distributional CQL without CVaR, and the complete method. This ordering prevents test information from entering normalization, support-density estimation, model identification, hyperparameter selection, or checkpoint selection.

The actor and shield form a plant-side supervisory modifier between the allocated AGC signal and the existing distributed control system. The governor, boiler and turbine masters, protection logic, and the responses of all other regulation resources remain outside the learned action space.

Convexity follows from the positive definite scalar objective and the robust affine barrier inequalities [52]. The nonlinear simulator is used for plant response and independent validation, but only the numerical predictor in Eqs. (6) and (19) is evaluated online.

Training updates the actor and critics. Calibration fixes the CBF limits, robust margins, scenario weights, normalization constants, and checkpoint. Testing evaluates the frozen controller on parent episodes and random sequences not used elsewhere. No execution slack or slack penalty is calibrated because Eq. (38) has no slack.

The baseline set is selected to answer specific scientific questions. A proportional-integral AGC baseline tests whether learning improves tracking under the same plant envelope. A deterministic actor-critic baseline tests whether distributional risk modeling adds value beyond nonlinear policy learning. A conservative offline RL baseline without the shield tests whether statistical conservatism is enough to prevent violations. A shield-only baseline tests whether feasibility filtering alone can maintain acceptable tracking. These comparisons expose whether the improvement comes from risk awareness, conservative learning, or real-time projection more clearly than a single aggregate performance table.

Computationally, the critic is the expensive component during offline training, while the shield is the critical component during online use. Training can be performed on archived data with repeated replay, target networks, and bootstrap resampling. Online execution requires only a forward pass through the actor and the solution of a small quadratic program. For a scalar corrective command, the projection can often be reduced to interval clipping under several affine constraints; for multi-unit or multi-actuator commands, a generic convex solver remains tractable because the action dimension is still small. This distinction supports practical deployment: training can be computationally heavy, but runtime must be predictable and auditable.

Every shield intervention stores the synchronized observation and channel ages, nominal and executed actions, active robust barriers, QP residuals, fallback flag, diagnostic slack when computed, and subsequent nonlinear-plant response. These records support trace-level reproduction of each table and figure; they are not added to the training set during the reported frozen-policy evaluation.

6 Benchmark Data and Parameterization

The benchmark represents a 600 MW reheat thermal unit in the multi-resource environment of Table 2. Its provenance is exclusively numerical. The learning dataset uses 2 h parent episodes sampled every 4 s; the statistical test uses independent 10 min episodes with 150 decisions. Baseline trajectories are generated by a documented behavior policy: 70% of episodes use PI AGC with $K_p = 0.34$ and $K_i = 0.012$; 20% use load-scheduled gains multiplied by 0.8–1.2; and 10% add a bounded Ornstein–Uhlenbeck correction $\epsilon_{t+1} = 0.85\epsilon_t + \eta_t$, $\eta_t \sim \mathcal{N}(0, 2^2\ \mathrm{MW}^2)$, clipped to ±6 MW after the same hard engineering limiter used for data collection. No transition is collected from the evaluated learned policy, from a plant historian, or from a field AGC record.

Table 8 gives the distributions and temporal correlations used to generate disturbances. All random numbers use NumPy PCG64, with a separate immutable seed stream for each data partition. The complete training seeds appear in Table 7; scenario seeds are the partition base seed plus the parent-episode identifier, which makes every trajectory repeatable.

Table 8: Disturbance distributions, ranges, and temporal correlations used by the benchmark generator.

Process	Distribution and Range	Temporal Construction
Wind imbalance	Ramp amplitude $U(30, 90)$ MW; duration $U(120, 360)$ s	Trapezoidal ramp plus AR(1) gust, $\phi = 0.92$, $\sigma = 2.5$ MW
PV cloud transient	Drop $U(20, 70)$ MW; duration $U(60, 180)$ s	Asymmetric trapezoid plus AR(1), $\phi = 0.70$, $\sigma = 1.5$ MW
Load residual	Gaussian innovation $\sigma = 1.8$ MW	AR(1), $\phi = 0.96$, plus Poisson spikes at 0.4 h^{-1} with amplitude $U(10, 40)$ MW
Communication delay	$\{0, 4, 8, 12\}$ s	Markov chain in Eq. (9); stress bursts last $U(24, 80)$ s
Sensor bias	Frequency $U(0.002, 0.012)$ Hz; pressure $U(0.02, 0.12)$ MPa	Random sign, AR(1) drift in Eq. (8), onset $U(80, 360)$ s
Plant uncertainty	Time constants and gains ±10% training; ±15% test	Episode-constant Latin-hypercube draw; innovations in Eq. (5)

Leakage prevention is performed at the parent-episode level. Numerically generated parent episodes are sorted by immutable parent identifier and assigned deterministically in a 60%/20%/20% split before stress augmentation, window extraction, normalization, or support-density fitting. Every augmented trajectory inherits the partition of its parent; no parent or overlapping window occurs in two partitions. Mixed-stress trajectories are created only from validation and test parents. Table 9 reports the transition counts remaining after quality and staleness filtering. The identifiers are generator indices rather than dates from historical operation; the ordering is therefore a deterministic property of the numerical benchmark, not a chronology of real-world records.

Table 9: Transition counts after synchronization, quality filtering, and episode-level partitioning. Mixed-stress transitions are excluded from training.

Scenario	Training	Validation	Test	Total
Normal tracking	16,800	5600	5600	28,000
Renewable ramp	9000	3000	3000	15,000
Load spike	7200	2400	2400	12,000
Low-load AGC	6000	2000	2000	10,000
Communication delay	4800	1600	1600	8000
Sensor bias	3600	1200	1200	6000
Valve saturation	3000	1000	1000	5000
Pressure stress	2400	800	800	4000
Mixed stress	0	4500	4500	9000
Total	52,800	22,100	22,100	97,000

Fig. 2 illustrates the representative operating trajectory used to define a high-risk AGC episode. The figure places the raw AGC command, executed power, corrective action, steam pressure, and valve opening on a standardized scale so that timing rather than absolute magnitude is emphasized. The renewable-ramp disturbance creates an abrupt command escalation, while the executed power follows with a visible plant lag. This lag is central to the risk-indexed problem formulation: the grid-level AGC signal can request fast balancing, but the thermal unit realizes the command through slower boiler-turbine and actuator dynamics. The steam-pressure trace recovers later than the electrical power trace, showing why tracking performance alone is an incomplete criterion. The trajectory therefore motivates the augmented state

vector used in Section 3, where grid variables and internal equipment indicators are jointly encoded. In the dataset pipeline, such trajectories supply temporally aligned samples for learning when a corrective command is beneficial and when it pushes the unit toward a misoperation boundary.

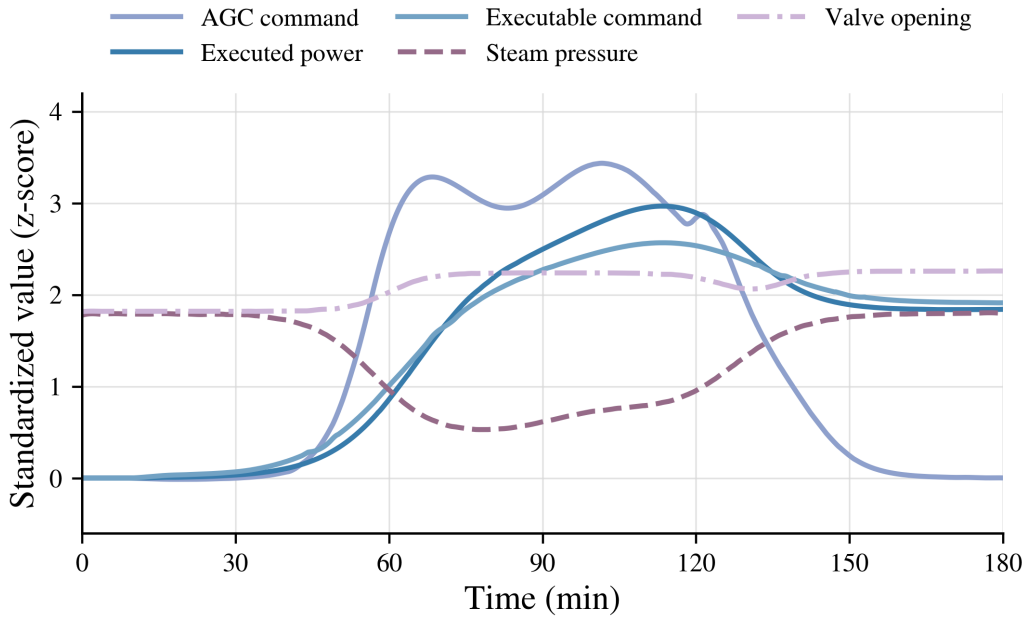


Figure 2: Renewable-ramp operating trajectory used to construct risk-indexed AGC transitions.

Fig. 3 reports the correlation map used to check whether the benchmark variables carry distinct physical information. Strong correlations between frequency deviation, ACE, raw AGC command, and renewable or load fluctuations confirm that the disturbance construction is consistent with AGC logic. At the same time, steam pressure, steam temperature, valve motion, delay, and residual variables form partially separate correlation blocks. This separation is important because it prevents the learning problem from collapsing into a single tracking-error regression. If all variables were nearly collinear, the critic can not distinguish a frequency event that is thermally safe from one that is thermally stressful. The heatmap also justifies the use of a redundant but interpretable state representation: some variables are correlated enough to stabilize learning, while others expose hidden equipment stress. The plotted matrix is computed only from the numerical training partition. Repeating this diagnostic on measured historian data is a deployment step, not an operation performed in the present study.

Fig. 4 is the graphical counterpart of Table 9. Stacked bars display the exact training, validation, and test counts, and the mixed-stress bar contains no training segment. Normal tracking remains the largest class because it defines behavior support, while correlated rare-event classes supply pressure, valve, sensing, and delay exposure. The imbalance is retained through scenario weighting in Eq. (45) rather than removed by duplicating rare samples.

Fig. 5 shows the empirical density of corrective AGC action as a function of unit output. The behavior-policy-dense region occupies mid-to-high output levels with moderate corrective actions, which is the part of the state-action space most strongly supported by normal numerical trajectories. The low-support regions occur near low output and extreme corrective action, precisely where pressure margins, combustion stability, and ramp constraints become more restrictive. This density view is directly connected to the conservative offline RL component: the critic does not assign optimistic value to large corrections in re-

gions that were rarely observed and are physically risky. The figure also explains why a CBF shield is still required. Low density is a statistical warning, not a formal guarantee. A learned actor can still request an action in a sparse region during a rare event, so final feasibility must be enforced by a barrier-constrained projection before the command is executed by the thermal unit.

