Offline Model-Based Reinforcement Learning for Energy-Efficient GPU Data-Center Cooling

Stanford CS224R Final Report

Naomie Chien
Department of Computer Science
Stanford University
naochien@stanford.edu

June 8, 2026

Extended Abstract

Motivation. Data center cooling consumes 30–40% of facility electricity yet remains governed by reactive, hand-tuned PID controllers that over-cool during idle periods and lag behind thermal transients. Learned controllers could close this gap, but live exploration in a production facility is thermally unsafe. Offline model-based reinforcement learning (MBRL) offers a principled alternative: train a dynamics simulator on historical telemetry, augment the dataset with synthetic rollouts, and optimize a policy entirely offline.

Approach. We build a complete offline MBRL pipeline on real telemetry from the Marconi100 (M100) GPU supercomputer at CINECA, a 980-node cluster logged at 10-second cadence. *Stage 1* trains a probabilistic ensemble of K=5 residual MLPs predicting the next-state residual $\Delta s = s_{t+1} - s_t$ with per-dimension Gaussian uncertainty from a 28-dimensional thermal/power state and a 5–7-dimensional cooling action. A physics-constrained chiller head enforces $p_{\text{chiller}} = \text{COP}(s, a) \cdot p_{\text{IT}}$, eliminating otherwise-dominant PUE prediction error. *Stage 2* trains three controllers entirely within the learned simulator: (i) Conservative Q-Learning (CQL), (ii) random-shooting Model Predictive Control (MPC), and (iii) a hybrid MPC+CQL using the CQL critic as a bootstrapped terminal value. Dataset coverage is expanded beyond the narrow PID action distribution via MOPO-style synthetic rollouts penalized by ensemble uncertainty.

Results. The ensemble achieves a one-step normalized RMSE of 0.033 and holds 30-step rollout RMSE to 0.29 (a 3× reduction over a single model) while the COP head drives PUE error to near zero. Over 64 held-out 5-minute episodes, MPC achieves a mean return of -7.26 (vs. -7.94 for logged PID), winning every episode and reducing simulated cooling energy by 8.15%. Hybrid MPC+CQL wins 96.9% of episodes with a 3.87% energy reduction. CQL yields a marginal but statistically significant improvement (+0.02 return), constrained by the narrow PID behavioral support. All controllers achieve zero ASHRAE A2 thermal violations.

Key findings. (1) A physics-constrained ensemble simulator is faithful enough for offline RL: an oracle over ground-truth dynamics scores -7.98, within 0.04 of the model's logged-policy estimate. (2) MPC's forward planning escapes the behavioral support of the offline dataset in a way CQL's conservatism cannot, yielding an **8.15% simulated cooling energy reduction** at zero thermal violations. (3) The structural gap between MPC and CQL traces to dataset coverage, not algorithm quality; richer action diversity is the critical path to stronger offline RL performance. (4) The dominant remaining modeling challenge is chiller power at free-cooling regime boundaries; a dedicated regime-switching head is the most promising near-term fix.

1 Introduction

Artificial intelligence infrastructure has become one of the fastest-growing sources of electricity demand globally. Data centers already account for roughly 1–2% of global electricity consumption, and projections from the International Energy Agency suggest demand will more than double by 2030 as GPU-based training and inference workloads continue to scale [International Energy Agency, 2025]. Cooling systems represent the largest controllable energy load within data centers, typically consuming 30–40% of total facility power. As rack power densities climb toward and beyond 100 kW per rack for modern AI accelerator pods, cooling has become a key operational and sustainability constraint.

The core difficulty is that AI workloads are thermally non-stationary in a way that conventional control architectures were not designed to handle. GPU clusters alternate between sustained high-utilization training runs and intermittent, bursty inference jobs, producing rapid thermal transients with rise times on the order of minutes. Rule-based Building Management System (BMS) controllers respond reactively: they systematically over-cool during low-load periods and lag behind load spikes, risking thermal violations that can trigger compute throttling. The result is a persistent tension between energy efficiency and thermal safety that manual setpoint tuning cannot resolve at scale.

