Towards Affordable Energy: A Gymnasium Environment for Electric Utility Demand-Response Program

Jose E. Aguilar Escamilla, Lingdong Zhou, Xiangqi Zhu, Huazheng Wang

Keywords: demand response, electricity markets, reinforcement learning, risk-aware learning, building energy simulation, Gymnasium

Summary

We present DR-GYM, a Gymnasium-compatible reinforcement learning environment for electric utility demand-response optimization. The environment simulates an electric utility economic model that issues hourly (or daily) electricity bill credits to a heterogeneous population of residential buildings, mediating between a volatile wholesale electricity market and price-vulnerable consumers. The reward function is configurable, and permits users to balance learning objectives over utility revenue, consumer costs, and grid reliability. We also allow optional risk-aware measures for penalizing risk in consumer bills, connecting the environment to the growing literature on risk-aware reinforcement learning (Garcia & Fernandez, 2015). The market model uses a regime-switching Markov spike process calibrated to CAISO and ERCOT price statistics. Building demand data is drawn from CityLearn's EnergyPlus-simulated ResStock profiles. Experiments with Proximal Policy Optimization show that a learned policy outperforms all hand-designed baselines on the composite reward and achieves lower consumer bills. DR-Gym is intended to serve as a rigorous testbed for reinforcement learning in the context of demand response and the interaction of electricity wholesale market and electricity distribution system economics .

Contribution(s)

1. We present DR-Gym, an open-source Gymnasium-compatible reinforcement learning environment for electric utility demand-response policy optimization. The environment models a single electric utility issuing hourly or daily bill credits to a heterogeneous population of buildings, with a configurable, multi-objective reward.
   Context: Existing RL environments for building energy (e.g., CityLearn) focus on device-level scheduling decisions for a single building; DR-Gym focuses on the market-level electric utility environment, acting across many buildings under wholesale price uncertainty.
2. The market model incorporates a regime-switching price-spike process with temperature coupling, validated against CAISO and ERCOT day-ahead price statistics. The customer model uses a heterogeneous logistic acceptance function calibrated to empirical demand-response trial data.
   Context: Price parameters are calibrated to CAISO 2019 day-ahead data; customer parameters are calibrated to ranges reported in Faruqui & Sergici (2010).
3. Experiments using data snapshots and Proximal Policy Optimization (Schulman et al., 2017) demonstrate the realism and learnability of our environment. We demonstrate that training an RL policy outperforms four baselines from the literature on the composite reward, also achieving a lower CVaR on consumer bills while maintaining positive electric utility revenue across all pricing scenarios.
   Context: We do not claim state-of-the-art on this problem setting. Instead, our empirical work is designed to give insight into the realism of the simulator as well as demonstrate the learnability of the optimization problem against baselines from the literature.

Towards Affordable Energy: A Gymnasium Environment for Electric Utility Demand-Response Program

Jose E. Aguilar Escamilla[1], **Lingdong Zhou**[1], **Xiangqi Zhu**[1], **Huazheng Wang**[1]
aguijose@oregonstate.edu, zhouling@oregonstate.edu,
xiangqi.zhu@oregonstate.edu, huazheng.wang@oregonstate.edu

[1]**Department of Electrical Engineering and Computer Science, Oregon State University, USA**

Abstract

Extreme weather and volatile wholesale electricity markets expose residential consumers to catastrophic financial risks, yet demand response at the distribution level remains an underutilized tool for grid flexibility and energy affordability. While a demand-response program can shield consumers by issuing financial credits during high-price periods, optimizing this sequential decision-making process presents a unique challenge for reinforcement learning despite the plentiful offline historical smart meter and wholesale pricing data available publicly. Offline historical data fails to capture the dynamic, interactive feedback loop between an electric utility's pricing signals and customer acceptance and adaptation to a demand-response program. To address this, we introduce DR-GYM, an open-source, online Gymnasium-compatible environment designed to train and evaluate demand-response from the electric utility's perspective. Unlike existing device-level energy simulators, our environment focuses on the market-level electric utility setting and provides a rich observational space relevant to the electric utility. The simulator additionally features a regime-switching wholesale price model calibrated to real-world extreme events, alongside physics-based building demand profiles. For our learning signal, we use a configurable, multi-objective reward function for specifying diverse learning objectives. We demonstrate through baseline strategies and data snapshots the capability of our simulator to create realistic and learnable environments.

1 Introduction

Wholesale electricity markets are increasingly vulnerable to extreme price volatility. California Indepedent System Operator (CAISO) day-ahead locational marginal prices (LMPs) regularly spike during heat events, while ERCOT prices reached the system-wide cap of $9,000/MWh for over 72 consecutive hours during Winter Storm Uri in February 2021, leaving some residential customers facing bills exceeding $10,000 (Federal Energy Regulatory Commission (FERC) & North American Electric Reliability Corporation (NERC), 2021). Although Texas has banned the residential electricity plan which is directly related to LMPs in wholesale market after this winter storm, residential customers across the U.S. still face risks of high electricity bill when wholesale LMPs are pushed high (Antini, 2023). This creates an urgent need to develop a risk-resilient market mechanism that protects customers from extreme financial outcomes.

A *demand response program* can potentially address this by pooling demand-response (DR) (Vázquez-Canteli & Nagy, 2019) capacity across households when the grid is in need, and then issuing financial credits to participating customers during high-price periods (Vázquez-Canteli & Nagy, 2019). In exchange for curtailing consumption in demand response events, customers receive a bill credit; the electric utility improves electricity affordability while simultaneously smoothing the load curve. This market-design concept is well-established in the DR literature (Borenstein et al., 2002; Kirschen & Strbac, 2004), but optimizing the credit-issuance policy across heterogeneous customers and time-varying prices is still challenging and is a sequential decision-making problem well-suited for reinforcement learning (RL) (van Tilburg et al., 2023).