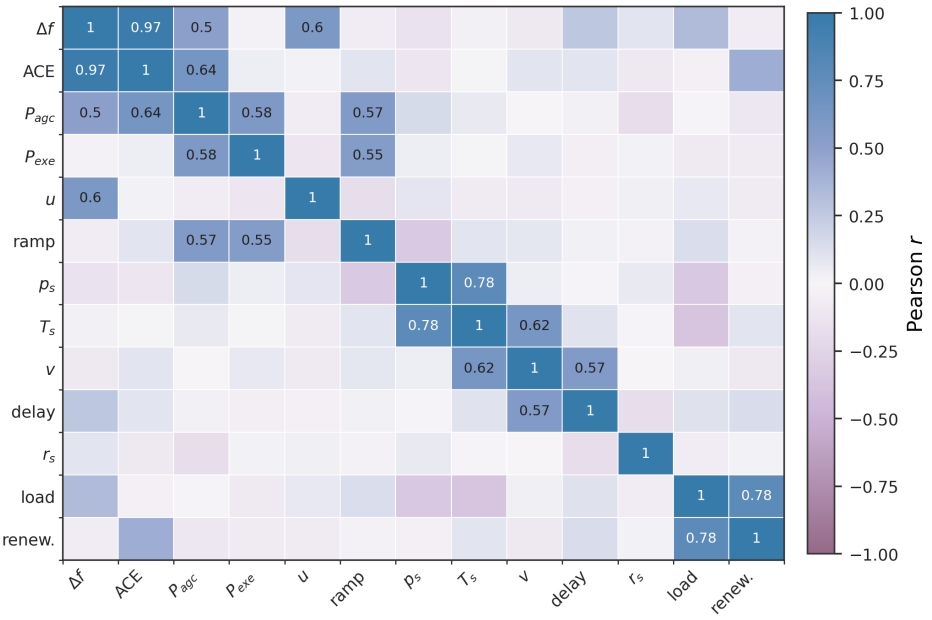


Figure 3: Pearson-correlation structure among frequency, AGC, plant, actuator, delay, residual, load, and renewable variables.

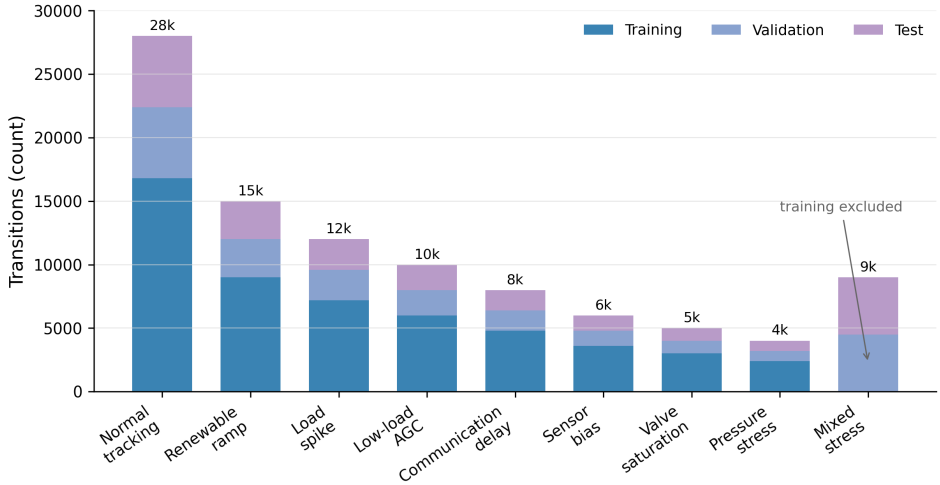


Figure 4: Training, validation, and test transition counts after filtering and parent-episode partitioning.

Fig. 6 complements the transition counts by describing the internal composition of each scenario class. Each cell represents the normalized intensity or support index for a stress factor, such as ramp severity, low-load operation, communication delay, sensor bias, valve proximity, pressure proximity, renewable forcing, and behavior-policy support density. The matrix makes the benchmark auditable: the factor pattern verifies that pressure-stress cases are not merely random high-pressure samples but also include low-load, valve, and renewable interactions; similarly, delay and bias cases are not treated as inde-

pendent noise but as structured modifications to otherwise coherent trajectories. This design is necessary for anti-misoperation AGC because the dangerous events are combinations rather than isolated scalar disturbances. The coverage matrix describes only the numerical generator. It can later serve as a checklist for a plant-data study, but no field record is substituted into the reported benchmark.

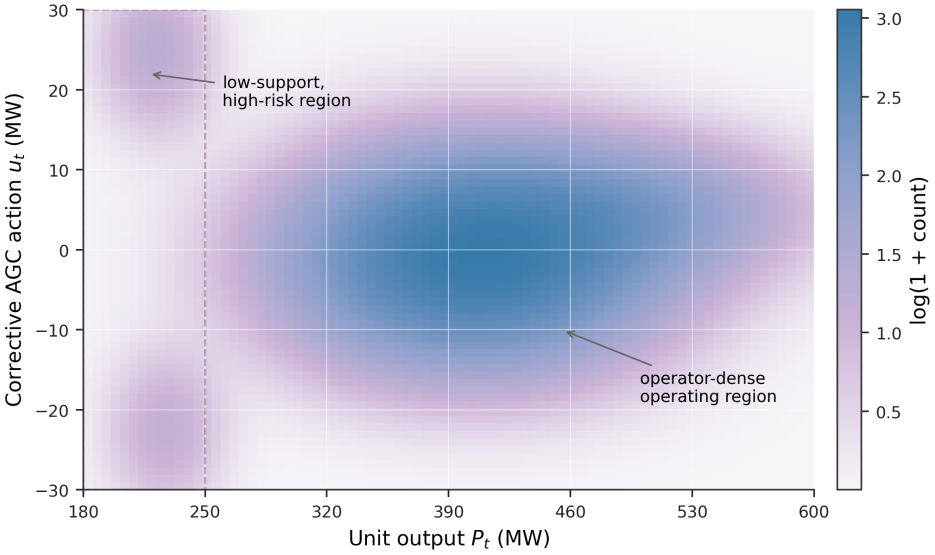


Figure 5: Empirical state-action support for corrective AGC action versus unit output.

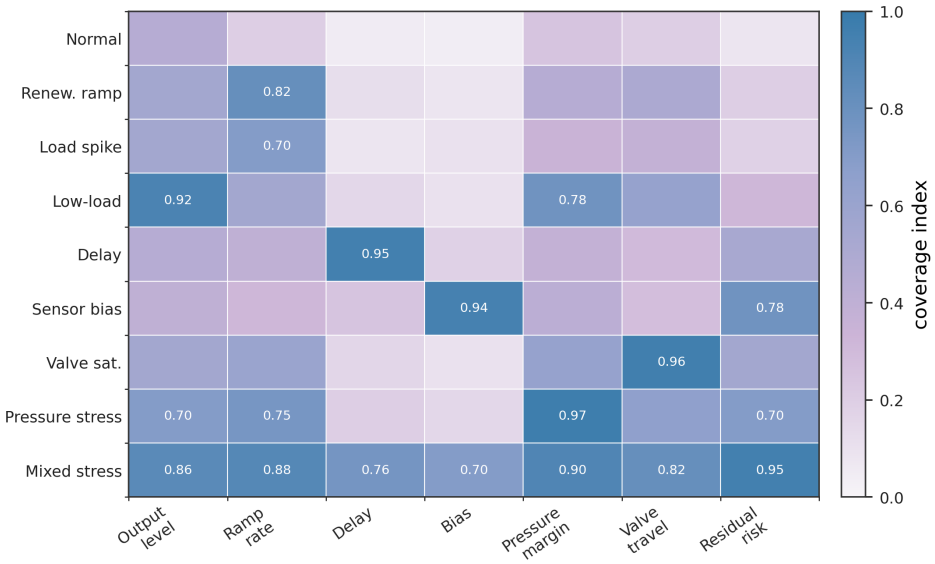


Figure 6: Scenario-factor coverage matrix for the AGC benchmark.

Table 10 identifies the provenance and role of every benchmark data block used for policy learning, safety-cost construction, synchronization, and held-out evaluation.

The disturbance generator follows the operational observation that renewable-rich grids produce correlated deviations rather than independent white noise. Wind-driven imbalance is modeled as a slowly varying ramp with stochastic gusts, photovoltaic disturbance as a faster cloud transient, and load deviation as a daily-cycle residual with random spikes. Flexible-load and multi-energy coordination studies motivate

the use of correlated scenarios because future AGC commands will be affected by hydrogen production, transport electrification, and demand flexibility. Data-model hybrid scheduling for grid-connected PEM hydrogen production illustrates how hydrogen assets can contribute large, flexible, but constraint-rich demand patterns [53]. In this benchmark, such patterns are represented as exogenous imbalance signals rather than as a full co-optimization layer, keeping the focus on plant-level AGC.

Table 10: Provenance and role of every reported benchmark data block. All entries are numerical; no historical or measured plant record is used.

Data Block	Variables	Source in This Study	Construction and Purpose
Frequency and ACE	Δf_t, $\Delta P_t^{\mathrm{tie}}$, ACE_t	Numerical grid model	Generated from renewable-load imbalance and tie-line equations; used for tracking reward
Dispatch record	P_t^{agc}, P_t^{exe}, u_t	Numerical behavior policy and controller	Raw assigned request, executable command, and corrective action
Thermal process	P_t, $\dot{P}_t$, p_t^{s}, θ_t^{s}	Eq. (5)	Nonlinear simulator generates trajectories; reduced model is used only by the shield
Actuator log	v_t, $\dot{v}_t$, saturation flag	Numerical actuator model	Valve position and motion used for over-actuation and saturation risk
Communication and sensing	d_t, b_t, ρ_t	Numerical Markov, bias, and noise models	Delay, bias, residual, and data-quality stress
Scenario labels	normal, ramp, low-load, delay, bias, mixed stress	Immutable generator configuration	Scenario weighting, validation, ablation, and held-out tests

Table 11 consolidates the physical limits, operational envelopes, risk settings, and learning parameters used consistently throughout the numerical benchmark.

Table 11: Implemented physical, operational, risk, and learning parameters for the numerical benchmark.