Online reinforcement learning (RL) could in principle learn an anticipatory control policy, but live exploration in a production data center is unsafe: a bad action can violate thermal limits and throttle a 980-node cluster. *Offline* model-based RL (MBRL) offers a principled alternative: learn a simulator from historical telemetry and optimize a policy entirely within it, with no live risk to the facility.

This paper presents a complete offline MBRL pipeline for GPU data-center cooling. Our contributions are:

1. A physics-constrained probabilistic ensemble dynamics model trained on real Marconi100 supercomputer telemetry, with a thermodynamic chiller head that enforces $p_{\text{chiller}} = \text{COP}(s, a) \cdot p_{\text{IT}}$ and eliminates dominant PUE prediction error.
2. A MOPO-style synthetic data augmentation strategy that expands the offline dataset beyond the narrow PID action distribution while penalizing high-uncertainty transitions.
3. A comparative evaluation of CQL, random-shooting MPC, and a hybrid MPC+CQL controller, all trained and evaluated entirely offline, demonstrating up to 8.15% simulated cooling energy reduction at zero thermal violations.

2 Related Work

Classical and model-predictive control. Industry-standard cooling relies on PID- and rule-based BMS controllers: interpretable and safe, but reactive and manually tuned, leading operators to adopt conservative, energy-wasteful setpoints. Lazic et al. [2018] apply MPC to a Google data center, outperforming PID by learning a sparse linear dynamics model from a few hours of randomized exploration. MPC requires an accurate hand-built plant model, however, and scales poorly with the number of controlled variables, which is the system-identification burden that a learned model removes. Mirhoseininejad et al. [2021] extend MPC to multi-setpoint control with a data-driven thermal model, demonstrating the importance of frequent model updates as hardware configurations change.

Neural dynamics models and model-based RL. Model-based RL learns a neural dynamics model and plans through it. Nagabandi et al. [2018] establish that even simple learned models can support sample-efficient control. Chua et al. [2018] (PETS) show that a probabilistic ensemble capturing both aleatoric (output-variance) and epistemic (inter-member disagreement) uncertainty greatly improves long-horizon planning performance. We adopt the PETS ensemble design but add a domain-specific physics constraint and apply it to real facility telemetry. Open-loop rollouts accumulate error as predictions are fed back as inputs; Janner et al. [2019] (MBPO) mitigate this with short rollouts branched from real states, the principle behind our 5-step planning and rollout horizon.

Offline RL and conservatism. Model-free offline RL collapses under distributional shift because Q-values are overestimated for actions absent from the logged data. Kumar et al. [2020] (CQL) penalize out-of-distribution actions with an explicit regularization term on the Q-function. Yu et al. [2020] (MOPO) penalize the reward by ensemble uncertainty during synthetic rollouts, allowing policy training on model-generated data while staying near the behavioral distribution. Yu et al. [2021] (COMBO) extends this to conservative offline model-based optimization without requiring a separately trained behavior policy. Our pipeline integrates CQL with MOPO-style rollout augmentation.

RL for data center cooling. Luo et al. [2022] report 9–13% cooling energy savings in live Google deployments using online RL. Zhan et al. [2025] demonstrate that offline RL can achieve 14–21% cooling energy reductions with zero thermal violations in over 2000 hours of real operation. Naug et al. [2024] introduce SustainDC, a benchmarking environment that evaluates agents simultaneously on energy, carbon intensity, and water use. Our work contributes a reproducible offline pipeline on real supercomputer telemetry, enabling systematic comparison of model-based and policy-gradient approaches without live risk.

3 Background: Problem Formulation

We model the GPU cooling pod as a discrete-time MDP with 10-second steps. The state $s_t \in \mathbb{R}^{28}$ covers six sensor streams: the liquid cooling circuit (supply/return temperatures, flow, valve positions, pump speed), the air cooling circuit (supply/return temperatures, fan speed, compressor utilization, free-cooling status and fluid temperature), facility power metering (IT load p_{IT}, CDZ power, chiller power p_{chiller}, pump power, PUE), cluster workload (GPU/CPU/memory utilization), node-level thermal readings (rack inlet and GPU core), outdoor weather, and cyclic time-of-day features. The action $a_t \in \mathbb{R}^5$ controls CRAH fan speed, two liquid valve positions, liquid pump speed, and free-cooling valve position (extended to $a_t \in \mathbb{R}^7$ for MPC, adding air and liquid supply temperature setpoints).