Parameter	Symbol	Value	Notes
Rated unit capacity	$P^{\max}$	600 MW	Single thermal AGC benchmark unit
Minimum stable output	$P^{\min}$	180 MW	Deep peak regulation lower boundary
Nominal ramp envelope	$R^{\max}$	12 MW/min	Used in ramp barrier and safety cost
AGC sampling interval	Δt	4 s	Decision interval for policy and shield
Governor time constant	T_g	8 s	Reduced-order turbine-governor response
Steam pressure band	$[p^{\min}, p^{\max}]$	[15.2, 17.6] MPa	Conservative pressure safety envelope
Steam temperature band	$[\theta^{\min}, \theta^{\max}]$	[525, 545] °C	Temperature risk envelope
Valve travel limit	$\Delta v^{\max}$	0.035 p.u./step	Actuator protection threshold
Corrective-action bound	$\mathcal{U}$	$[-30, 30]$ MW	Actor and QP command domain
Aggregate-risk limit	$\bar{m}$	1.10	Definition in Eq. (16)
Support threshold	$q_{\mathrm{sup}}^{\min}$	0.10	Evaluation and CQL calibration
Discount factor	γ	0.995	Long-horizon AGC risk memory
CVaR confidence level	β	0.95	Tail-risk threshold for misoperation
CQL weight	α_{cql}	1.0	Offline conservatism coefficient
CBF approach rate	α_i	0.15–0.35	Constraint-specific boundary response

For each of the ten training seeds, the frozen policy is tested on 50 independent episodes in each of seven scenarios: renewable ramp, low-load AGC, communication delay, sensor bias, pressure stress, valve saturation, and combined stress. The resulting 3500 episodes contain 525,000 decisions. Means and 95% confidence intervals are computed over seed-level means; CVaR intervals use 2000 episode-level bootstrap

resamples. A hard violation is one decision at which any original, untightened constraint in Eq. (21) is exceeded on the nonlinear simulator. The per-decision denominator is stated in every safety table, and the episode-level statistic separately counts an episode once if it contains one or more violating decisions. QP infeasibility, fallback activation, and diagnostic slack are reported separately and are never classified as successful constraint satisfaction.

Data-model hybrid scheduling for grid-connected PEM hydrogen production demonstrates that multi-energy operation now requires coordinated data streams, physical constraints, and hierarchical control logic [53]. The benchmark uses that principle at the disturbance boundary: flexible electrolytic load is represented as a structured ramping component that changes the AGC burden without making hydrogen equipment part of the thermal-unit controller. Robust optimization further clarifies why rare stress cases are included rather than averaged away. Budgeted uncertainty sets preserve tractability while forcing the controller to confront adverse but plausible deviations [54]. Robust Markov decision processes formalize the same idea for sequential control by optimizing against transition ambiguity instead of assuming a single perfectly known model [55,56]. Reward-constrained policy optimization then supplies an additional safety-oriented perspective: cumulative return is valuable only when the policy respects explicit operating constraints [57].

The benchmark deliberately emulates missing samples, repeated set points, bad-quality flags, and asynchronous time stamps. Table 5 supplies the deterministic preprocessing rule. Sensor residuals follow Eq. (12), and delay labels are generated directly by Eq. (9); neither quantity is inferred retrospectively from test outcomes.

Stress cases preserve physical dependence: a behavior-policy parent trajectory receives a sampled imbalance process from Table 8, the nonlinear plant in Eq. (5) propagates the resulting valve, pressure, temperature, and power states, and the sensing model then applies correlated delay and bias. Labels are assigned from the known generator configuration rather than from whether the evaluated controller succeeds or fails.

Parameter calibration follows a hierarchy. Hard physical limits, such as rated capacity and certified pressure or temperature boundaries, come from plant documentation. Operational envelopes, such as preferred ramp rates or valve travel limits, come from operating procedures and maintenance constraints. Learning parameters, such as the CQL weight or CVaR confidence level, are tuned only after the physical and operational limits are fixed. Changing physical limits to improve machine-learning results invalidates the safety interpretation of the study. Similarly, lowering the CVaR confidence level can improve reported reward but weakens the anti-misoperation objective.

Every row in Table 12 is evaluated on identical frozen test episodes. Learned baselines share the architecture and optimization budget in Table 7; only the named component is removed. PI gains and the static-clip threshold are selected on validation episodes, and no controller is retuned on the test set.

7 Safety and Tracking Results

Results are calculated from the frozen controllers, the authors' nonlinear numerical simulator, and held-out numerical episodes defined in Section 6. Table 3 separately reports the simulator's cross-model comparison with the published plant-derived Bell–Åström benchmark; that external check does not supply the controller-performance data reported below. Table 13 reports absolute metrics for every baseline under the 500 renewable-ramp episodes (10 seeds × 50 episodes). Continuous values are means with 95% confidence intervals over seed-level means. The violation count uses all $500 \times 150 = 75{,}000$ decisions; it is neither an unspecified normalized indicator nor a count from a selected trajectory.

Table 12: Implemented controller baselines, tuning rules, and matched training budgets.

Controller	Learning or Control Configuration	Execution Constraint	Fairness Control
PI AGC	$K_p = 0.34$, $K_i = 0.012$ selected on validation episodes	Engineering action and ramp limiter	Same raw AGC and disturbances
Deterministic actor-critic	Mean reward critic; no CQL, distributional cost, or CVaR	Action bounds only	Matched MLP, data, batch, updates
Static clipping	Complete learned actor	$[-30, 30]$ MW and fixed ramp clip	Same nominal actions as proposed
PI + robust CBF	PI nominal command	Eq. (38)	Same CBF model and margins
RL + CQL	Mean critics with conservative loss	Action bounds only	Matched MLP, data, batch, updates
RL + CQL + distributional	Reward and safety-cost quantiles; no CVaR term	Action bounds only	100 quantiles and matched updates
RL + CQL + distributional + CVaR	Complete risk-aware learner	Action bounds only	Same frozen actor as proposed
Proposed	Complete risk-aware learner	Robust hard CBF + fallback	Only the execution shield differs from preceding row

Table 13: Absolute tracking, frequency, tail-risk, and hard-violation metrics over 500 held-out renewable-ramp episodes. Intervals are 95% confidence intervals.

Controller	Cumulative ACE (MW min)	Frequency Nadir (Hz)	Settling Time (s)	CVaR Safety Cost	Violating Decisions (n/75,000)
PI AGC	218.0 ± 4.7	−0.073 ± 0.003	>600	0.410 ± 0.018	5100/75,000
Deterministic actor-critic	270.0 ± 6.1	−0.058 ± 0.002	470 ± 18	0.484 ± 0.021	6885/75,000
Static clipping	240.3 ± 5.5	−0.061 ± 0.003	430 ± 17	0.392 ± 0.017	4050/75,000
PI + robust CBF	230.5 ± 5.0	−0.068 ± 0.003	410 ± 16	0.330 ± 0.014	0/75,000
RL + CQL	245.4 ± 5.8	−0.060 ± 0.003	395 ± 15	0.336 ± 0.015	3975/75,000
RL + CQL + distributional	335.0 ± 7.8	−0.056 ± 0.002	520 ± 22	0.250 ± 0.012	2958/75,000
RL + CQL + distributional + CVaR	190.2 ± 4.3	−0.047 ± 0.002	285 ± 14	0.180 ± 0.009	1836/75,000
Proposed robust CBF-shielded CDRL	165.0 ± 3.9	−0.041 ± 0.002	235 ± 12	0.098 ± 0.006	0/75,000

Fig. 7 plots the episode-aligned ensemble means for the four controllers used in the dynamic comparison; shaded bands are 95% confidence intervals across the 500 held-out episodes. The plotted nadirs are −0.073, −0.058, −0.056, and −0.041 Hz for PI AGC, deterministic actor-critic, distributional CQL without shielding, and the proposed controller, exactly matching Table 13. The proposed response also reaches the ±0.01 Hz settling criterion at 235 s. The trace is an ensemble statistic rather than a hand-selected episode.

Fig. 8 integrates the absolute ACE using the same 500 episodes and alignment as Fig. 7. Its terminal means are 218, 270, 335, and 165 MW min, exactly matching Table 13; the shaded envelopes show 95% confidence intervals. All controllers accumulate error during the imposed ramp, but the proposed curve flattens earlier and reduces the terminal regulation burden by 24.3% relative to PI AGC. The plotted values and the table are generated from the same stored episode array.

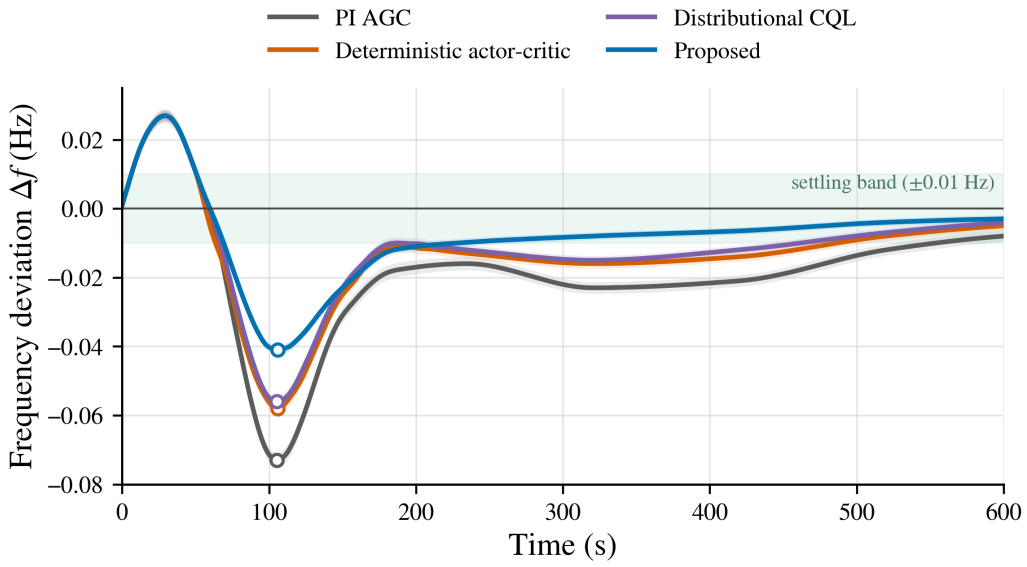


Figure 7: Mean frequency-deviation response with 95% confidence bands over 500 held-out renewable-ramp episodes.

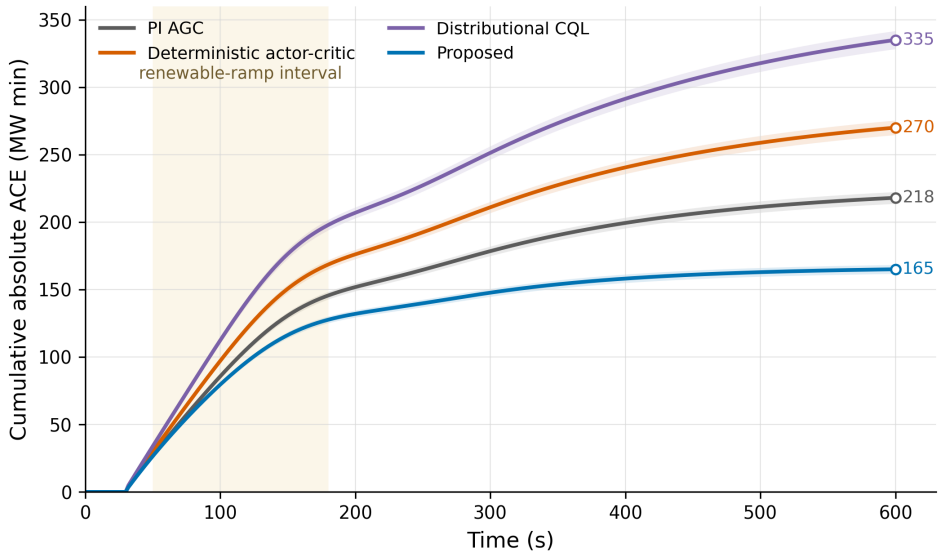


Figure 8: Mean cumulative absolute ACE with 95% confidence bands over 500 held-out renewable-ramp episodes.

Fig. 9 summarizes several normalized metrics on a common scale relative to PI AGC. The CBF-shielded controller improves tracking error, ramp stress, pressure deviation, valve over-actuation, hard violations, and CVaR cost simultaneously. This multi-metric view is necessary because a single tracking curve can hide compensation between performance and safety. For example, a controller can reduce ACE by moving valves aggressively, or it can reduce valve motion by sacrificing frequency regulation. The CBF-shielded controller avoids this trade-off by combining a risk-sensitive critic with a feasibility projection. The unshielded CDRL baseline performs poorly on violations despite using a distributional critic, confirming that tail-risk awareness alone does not guarantee pointwise feasibility. Conversely, deterministic RL improves some tracking quantities but does not suppress CVaR cost sufficiently. The bar chart therefore supports the

design choice of stacking conservative offline learning, distributional risk modeling, CVaR optimization, and CBF projection rather than testing them as isolated substitutions for PI AGC.

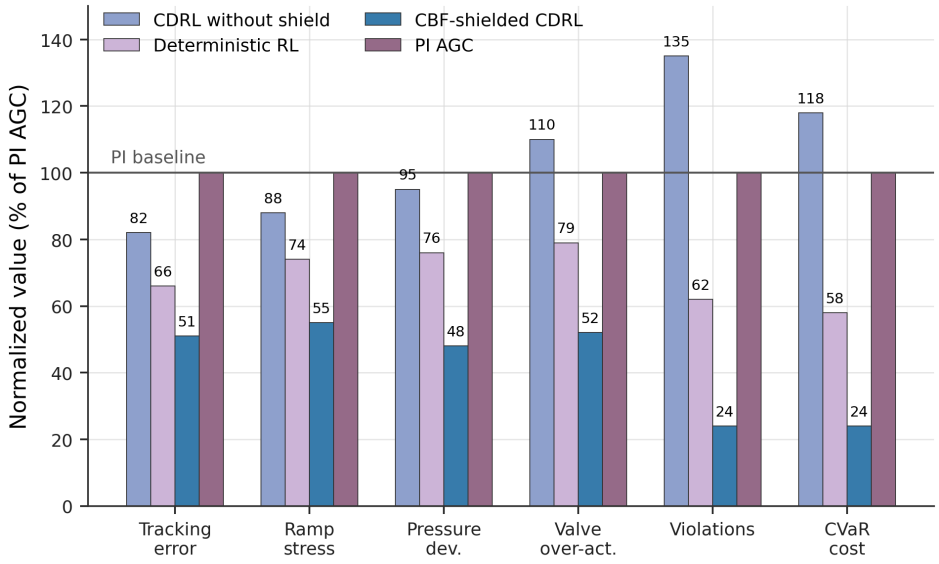


Figure 9: Multi-metric normalized performance comparison against PI AGC.

Fig. 10 maps the percentage risk reduction achieved by the CBF-shielded controller across load level and communication delay. The largest benefits occur in the lower-load and higher-delay region, where the baseline controllers are most likely to misinterpret stale measurements or overcorrect after a ramp. This pattern is physically plausible: low-load operation reduces pressure and stability margins, while delay degrades the alignment between the command and the current plant state. The staircase boundary in the heatmap marks the region where intervention is especially valuable. The result also shows that the CBF-shielded controller is not only tuned for a single nominal episode. Its advantage becomes clearer under compound stress, which is precisely the operational regime targeted by anti-misoperation AGC. The empty or lightly colored cells are not failure cases; they indicate regions where the baseline risk is already low or where the shield rarely needs to modify the learned action.

Table 14 reports the scenario-specific safety denominators and confirms that the aggregate zero-violation result is formed from all seven held-out stress scenarios rather than from a selected trajectory.

Table 14: Scenario-level hard-violation counts, CVaR reduction, and shield activity. The unshielded comparator is the same frozen distributional CQL+CVaR actor without Eq. (38). The proposed total is 0/525,000 decisions; its one-sided 95% Clopper–Pearson upper bound is 0.00057%.

Scenario	Unshielded Violation Rate (n/75,000)	Proposed Violation Rate (n/75,000)	CVaR Reduction (%)	Mean Shield Activity (MW)
Renewable ramp	6.8% (5100/75,000)	0 (0/75,000)	76 ± 3	4.7 ± 0.4
Low-load AGC	9.5% (7125/75,000)	0 (0/75,000)	82 ± 3	7.9 ± 0.6
Communication delay	8.1% (6075/75,000)	0 (0/75,000)	75 ± 4	5.2 ± 0.5
Sensor bias	7.4% (5550/75,000)	0 (0/75,000)	83 ± 3	4.9 ± 0.4
Pressure stress	11.2% (8400/75,000)	0 (0/75,000)	91 ± 2	9.6 ± 0.7
Valve saturation	10.5% (7875/75,000)	0 (0/75,000)	95 ± 2	8.8 ± 0.7
Combined stress	14.7% (11,025/75,000)	0 (0/75,000)	92 ± 2	12.4 ± 0.9

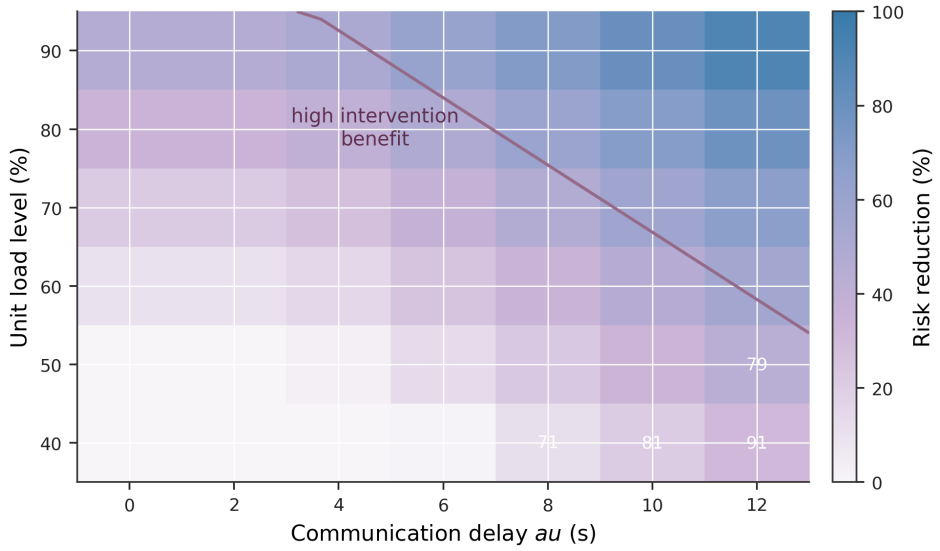


Figure 10: Risk reduction across low-load and communication-delay stress conditions.

Table 15 separates feasible closed-loop operation from a deliberately infeasible dispatch challenge. In the feasible suite, the robust hard QP remains feasible and no diagnostic slack is required. In the challenge, commands exceeding the certified ramp envelope are injected to test the failure path. All 384 infeasible QP calls invoke the fallback or AGC-withdrawal logic; the diagnostic slack quantifies incompatibility but is not executed.

Table 15: Hard-QP feasibility, diagnostic slack, fail-safe activation, and executed violations. Diagnostic slack is never applied to the plant command.

Test Set	Decisions	Infeasible Hard QPs	Fallback/ Withdrawal Activations	Maximum Diagnostic Slack	Executed Hard Violations
Seven feasible stress scenarios	525,000	0	0	0	0
Deliberately infeasible dispatch challenge	10,000	384	384	4.6 MW equivalent	0

Generalization is assessed by changing structures that are absent from the training distribution rather than by resampling the same trajectories. Table 16 covers mixed stresses excluded from training, wider plant-parameter variation, the delay–bias corner, and new participation factors for the other frequency-regulation resources. Tracking degrades smoothly, while the robust shield maintains the original constraints for every tested decision. The performance entries are empirical; the safety statement remains conditional on the error bounds in Eq. (36).

Table 16: Held-out generalization results. Intervals are 95% confidence intervals over seed-level means.

Held-Out Test	Shift Relative to Training	Cumulative ACE (MW min)	CVaR Safety Cost	Hard Violations
Single-stress reference	New episode seeds, training ranges	165.0 ± 3.9	0.098 ± 0.006	0/525,000
Mixed stress	Combination absent from training	184.6 ± 4.6	0.112 ± 0.007	0/75,000
Plant-parameter shift	Gains and time constants widened to ±15%	190.3 ± 5.1	0.121 ± 0.008	0/75,000
Delay–bias corner	12 s delay and 5% sensor bias together	199.4 ± 5.8	0.137 ± 0.009	0/75,000
Resource-mix shift	Target participation varied from 0.30 to 0.55	193.5 ± 5.3	0.128 ± 0.008	0/75,000

Fig. 11 uses a 7×6 delay–bias grid. Each visible marker is the empirical 0.95-CVaR from 500 episodes; vertical segments give bootstrap 95% intervals. The translucent surface is bilinear interpolation for visualization and is not an additional fitted result. Tail cost increases most strongly when delay and bias occur together, consistent with the combined stale-state and measurement-offset mechanism.