The goal of an offline RL agent is to find a policy $\pi^*(a|s)$ maximizing expected cumulative reward using only a fixed dataset $\mathcal{D}$ of transitions (s, a, r, s') collected under the facility's existing PID controller, with no further environment interaction.

Six dimensions (IT load, GPU/CPU/memory utilization, outdoor temperature and humidity) are exogenous: the cooling policy cannot influence them. These channels are propagated forward unchanged during synthetic rollouts, which is physically correct and prevents spurious feedback.

4 Dataset

Source. We use real telemetry from Marconi100 (M100), a 980-node GPU supercomputer at CINECA monitored at 10-second cadence across seven ExaMon namespaces [Borghesi et al., 2023]. We use three operating windows (February, May, and September 2022) that span seasonal variation in outdoor conditions and cooling regime, spanning summer free-cooling and winter mechanical cooling. Raw streams from the Schneider liquid-cooling loop (RDHx), Vertiv air-cooling units (CDZ), facility power meters, and Slurm workload logs are joined onto a common 10-second grid.

Preprocessing. Each channel is standardized to zero mean and unit variance using training-split statistics. Targets are the per-step residuals $\Delta s = s_{t+1} - s_t$ rather than absolute next states; because most variables change slowly at 10-second cadence, residuals are zero-centered and small-magnitude, giving a better-conditioned target. Splits are made at episode boundaries to prevent temporal leakage. The 30-step ($\times 10\,\text{s} = 5\,\text{min}$) evaluation horizon matches the commercial cooling actuation cycle.

Regime structure. A key property of the data is bimodality. The pod operates in both *mechanical* and *free* cooling: when free cooling engages, compressor utilization collapses toward zero while chiller power drops to a distinct lower cluster. The dynamics model must implicitly learn this regime switch from continuous state input, which motivates the structured chiller head described below.

Dataset limitation. Because the historical logs were collected entirely under PID control, the dataset covers a narrow slice of the possible action space. Fan speed, pump speed, and liquid valve positions exhibit almost no variation in the logged data; free-cooling valve coverage is especially sparse. This behavioral concentration is the fundamental challenge for offline RL and motivates synthetic augmentation.

5 Methods

5.1 Stage 1: Physics-Constrained Ensemble Dynamics Model

Residual MLP backbone. Each ensemble member is a residual MLP (≈546K parameters): a linear projection to a 256-dimensional hidden space, four pre-activation residual blocks (LayerNorm → SiLU → Linear, twice per block, dropout 0.15), and a near-zero-initialized output projection so the model starts close to the identity map $\hat{s}_{t+1} \approx s_t$. The near-zero initialization is particularly important: it ensures the model does not need to first "unlearn" a random output before fitting the small-magnitude residuals.

Probabilistic outputs. Each member outputs a Gaussian over the residual: a mean $\mu_\theta(s,a)$ and a softplus-bounded log-variance $\log \sigma^2_\theta(s,a)$ per dimension, trained by Gaussian NLL:

$$\mathcal{L}_{\text{NLL}}(\theta) = \frac{1}{2} \sum_d \left[\frac{(\Delta s_d - \mu_{\theta,d})^2}{\sigma^2_{\theta,d}} + \log \sigma^2_{\theta,d} \right]. \tag{1}$$

This allows heteroscedastic noise modeling: the model can express higher uncertainty on inherently noisy channels (chiller power, humidity) and lower uncertainty on slowly varying thermal channels.

K=5 ensemble and uncertainty decomposition. We train K=5 members with independent random initialization and bootstrap-resampled minibatches. At inference the ensemble mean is the prediction. Aleatoric uncertainty is the mean predicted variance $\frac{1}{K}\sum_k \sigma^2_{\theta_k}(s,a)$; epistemic uncertainty is the inter-member variance of the means, $\frac{1}{K}\sum_k (\mu_k - \bar{\mu})^2$. Downstream MOPO and MPC planners use epistemic uncertainty to avoid regions where the model is unreliable.

Physics-constrained chiller head. Chiller power is the hardest channel to predict and dominates PUE, which is the efficiency KPI in the control reward. Rather than regress p_{chiller} freely, we enforce the thermodynamic identity:

$$p_{\text{chiller}}(s,a) = m_\phi(s,a) \cdot p_{\text{IT}}, \tag{2}$$

where m_ϕ is a small MLP (≈2.8K parameters) whose output is sigmoid-squashed into $[0.02, 0.60]$ so it cannot exceed physically plausible cooling-to-IT power ratios. PUE is then recomputed analytically:

$$P_{\text{cool}} = P_{\text{CDZ}} + P_{\text{chiller}} + P_{\text{pump}}, \tag{3}$$

$$\text{PUE} = \frac{P_{\text{IT}} + P_{\text{cool}}}{P_{\text{IT}}}. \tag{4}$$

The head is initialized near the empirical M100 cooling ratio (≈0.22). Because the M100 telemetry does not separately record commanded versus measured free-cooling valve position, the valve channel is masked out of m_ϕ's input and a fixed linear suppression term is applied instead.

5.2 Stage 2: Offline Control

Reward function. The reward combines energy efficiency with thermal safety and physical operating constraints:

$$r_{\text{primary}} = -\alpha\,(\text{PUE} - 1) - \beta \sum_r \max\big(0,\, T_{\text{inlet},r} - 27\big) - \gamma\, \|a_t - a_{t-1}\|^2, \tag{5}$$

where 27 °C is the ASHRAE A2 server inlet limit. Five soft penalties enforce physical constraints (cooling capacity, coil short-cycling, hot-aisle return cap, minimum airflow, humidity band). Total reward is $r = r_{\text{primary}} + r_{\text{soft}}$.

MOPO synthetic augmentation. To expand coverage beyond the narrow PID action distribution, we follow the MOPO framework [Yu et al., 2020]: every 10K training steps, the current policy is unrolled through the dynamics model for H=5 steps from real dataset states, producing synthetic transitions penalized by ensemble disagreement:

$$\tilde{r}(s,a) = r(s,a) - \lambda\, \sigma_{\text{ens}}(s,a), \tag{6}$$

where $\sigma_{\text{ens}}(s,a) = \sqrt{\text{Var}_k[\mu_k(s,a)]}$ is the ensemble standard deviation and λ is a tuned pessimism coefficient. Synthetic transitions are mixed with real data at approximately 78% synthetic / 22% real after the buffer warms up.

CQL policy. The CQL policy operates on a 5-dimensional actuator space. CQL augments the standard Bellman objective with a regularization term that suppresses Q-values for out-of-distribution actions:

$$\mathcal{L}_{\text{CQL}}(Q) = \alpha(\mathbb{E}_{s,a\sim\mu}[Q(s,a)] - \mathbb{E}_{s,a\sim\beta}[Q(s,a)]) + \frac{1}{2}\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\big[Q(s,a) - \mathcal{B}^\pi Q(s,a)\big]^2, \tag{7}$$

where $\mathcal{B}^\pi$ is the Bellman operator and α is annealed from 1.0 to 0.1 over 100K gradient steps. The CQL conservatism prevents the policy from exploiting hallucinated Q-values in regions the PID controller never explored.

MPC planner. At each step, 512 candidate action sequences are sampled uniformly from the actuator space, rolled out for H=5 steps through the learned ensemble, and the sequence with the highest cumulative discounted reward is executed. MPC evaluates actions by simulating their consequences directly rather than querying a pre-trained Q-function, so it is unconstrained by the behavioral distribution of the offline dataset and can freely exploit actions the PID controller never took. The MPC variant operates on a 7-dimensional action space, additionally commanding air and liquid supply temperature setpoints.

Hybrid MPC+CQL. Rather than summing rewards over the full planning horizon, the hybrid uses the CQL critic to provide a bootstrapped terminal value at planning depth H=1. This combines MPC's willingness to explore novel actions with CQL's learned estimate of long-run return, reducing per-episode variance while maintaining most of MPC's energy-efficiency gains.