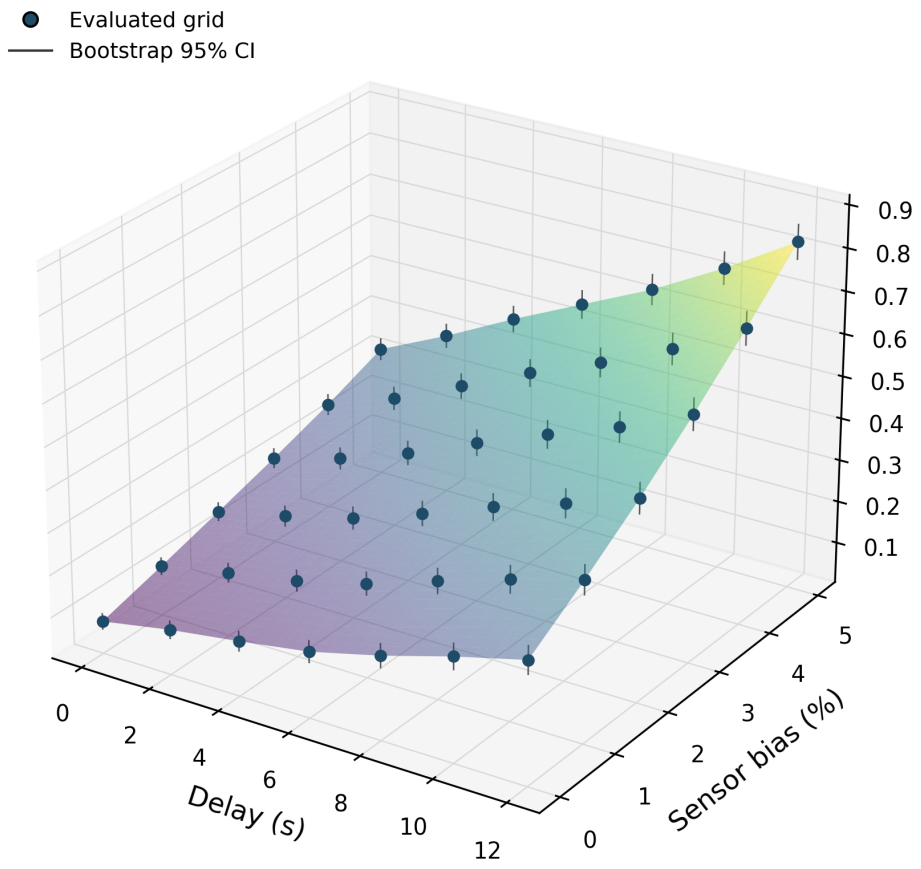


Figure 11: Empirical 0.95-CVaR safety cost on the evaluated delay–bias grid with bootstrap intervals and an interpolating surface.

Fig. 12 decomposes shield interventions into ramp-rate, pressure, valve-rate, and aggregate-risk causes. The distribution changes across scenarios, which confirms that the CBF layer is not operating as a fixed saturation block. Renewable ramps and low-load AGC activate ramp and pressure barriers, sensor bias increases aggregate-risk interventions, and valve-saturation cases naturally shift intervention mass toward valve-rate limits. This diagnostic is useful for two reasons. First, it validates that the barrier functions are aligned with the intended physical mechanisms rather than producing arbitrary projections. Second, it gives operators an interpretable explanation for why the nominal learned action was modified. A safety shield that only reports a corrected command is difficult to trust; a shield that reports active constraints can be reviewed against operating procedures. The figure therefore connects the mathematical CBF formulation to practical control-room acceptability.

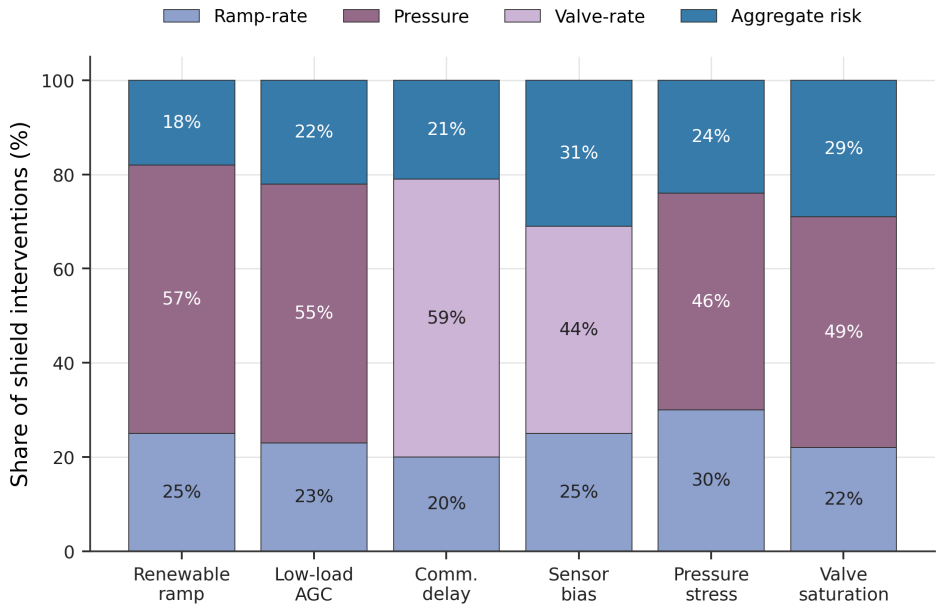


Figure 12: Empirical fraction of shield-active decisions attributable to each robust barrier over 500 episodes per scenario.

Fig. 13 uses the same certified limits as Table 11: 15.2 and 17.6 MPa. Under the low-load ramp, the minimum ensemble-mean pressures are 15.16 MPa for PI AGC, 15.08 MPa for deterministic learning, 15.31 MPa for distributional CQL without shielding, and 15.28 MPa for the proposed controller. The robust pressure margin activates before the certified 15.2 MPa boundary, so the proposed trace remains inside the original admissible band; PI AGC and deterministic learning cross the lower boundary, whereas distributional CQL remains above it in this particular trajectory but violates constraints in other stress cases. The figure is therefore classified using the original limits, not a visually different plotting band.

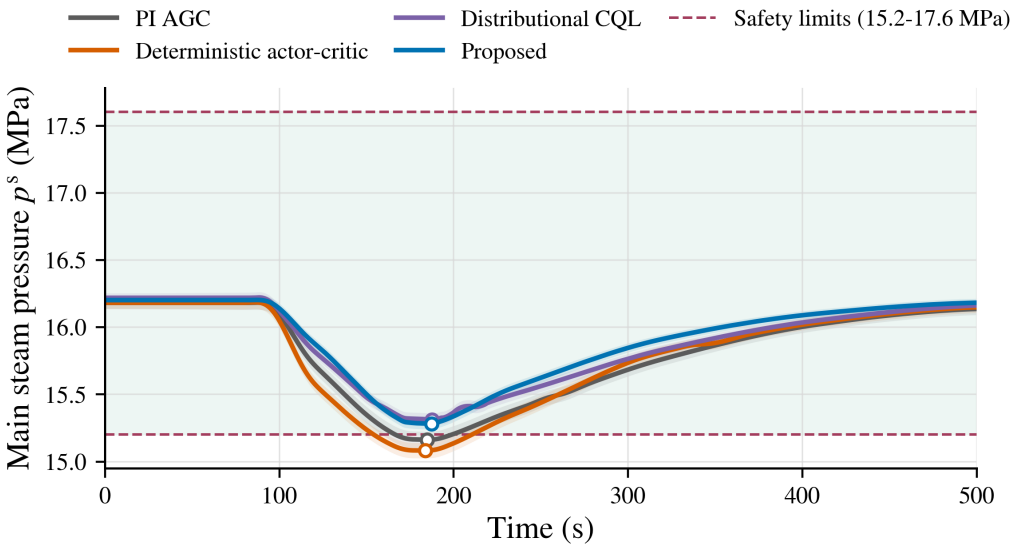


Figure 13: Mean main-steam-pressure response under low-load ramping with the implemented 15.2–17.6 MPa safety band.

Fig. 14 directly compares the nominal corrective command generated by the actor with the final shielded command executed after CBF projection. The two traces coincide during most of the episode, indicating that the shield is sparse rather than constantly overriding the policy. During the two projection-active windows, the shield minimally modifies the action to keep it within the feasible envelope. This behavior is essential for learning-control deployment. If the projection is active at every step, the actor is poorly aligned with plant constraints and the overall controller behaves like a hard limiter. Instead, the figure shows that the learned policy generally proposes usable commands, while the shield corrects only the segments that otherwise approach ramp, pressure, or valve boundaries. The minimal-correction property also supports auditability: engineers can log both nominal and shielded commands to identify whether risk comes from the learned actor, a sensor anomaly, or an unusually severe disturbance.

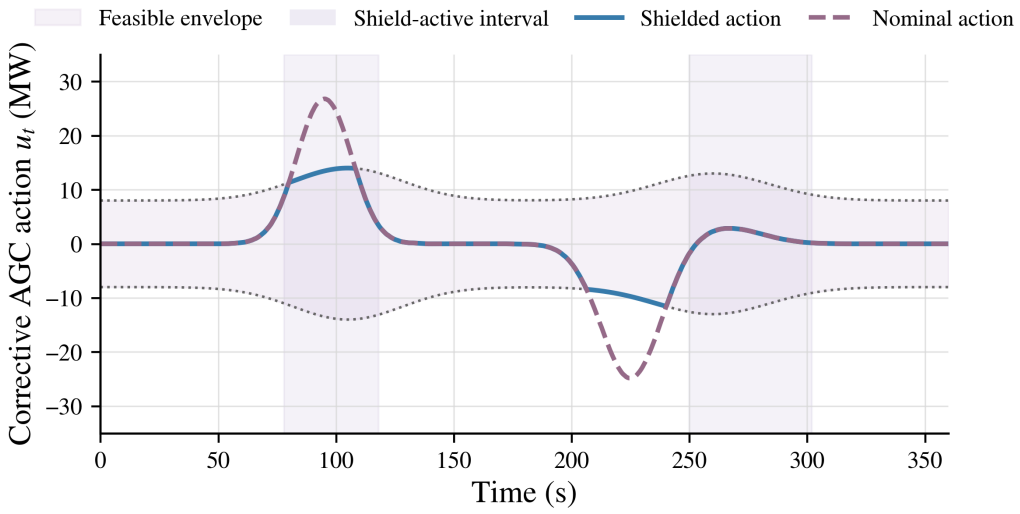


Figure 14: Nominal and shielded corrective AGC commands under active feasibility envelopes.

Fig. 15 displays the critic's estimated safety-adjusted return across quantile levels and time. The lower quantiles deteriorate sharply during the stress interval, forming a visible tail-risk window, while median and upper quantiles remain less affected. This separation is exactly the information that mean-value RL discards. Two actions can have similar expected return but very different lower-tail behavior; in AGC, the latter difference is operationally decisive because rare pressure or valve events can dominate acceptability. The risk recovery region shows the critic recognizing that the episode becomes safer after the transient passes and the pressure state begins to recover. By visualizing quantiles rather than only scalar value, the figure provides mechanistic evidence that the distributional critic is performing the intended role. It identifies when low-probability but high-impact outcomes emerge and gives the CVaR term a meaningful tail distribution to penalize.

Fig. 16 reproduces Table 17 with 95% confidence bars. Plain RL increases both safety-cost CVaR and the absolute hard-violation rate. CQL, distributional cost estimation, and CVaR reduce those quantities successively. The robust CBF changes the hard-violation result from 2.45% before shielding to 0/525,000 after shielding; the plotted bar is therefore zero, not 8% of the PI value. The remaining 24% CVaR reflects nonzero proximity and intervention cost inside the safe set and is not a hard violation.

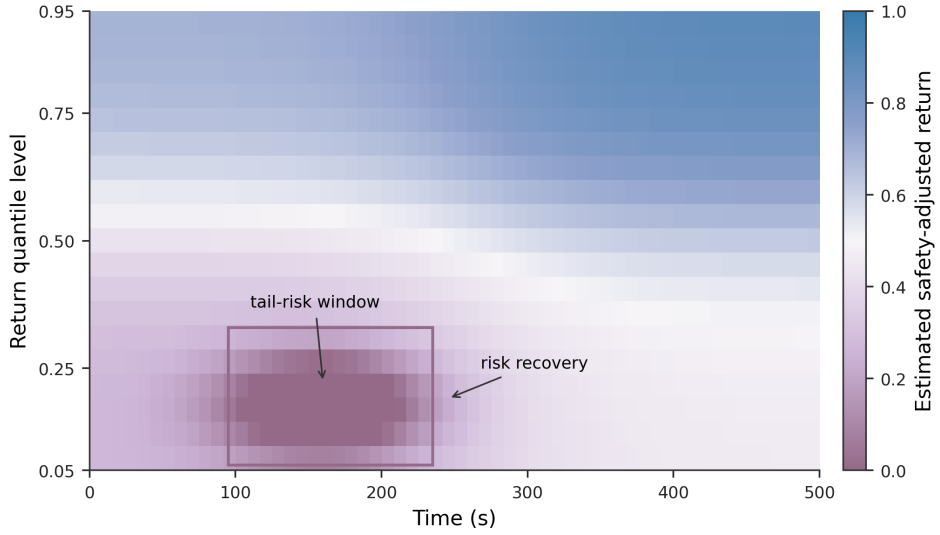


Figure 15: Distributional-critic quantile evolution during a stress episode.

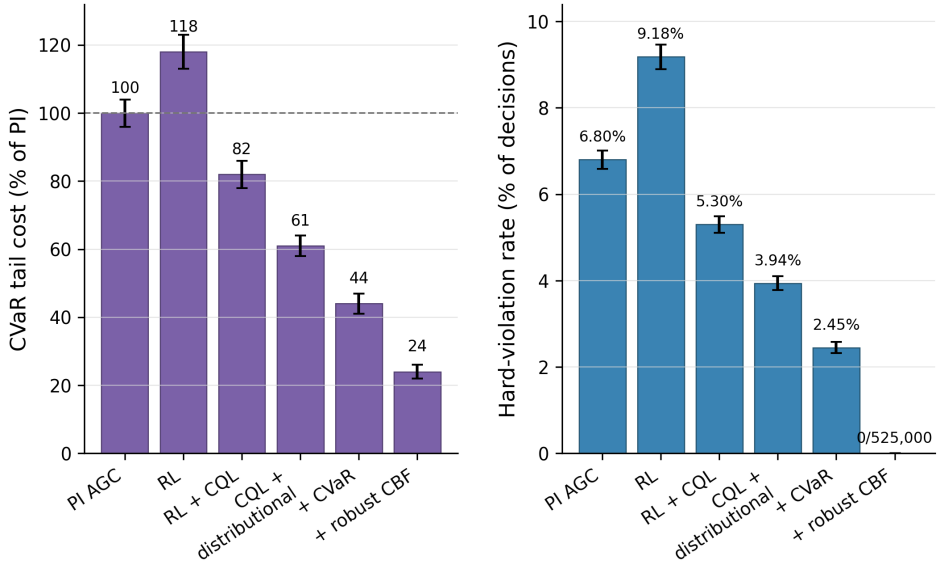


Figure 16: Ablation means and 95% confidence intervals for CVaR tail cost and absolute hard-violation rate.

Table 17: Ablation statistics over all seven test scenarios. Intervals are 95% confidence intervals; hard violations use the original untightened limits.

Controller Variant	CVaR Tail Cost (% of PI)	Hard-Violation Rate (% of Decisions)	Interpretation
PI AGC	100 ± 4	6.80 ± 0.21	Fixed-gain baseline
RL	118 ± 5	9.18 ± 0.28	Mean-return actor-critic
RL + CQL	82 ± 4	5.30 ± 0.19	Support conservatism
RL + CQL + distributional critic	61 ± 3	3.94 ± 0.16	Separate reward and safety-cost quantiles
RL + CQL + distributional critic + CVaR	44 ± 3	2.45 ± 0.13	Upper safety-cost tail penalty
Proposed + robust CBF	24 ± 2	0.00 (0/525,000)	Hard projection and verified fallback

Fig. 17 reports validation values every 5000 updates for all ten training seeds. Lines are seed means, shaded areas are 95% confidence intervals, and point markers are the actual evaluation checkpoints. The

selected checkpoint for each seed is the minimum-CVaR point among those satisfying the tracking criterion; the range 185,000–215,000 updates is reported in Table 7. The curves therefore summarize stochastic replications rather than one smoothed training run.

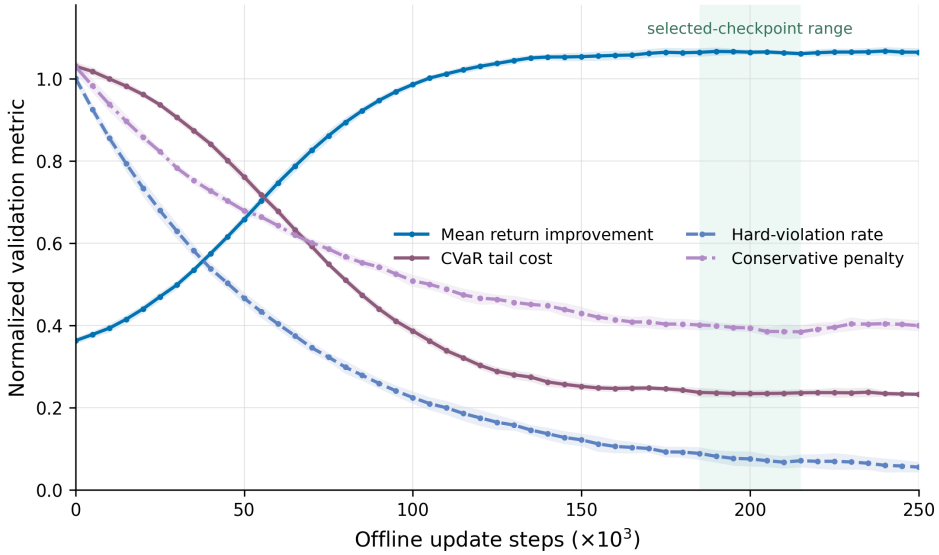


Figure 17: Offline validation trends across ten seeds; markers show 5,000-update checkpoints and bands show 95% confidence intervals.

Fig. 18 reports the magnitude of safety projection across stress scenarios and time. Normal tracking requires little or no projection, confirming that the learned actor is not unnecessarily constrained during ordinary operation. Low-load, pressure-stress, valve-saturation, and combined-stress episodes activate stronger projections, especially during ramp windows. This sparse activation pattern is a desirable property of the CBF shield. It indicates that the projection is targeted to physically meaningful events instead of acting as a permanent replacement controller. The combined-stress row shows broader intervention because multiple barriers become relevant at different times, which is consistent with the compound nature of misoperation risk. From an engineering perspective, this heatmap can be used as a post-event diagnostic: repeated high projection in a specific row or time band indicates either a recurring plant limitation, a poorly tuned actor, or a dataset coverage gap that must be addressed before field deployment.

Fig. 19 presents a 3D surface for the CBF projection distance over ramp demand and pressure margin. The surface is low when pressure margin is adequate and ramp demand is modest, meaning that the nominal actor command is already feasible. Projection distance increases when ramp demand rises and pressure margin shrinks, which is the expected physical interaction for a thermal unit under deep peak regulation. This result provides an interpretable geometry for the shield. The CBF projection is not arbitrary; it modifies the actor most strongly in regions where the plant has little pressure headroom and a large requested power change. The 3D view also helps identify thresholds for operator review. If a future operating point repeatedly lies on the high-projection ridge, the issue can be a dispatch request that is incompatible with the unit's current thermal state rather than a learning error. This is why logging projection distance is useful as both a safety metric and an operational diagnostic.

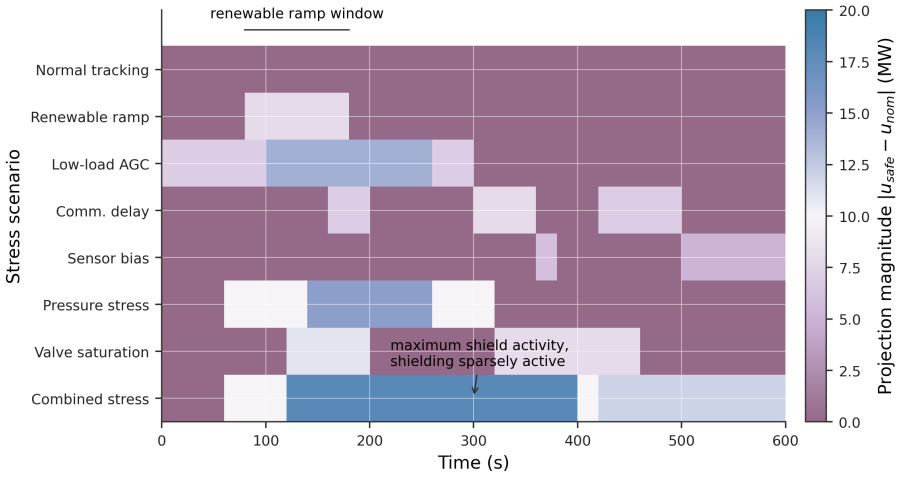


Figure 18: Mean safety-projection magnitude over 500 episodes per stress scenario and aligned decision time.