6 Experiments and Results

6.1 Dynamics Model Performance

Training. We train two separate dynamics ensembles, one for the CQL 5-dimensional action space and one for the MPC 7-dimensional action space, with AdamW, cosine learning rate decay, gradient clipping, and early stopping (patience 8 epochs). The CQL dynamics model converges over 18 epochs (best validation NLL -2.19 at epoch 10); the MPC model converges faster, triggering early stopping at epoch 12 (best val NLL -1.90 at epoch 4). The wider train/val NLL gap in the MPC model reflects its larger, sparser action space: the additional setpoint dimensions are rarely varied in the logged PID trajectories.

The CS229-reported model (the primary ensemble used for downstream evaluation) trains for 38 epochs with best validation NLL -3.25 at epoch 30, showing modest train/val gap (-3.65 train vs. -3.25 val), indicating generalization without significant overfitting.

One-step accuracy and ensemble effect. Table 1 compares a single probabilistic residual MLP against the K=5 ensemble at increasing rollout horizons (normalized RMSE). The ensemble roughly halves one-step error and provides a 3× reduction at the 30-step (5-minute) horizon, keeping rollout RMSE well below 0.3 throughout the evaluation window.

Table 1: Rollout RMSE (normalized) vs. horizon, single model vs. ensemble.

Horizon	Single model	Ensemble (K=5)
1-step	0.065	0.033
5-step	0.446	0.108
10-step	0.446	0.192
20-step	0.681	0.214
30-step	0.876	0.291

Per-dimension error and the COP-head ablation. Without the chiller head, PUE is freely regressed and achieves RMSE ≈ 7.4 on a quantity that physically sits near 1.5, effectively a broken predictor. The COP head ties PUE to predicted chiller power by construction, driving PUE error to ≈ 0 and concentrating residual power error in chiller power. All thermal channels (liquid/air supply and return, node inlet, GPU core) are predicted to $< 0.1\,^{\circ}$C. The constraint buys *consistency* rather than smaller aggregate error: it makes the reward's efficiency signal trustworthy and confines the remaining difficulty to one meaningful channel.

Oracle validation. An oracle that rolls logged actions through ground-truth dynamics returns -7.98, within 0.04 of the model's logged-policy estimate (-7.94). This near-zero gap confirms that the learned simulator is faithful to the real system and that downstream performance gains are not artifacts of model bias.

6.2 Synthetic Data Quality

The MOPO ring buffer exhibits well-calibrated behavior: higher-uncertainty synthetic transitions receive proportionally stronger pessimism penalties, limiting their influence on policy updates. Real and synthetic reward distributions overlap substantially and share a bimodal structure peaking near -0.25, confirming that the dynamics model generates realistic transitions. The bimodal structure (corresponding to mechanical vs. free-cooling regimes) is faithfully reproduced, validating the synthetic augmentation strategy.

6.3 Offline Control Results

Aggregate performance. Table 2 summarizes mean returns and win rates over 64 held-out 5-minute episodes. All controllers achieve zero ASHRAE A2 thermal violations.

Table 2: Performance over n=64 held-out validation episodes (30-step horizon). Δ is mean improvement over logged PID. p-values from paired two-tailed t-test (df = 63).

Controller	Return (mean $\pm$ std)	Δ	Beats Logged (%)	p
Logged (facility PID)	-7.94 ± 1.27	–	–	–
CQL	-7.92 ± 1.25	+0.02	70.3	< 0.001
Hybrid MPC+CQL	-7.51 ± 1.17	+0.43	96.9	< 0.001
MPC	-7.26 ± 1.23	+0.68	100.0	< 0.001

Energy efficiency. Table 3 compares PUE and cooling energy. MPC achieves the strongest improvement: 8.15% cooling energy reduction and 1.68% PUE improvement. Hybrid MPC+CQL achieves 3.87% energy reduction. CQL produces only marginal improvement (-0.30%, $p = 0.12$).

Action-level analysis. CQL deviates modestly and selectively from the PID baseline. The largest deviations occur on the two liquid valve positions ($\Delta_{\text{abs}} = 5.4\%$ and 8.1%) and the free-cooling valve (6.0%), while pump speed deviates least (0.23%). This hierarchy reflects CQL's conservatism: dimensions densely covered

Table 3: Controller comparison on cooling energy. Negative values indicate reduced consumption.