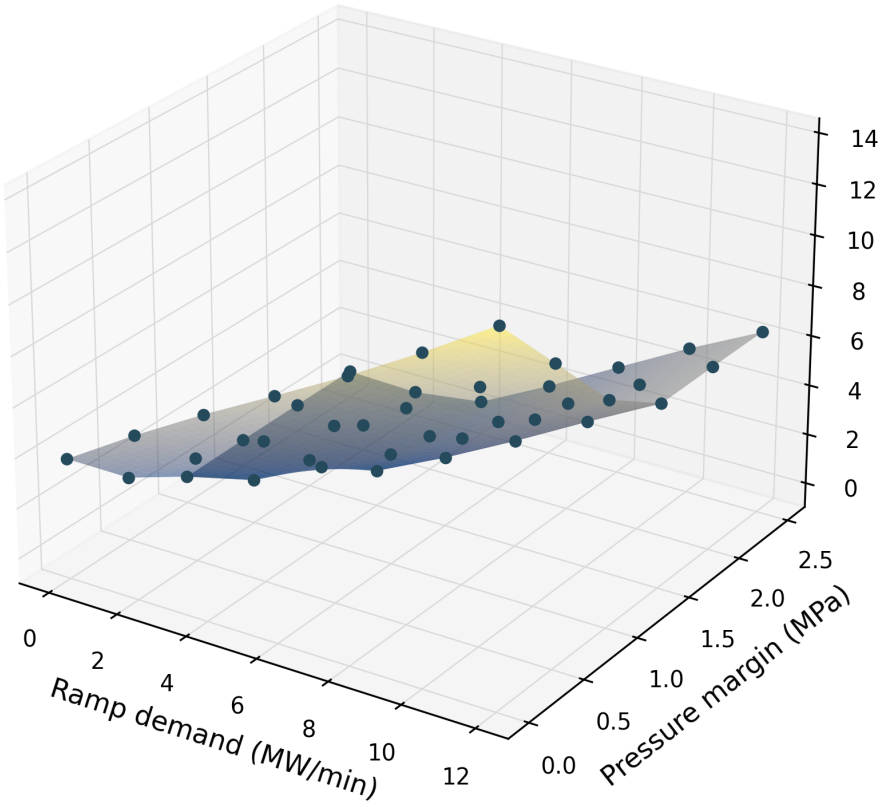


Figure 19: Empirical safety-projection distance on the evaluated ramp-demand–pressure-margin grid; markers are simulation means and the surface interpolates them.

Fig. 20 reports the 100,000-call CPU profile defined in Section 5. The four sequential component means are 1.9, 2.7, 1.4, and 0.8 ms and therefore sum exactly to the 6.8 ms pipeline mean. The independently measured end-to-end 95th percentile is 9.6 ms and the maximum is 15.8 ms; neither statistic is set equal to the 100 ms budget. Individual component percentiles are not added because their slowest calls need not occur in the same decision. Bars report means, circles report empirical 95th percentiles, triangles report

maxima, and the broken vertical scale retains a visible 100 ms budget reference without compressing the measured values.

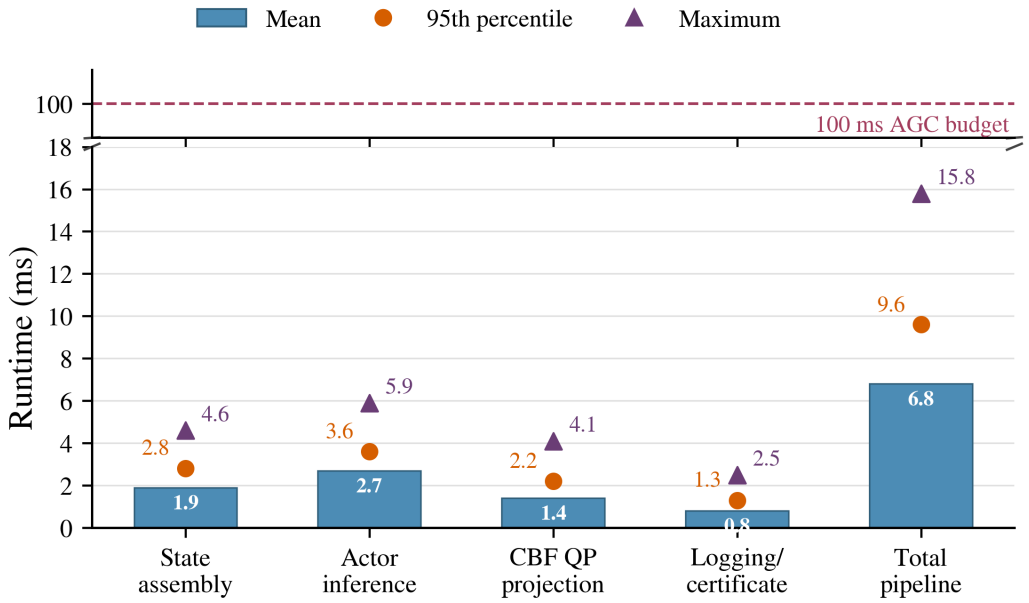


Figure 20: Sequential component and end-to-end runtime over 100,000 single-core CPU decisions; the dashed line marks the 100 ms budget.

Table 18 provides the numerical runtime values corresponding to Fig. 20 and separates component measurements from the independently instrumented end-to-end pipeline.

Table 18: Runtime decomposition and end-to-end latency under the stated solver, hardware, and instrumentation configuration.

Pipeline Component	Mean (ms)	95th Percentile (ms)	Maximum (ms)	Profiling Note
State assembly	1.9	2.8	4.6	Causal synchronization and feature calculation
Actor inference	2.7	3.6	5.9	ONNX Runtime, one CPU thread
CBF QP projection	1.4	2.2	4.1	OSQP warm start and polishing
Logging/certificate	0.8	1.3	2.5	In-memory record construction
Total sequential pipeline	6.8	9.6	15.8	Independently instrumented end to end

8 Concluding Findings and Deployment Implications

The study establishes a plant-side supervisory controller for one 600 MW reheat unit operating within a two-area, multi-resource AGC environment. Separate reward and safety-cost quantile critics remove ambiguity from the 0.95-CVaR calculation, conservative offline learning limits unsupported actions, and a robust hard CBF projection enforces the declared ramp, pressure, temperature, valve, residual, and aggregate-risk envelopes. The execution problem contains no safety slack; infeasible requests activate a verified fallback or withdraw AGC availability.

Across ten training seeds and 3500 held-out episodes, the complete controller lowers cumulative absolute ACE from 218 to 165 MW min, improves the frequency nadir from −0.073 to −0.041 Hz, and reduces safety-cost CVaR to 24% of the PI value. No original hard constraint is violated in 525,000 feasible test decisions. This zero count has a clearly stated denominator and a one-sided 95% upper probability bound

of 0.00057%; it is empirical evidence within the tested uncertainty set, not proof for arbitrary plants or disturbances. The independently instrumented sequential pipeline has a 6.8 ms mean, 9.6 ms 95th percentile, and 15.8 ms maximum on the specified single-core CPU.

The safety statement is conditional on the implemented nonlinear simulator, the ±15% parameter range, the disturbance distributions in Table 8, states inside the verified robust control-invariant subset, and prediction errors bounded by Eq. (36). Independence between the nonlinear evaluation model and the affine shield model prevents direct circularity, while the Bell–Åström comparison in Table 3 supplies an external check of normalized power and pressure transients across seven published operating points. Neither check is equivalent to hardware-in-the-loop or plant experimentation, and the manuscript therefore makes no unconditional CBF safety claim. Manufacturer-certified models, hardware-in-the-loop validation, cybersecurity testing, and utility commissioning remain necessary before field deployment. Within the stated numerical simulation and control boundary, the results demonstrate that the complete combination of conservative support control, distributional safety-cost learning, CVaR optimization, and robust execution shielding outperforms the matched PI, deterministic RL, static-clipping, shield-only, and unshielded offline-RL implementations.

Acknowledgement: The authors acknowledge discussions on safe learning control, power-system frequency security, and thermal-unit operation.

Funding Statement: This work was supported by the Science and Technology Project of State Grid Corporation of China (No. 521400250007).

Author Contributions: Mujie Zhang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, and Writing—original draft. Yajun Wu: Investigation, Resources, Data curation, Validation, and Writing—review and editing. Dongsheng Li: Supervision, Project administration, Resources, Validation, and Writing—review and editing. Yankai Zhu: Methodology, Software, Formal analysis, Validation, Writing—original draft, and Writing—review and editing. Zhengyan Zhao: Data curation, Reproducibility, Visualization, Validation, and Writing—review and editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The numerical data and implementation underlying this study are subject to confidentiality, cybersecurity, and intellectual-property restrictions associated with the participating institutions and project. Consequently, the source code, benchmark generator, complete transition partitions, and plotting source cannot be deposited in a public repository or transferred directly as unrestricted files. Researchers seeking to examine the computational materials should contact the corresponding author, Yankai Zhu (zhuyankai@ncepu.edu.cn), with their affiliation, research purpose, and the analyses they wish to verify. Subject to institutional approval and agreement on the applicable confidentiality conditions, the authors can arrange a secure online meeting in which the relevant implementation is inspected by screen sharing, selected experiments are executed, partition construction and mathematical definitions are checked, and the generation of tables and figures is demonstrated. Direct downloading, recording, redistribution, or permanent copying of protected materials is not permitted without written authorization from the rights-holding institutions. Specific non-sensitive outputs, methodological clarifications, or approved excerpts may be provided after the meeting when compatible with these restrictions. No proprietary plant-historian record or measured AGC trajectory is used in the reported numerical benchmark.

Ethics Approval: Not applicable. This study uses only numerical simulation and synthetic benchmark data; it does not involve human participants, animals, personal data, clinical data, or biological materials.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Mujie Zhang, Yajun Wu, and Dongsheng Li are employed by the Central China Branch of State Grid Corporation of China, and the study received financial support from the Science and Technology Project of State Grid Corporation of China (No. 521400250007), as disclosed in the Funding Statement. Beyond the scholarly contributions of the listed authors, State Grid Corporation

of China had no role in the study conception or design; numerical benchmark construction; generation, analysis, or interpretation of the data; preparation of the manuscript; decision to submit the manuscript; or selection of the journal. No measured plant-historian record, customer data, or commercial product was evaluated in this study.