Controller	PUE	ΔPUE (%)	ΔE_{cool} (%)
Logged (facility PID)	1.259	0.00	0.00
CQL	1.259	-0.06	-0.30
Hybrid MPC+CQL	1.249	-0.80	-3.87
MPC	1.238	-1.68	-8.15

in logged data receive confident Q-estimates and stay close to baseline, while episodically commanded valve dimensions are pushed to explore, but in regions where the dynamics model is less reliable.

MPC produces substantially larger and directionally distinct deviations. Fan speed is compressed to a narrow band around 53–54% regardless of what PID commanded (60–75%), and pump speed is reduced from near-constant 100% to 51–67%. Both liquid valves are throttled toward their lower limits, and MPC increases free-cooling valve opening relative to PID, actively substituting passive cooling for mechanical load. The hybrid MPC+CQL follows the MPC directional pattern with greater spread, with the CQL critic adding useful regularization that reduces per-episode variance ($\sigma = 1.17$ vs. 1.23 for pure MPC).

The structural asymmetry between MPC and CQL. The large performance gap between MPC and CQL is a fundamental consequence of the offline learning setting, not a failure of the CQL implementation. MPC evaluates actions by simulating their consequences forward through the dynamics model at each step; it can freely propose and score action sequences that the PID controller never executed. CQL is constrained by the Q-function's behavioral support: it can only extrapolate to actions that are plausibly nearby the logged distribution, and the MOPO augmentation partially but not fully resolves this. The oracle (best historically observed actions within the PID telemetry) itself returns -7.98, slightly worse than logged, confirming that it is a ceiling on retrospective action selection within the behavioral distribution, not a ceiling on what is physically achievable. MPC surpasses it precisely by escaping this distribution.

Free-cooling valve utilization illustrates this asymmetry most clearly. CQL assigns pessimistically low Q-values to high free-cooling openings because the ensemble was never trained on dense transitions in that regime, a generalization failure rather than a reward specification failure. MPC and the hybrid sidestep this by evaluating novel valve positions through forward simulation, though both remain constrained by ensemble uncertainty on far-out-of-distribution states.

7 Discussion and Limitations

Dataset coverage as the binding constraint. The key limitation is not the optimization algorithm but the narrow behavioral distribution of the historical telemetry. The M100 dataset was collected entirely under PID control, so the offline policy observes only a small slice of potential facility state-action space. The conservative biases of PID (never aggressively engaging free cooling, always running pumps near 100%) are inherited by CQL. MOPO augmentation partially mitigates this by generating synthetic transitions under more diverse actions, but the dynamics model's epistemic uncertainty remains high in these out-of-distribution regions, limiting how much synthetic data the pessimism-penalized policy will exploit.

Chiller power and regime switching. The remaining failure mode of the dynamics model is chiller power at the boundaries of the mechanical-to-free-cooling regime switch. The bimodal operating structure means the model must learn a near-discontinuous transition from a continuous state input, and the two regimes are unevenly represented in the data. Errors at these boundaries propagate to PUE estimates during long rollouts and reduce the planner's confidence in aggressive free-cooling actions. A dedicated regime classifier or a mixture-of-experts chiller head could address this in future work.

Evaluation is simulated, not deployed. All controller evaluations are performed within the learned dynamics model, not on the live facility. Results should be interpreted as simulated savings potential; real-

world deployment would require a shadow-mode validation period and a conservative rollout strategy to confirm that the learned policy is thermally safe before closing the control loop.

Operational integration. Data centers are typically managed through fragmented software stacks (independent BMS, EPMS, and DCIM platforms) with no shared optimization objective. A deployed RL controller would need to interface with all three, requiring careful API design, latency characterization, and fallback procedures. Safety interlocks that override the learned policy when thermal margins are tight are a prerequisite for any production deployment.