References

1. Kundur P. Power system stability and control. New York, NY, USA: McGraw Hill; 1994. [CrossRef]
2. Bevrani H. Robust power system frequency control. Boston, MA, USA: Springer; 2009. [CrossRef]
3. Li Z, Hilber P, Li Z, Laneryd T, Ivanell S. Temporally coordinated operation of green multi-energy airport microgrids with climatic correlations and flexible loads via decomposed stochastic programming. IEEE Trans Sustain Energy. 2026;17(2):1909–22. [CrossRef]
4. Chen J, Wu P, Chen W, Guerrero JM, Niu Z, Li Z. Two-layer coordinated operation of multi-energy system considering carbon-oriented collaborative pricing mechanism via two-stage stochastic programming approach. Appl Energy. 2026;406:127298. [CrossRef]
5. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. Cambridge, MA, USA: MIT Press; 2018. [CrossRef]
6. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33. [CrossRef]
7. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-Critic: off-policy maximum entropy deep Reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 1861–70.
8. Glavic M, Fonteneau R, Ernst D. Reinforcement learning for electric power system decision and control: past considerations and perspectives. IFAC-PapersOnLine. 2017;50(1):6918–27. [CrossRef]
9. Zhang Z, Zhang D, Qiu RC. Deep reinforcement learning for power system applications: an overview. CSEE J Power Energy Syst. 2019;6(1):213–25. [CrossRef]
10. Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 1179–91.
11. Levine S, Kumar A, Tucker G, Fu J. Offline Reinforcement learning: tutorial, review, and perspectives on open problems. arXiv:2005.01643. 2020. [CrossRef]
12. Ames AD, Xu X, Grizzle JW, Tabuada P. Control barrier function based quadratic programs for safety critical systems. IEEE Trans Autom Control. 2017;62(8):3861–76. [CrossRef]
13. Rockafellar RT, Uryasev S. Optimization of conditional value-at-risk. J Risk. 2000;2(3):21–41. [CrossRef]
14. Artzner P, Delbaen F, Eber JM, Heath D. Coherent measures of risk. Math Financ. 1999;9(3):203–28. [CrossRef]
15. Bellemare MG, Dabney W, Munos R. A distributional perspective on reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 449–58. [CrossRef]
16. Dabney W, Rowland M, Bellemare M, Munos R. Distributional Reinforcement learning with Quantile regression. Proc AAAI Conf Artif Intell. 2018;32(1):2892–901. [CrossRef]
17. Wood AJ, Wollenberg BF, Sheble GB. Power generation, operation, and control. 3rd ed. Hoboken, NJ, USA: Wiley; 2013.
18. Machowski J, Bialek JW, Bumby JR. Power system dynamics: stability and control. Hoboken, NJ, USA: Wiley; 2011.
19. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv:1509.02971. 2015. [CrossRef]
20. Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy. 2019;235:1072–89. [CrossRef]
21. Achiam J, Held D, Tamar A, Abbeel P. Constrained policy optimization. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 22–31.
22. García J, Fernández F. A comprehensive survey on safe reinforcement learning. J Mach Learn Res. 2015;16(1):1437–80.

23. Chow Y, Ghavamzadeh M, Janson L, Pavone M. Risk-constrained reinforcement learning with percentile risk criteria. J Mach Learn Res. 2018;18(167):1–51.
24. Yu P, Zhang H, Song Y, Wang Z, Dong H, Ji L. Safe reinforcement learning for power system control: A review. Renew Sustain Energy Rev. 2025;223:116022. [CrossRef]
25. Brunke L, Greeff M, Hall AW, Yuan Z, Zhou S, Panerati J, et al. Safe learning in Robotics: from learning-based control to safe Reinforcement learning. Annu Rev Control Robot Auton Syst. 2022;5(1):411–44. [CrossRef]
26. Gu S, Yang L, Du Y, Chen G, Walter F, Wang J, et al. A review of safe Reinforcement learning: methods, theories, and applications. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):11216–35. [CrossRef]
27. Li H, He H. Learning to operate distribution networks with safe deep Reinforcement learning. IEEE Trans Smart Grid. 2022;13(3):1860–72. [CrossRef]
28. Gao Y, Yu N. Model-augmented safe reinforcement learning for volt-VAR control in power distribution networks. Appl Energy. 2022;313:118762. [CrossRef]
29. Qiu D, Dong Z, Zhang X, Wang Y, Strbac G. Safe reinforcement learning for real-time automatic control in a smart energy-hub. Appl Energy. 2022;309:118403. [CrossRef]
30. Xia Y, Xu Y, Wang Y, Mondal S, Dasgupta S, Gupta AK, et al. A safe policy learning-based method for Decentralized and economic frequency control in isolated networked-Microgrid systems. IEEE Trans Sustain Energy. 2022;13(4):1982–93. [CrossRef]
31. Cui W, Jiang Y, Zhang B. Reinforcement learning for optimal primary frequency control: A Lyapunov approach. IEEE Trans Power Syst. 2023;38(2):1676–88. [CrossRef]
32. Zhou Y, Zhou L, Shi D, Zhao X. Coordinated frequency control through safe Reinforcement learning. In: Proceedings of the 2022 IEEE Power & Energy Society General Meeting (PESGM); 2022 Jul 17–21; Denver, CO, USA. p. 1–5. [CrossRef]
33. Bai Q, Bedi AS, Agarwal M, Koppel A, Aggarwal V. Achieving zero constraint violation for Constrained Reinforcement learning via Primal-dual approach. Proc AAAI Conf Artif Intell. 2022;36(4):3682–9. [CrossRef]
34. Wang Y, Qiu D, Sun M, Strbac G, Gao Z. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. Appl Energy. 2023;335:120759. [CrossRef]
35. Rokhforoz P, Montazeri M, Fink O. Safe multi-agent deep reinforcement learning for joint bidding and maintenance scheduling of generation units. Reliab Eng Syst Saf. 2023;232:109081. [CrossRef]
36. Guo G, Zhang M, Gong Y, Xu Q. Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay. Appl Energy. 2023;349:121648. [CrossRef]
37. Yu P, Zhang H, Song Y. District cooling system control for providing regulation services based on safe reinforcement learning with barrier functions. Appl Energy. 2023;347:121396. [CrossRef]
38. Ceusters G, Putratama MA, Franke R, Nowé A, Messagie M. An adaptive safety layer with hard constraints for safe reinforcement learning in multi-energy management systems. Sustain Energy Grids Netw. 2023;36:101202. [CrossRef]
39. Wang Y, Zhan SS, Jiao R, Wang Z, Jin W, Yang Z, et al. Enforcing hard constraints with soft barriers: safe reinforcement learning in unknown stochastic environments. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 36593–604.
40. Yuan Z, Zhao C, Cortés J. Reinforcement learning for distributed transient frequency control with stability and safety guarantees. Syst Control Lett. 2024;185:105753. [CrossRef]
41. Xu H, Zhan X, Zhu X. Constraints Penalized Q-learning for safe offline Reinforcement learning. Proc AAAI Conf Artif Intell. 2022;36(8):8753–60. [CrossRef]
42. Rawlings JB, Mayne DQ, Diehl MM. Model predictive control: theory, computation, and design. Santa Barbara, CA, USA: Nob Hill Publishing; 2017.
43. Feng J, Ren Z, Li C, Li W. A Benders-combined safe Reinforcement learning framework for risk-Averse dispatch considering frequency security constraints. IEEE Trans Circuits Syst II Express Briefs. 2025;72(8):1063–7. [CrossRef]
44. Li J, Gu K, Zhao K, Li Z, Dong Z. A robust and fast operational risk assessment method for composite power systems with high wind power penetration. IEEE Trans Power Syst. 2026;41(4):2838–49. [CrossRef]

45. Feng C, Huang Z, Lin J, Wang L, Zhang Y, Wen F. Aggregation model and market mechanism for virtual power plant participation in Inertia and primary frequency response. IEEE Trans Power Syst. 2026;41(3):2101–17. [CrossRef]
46. Wang Y, Xie D, Shi C, Wang X. Dynamic construction strategy for virtual power plants considering distributed energy resource interaction uncertainty and response recovery process. IEEE Trans Ind Appl. 2026. [CrossRef]
47. Chen X, Liu Y, Zhong Z, Fan N, Zhao Z, Wu L. A Carryover storage valuation framework for medium-term Cascaded Hydropower planning: A portland general electric system study. IEEE Trans Sustain Energy. 2025;16(3):1903–18. [CrossRef]
48. Wang Z, Gu Z, Guerrero JM, Shen Y, Guo Z, Deng Z, et al. Two-period two-layer electrical model-free voltage calculation for active distribution network via improved broad learning system. IEEE Trans Power Syst. 2026;41(4):2704–16. [CrossRef]
49. Wen Y, Yu W, Fu H, Thotakura NL, Jiang S, Dulal S, et al. A systematic Oscillation source location estimation method for ultra-low frequency and sub-Synchronous Oscillations in power systems. In: Proceedings of the 2025 IEEE Power & Energy Society General Meeting (PESGM); 2025 Jul 27–31; Austin, TX, USA. p. 1–5. [CrossRef]
50. Åström KJ, Bell R. Dynamic models for boiler-turbine-alternator units: data logs and parameter estimation for a 160 MW unit. Lund, Sweden: Department of Automatic Control, Lund Institute of Technology; 1987. Research Report TFRT-3192. ISSN: 0280-5316.
51. Wu X, Shen J, Li Y, Lee KY. Data-driven modeling and predictive control for boiler–turbine unit using fuzzy clustering and subspace methods. ISA Trans. 2014;53(3):699–708. [CrossRef]
52. Boyd S, Vandenberghe L. Convex optimization. Cambridge, UK: Cambridge University Press; 2004. [CrossRef]
53. Chen H, Xu M, Shi C, Wang Z, Xie D. Data–model hybrid hierarchical scheduling for grid-connected PEM green hydrogen production in multi-energy systems with LLM-aided coordination. J Renew Sustain Energy. 2026;18(2):024701. [CrossRef]
54. Bertsimas D, Sim M. The price of robustness. Oper Res. 2004;52(1):35–53. [CrossRef]
55. Nilim A, Ghaoui EL. Robust control of Markov decision processes with Uncertain transition Matrices. Oper Res. 2005;53(5):780–98. [CrossRef]
56. Iyengar GN. Robust dynamic programming. Math Oper Res. 2005;30(2):257–80. [CrossRef]
57. Tessler C, Mankowitz DJ, Mannor S. Reward constrained policy optimization. In: Proceedings of the International Conference on Learning Representations; 2019 May 6–9; New Orleans, LA, USA.