8 Future Work

Several directions could extend this work. First, richer action coverage (from a randomized exploration period, an EnergyPlus simulation, or a multi-facility dataset) would substantially broaden the behavioral support available to CQL and could close much of the gap with MPC. Second, a dedicated regime-switching model (e.g., a learned gating network that routes to separate dynamics heads for mechanical and free-cooling operation) could reduce chiller power error and improve rollout fidelity at regime boundaries. Third, online fine-tuning of the dynamics model during a shadow-mode deployment (where the policy observes real outcomes without actuating) could track distribution shift as hardware changes or seasonal operating regimes evolve. Finally, multi-objective evaluation using SustainDC-style metrics (energy, carbon, water) would give a more complete picture of sustainability impact than PUE alone.

9 Conclusion

We present a complete offline MBRL pipeline for GPU data-center cooling, trained and evaluated on real Marconi100 supercomputer telemetry. A physics-constrained K=5 ensemble dynamics model achieves 0.033 one-step RMSE, compounds rollout error 3× more slowly than a single model, and drives near-zero PUE prediction error via a thermodynamic chiller head. An oracle evaluation confirms the simulator is faithful to the real system. In downstream control evaluation, an uncertainty-pessimistic MPC planner reduces simulated cooling energy by 8.15% and wins every episode against the logged PID baseline; a hybrid MPC+CQL achieves 96.9% win rate with 3.87% energy savings; and CQL alone yields marginal but statistically significant improvement. The structural asymmetry between MPC and CQL traces to dataset coverage: MPC escapes the behavioral support of the PID-only logs through forward simulation, while CQL's conservatism confines both its risks and its gains to the historical operating regime. This gap identifies richer action coverage (from diverse exploration data or high-fidelity simulation) as the most important direction for future offline RL deployment in AI infrastructure.

Contributions. Naomie Chien designed and implemented the dynamics model architecture (residual MLP backbone, probabilistic NLL heads, K=5 bootstrap ensemble, physics-constrained COP head), the training and evaluation harness, the MOPO rollout buffer, and the MPC planner. Naomie also led integration of the dynamics model with the CQL pipeline. Nadja Yang developed the reward function, the hybrid MPC+CQL controller, and the energy and PUE evaluation framework. Pradyumna Singh contributed to CQL training, the action-level analysis, and the dataset preprocessing pipeline.

References

Borghesi, A., Bartolini, A., Milano, M., and Benini, L. (2020). *M100 ExaMon: a holistic monitoring dataset of the CINECA Marconi100 supercomputer.*

Borghesi, A., Di Santi, C., Molan, M., et al. (2023). *M100 ExaData: a data collection campaign on the CINECA's Marconi100 Tier-0 supercomputer.* Scientific Data 10, 288.

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). *Deep reinforcement learning in a handful of trials using probabilistic dynamics models (PETS).* NeurIPS 31.

International Energy Agency (2025). *Energy and AI.* Technical Report, IEA, Paris.

Janner, M., Fu, J., Zhang, M., and Levine, S. (2019). *When to trust your model: model-based policy optimization (MBPO).* NeurIPS 32.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). *Conservative Q-Learning for offline reinforcement learning.* NeurIPS 33, 1179–1191.

Lazic, N., Boutilier, C., Lu, T., et al. (2018). *Data center cooling using model-predictive control.* NeurIPS 31, 3818–3827.

Luo, J., Paduraru, C., Voicu, O., et al. (2022). *Controlling commercial cooling systems using reinforcement learning.* arXiv:2211.07357.

Mirhoseininejad, S., Badawy, G., and Down, D. (2021). *A data-driven, multi-setpoint model predictive thermal control system for data centers.* Journal of Network and Systems Management 29.

Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. (2018). *Neural network dynamics for model-based deep RL with model-free fine-tuning.* ICRA.

Naug, A., Guillen, A., Luna, R., et al. (2024). *SustainDC: Benchmarking for sustainable data center control.* NeurIPS 37, 100630–100669.

Yu, T., Thomas, G., Yu, L., et al. (2020). *MOPO: Model-based offline policy optimization.* NeurIPS 33.

Yu, T., Kumar, A., Rafailov, R., et al. (2021). *COMBO: Conservative offline model-based policy optimization.* NeurIPS 34.

Zhan, X., Zhu, X., Cheng, P., et al. (2025). *Data center cooling system optimization using offline reinforcement learning.* ICLR.