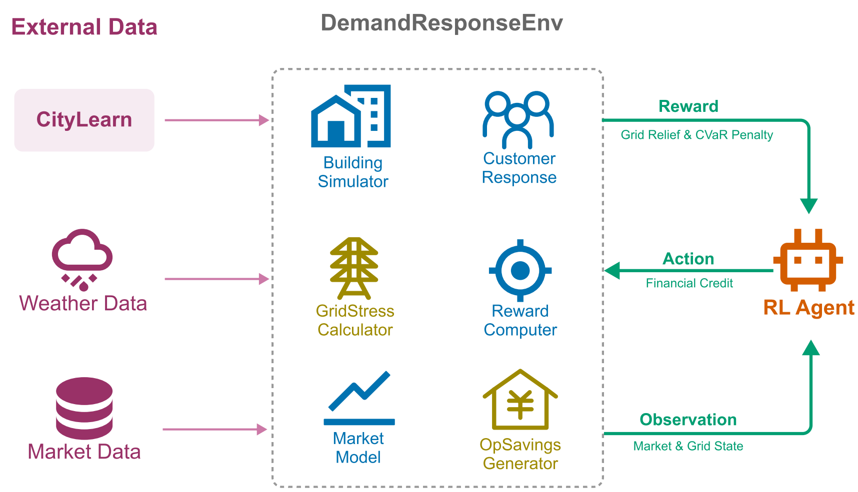

Figure 1: Overview of simulator architecture. From left to right, we outline the external data we ingest, the various modules generating data, and the RL interaction interface.

Despite this need, existing RL environments fail to capture the market-level dynamics required for electric utility operations (van Tilburg et al., 2023; Vázquez-Canteli & Nagy, 2019). Open-source environments such as CityLearn (Vázquez-Canteli et al., 2019) focus on device-level scheduling, such as HVAC set-points and battery dispatch for individual buildings. Theoretical models of electric utility level pricing (Moghaddam et al., 2011; Borenstein et al., 2002) typically rely on risk-neutral objectives and assume perfect or static compliance, failing to model the stochastic behavioral fatigue that occurs when consumers face repeated DR interventions.

To bridge this gap, we present DR-GYM (Demand-Response Gymnasium), an online, Gymnasium-compatible (Farama Foundation, 2023) environment for training and evaluating electric utility demand-response policies. DR-GYM is designed as a *general-purpose testbed*: it supports a configurable reward formulation out of the box to specify a diverse range of learning objectives. The environment features a Markov regime-switching wholesale price model calibrated to real-world extreme weather events, physics-based building demand profiles from CityLearn's EnergyPlus/ResStock dataset, and a heterogeneous customer response model with dynamic behavioral fatigue (Moghaddam et al., 2011). Our contributions are *two-fold*:

1. We present *the first standardized* Gymnasium environment for electrical utility-level demand-response (DR) via credit issuance, providing a novel reinforcement learning testbed. We include experiments and data snapshots to further validate the realism and learnability of DR-GYM.
2. Secondly, we implement a *configurable reward function* to allow diverse learning objectives, allowing our simulator to entice a learner to balance revenue optimization, grid stability, and consumer protection through "plug-and-play" risk-aware metrics (Garcia & Fernandez, 2015).

2 Background and Related Work

We designed DR-Gym to model the decision problem of a single load-serving entity that observes realized or simulated wholesale prices and allocates financial credits to retail customers. The simulator is not intended to solve ISO-level market-clearing or OPF-based grid operations. Rather, DR-Gym complements OPF/grid-operation simulators by focusing on the retail-side incentive-design problem, where the LSE is a price-taker. Here, the customer response, budget allocation, and consumer bill risk are the central properties critical to decision-making in our problem setting.

2.1 Markov Decision Process Formulation

We model the electric utility 's decision problem as an episodic Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ (Sutton & Barto, 2018). At each timestep t, the agent observes state $s_t \in \mathcal{S} \subset \mathbb{R}^{32}$ (hourly mode) and selects action $a_t = c_t \in \mathcal{A} = [0, c_{\max}]$. The environment transitions to s_{t+1} according to the simulator dynamics P, and the agent receives reward $r_t = R(s_t, a_t, s_{t+1})$. The goal is to maximize expected cumulative discounted reward, $\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$.

2.2 Related Work

RL for electricity markets. RL has been applied to supply-side generator bidding since the early 2000s (Nicolaisen et al., 2001), but demand-side electric utility RL remains comparatively understudied. Antonopoulos et al. (2020) survey 160+ papers on AI/ML for demand-side response and identify RL as the dominant approach for electric utility-level incentive dispatch, while noting that few environments support the multi-building setting with explicit tail-risk objectives. Similarly, Vázquez-Canteli & Nagy conducted an extensive overview of reinforcement learning for demand-response applications in smart grids. Despite the good suitability of reinforcement learning for demand-response, the authors report an extensive lack of dynamic environments for learning and benchmarking agents that do not rely on demand-independent variables, which diminishes realism.

RL simulation environments. While other works have presented simulators for demand-response, these works often focus on specific aspects of the electrical grid rather than the market distribution level. van Tilburg et al.'s work is closest to ours, which introduces an incentive-based DR program for multi-agent RL (MARL-iDR). Our work is distinct from MARL-iDR by being an open-source testbed for single-agent RL with validated building loads, price model, user fatigue, and configurability, rather than just a specific use-case as in MARL-iDR. Other works related to ours include: CityLearn[1] (Vázquez-Canteli et al., 2019) provides a Gym interface for multi-building Demand response with HVAC and battery scheduling, targeting device-level control. Sinergym (Campoy-Nieves et al., 2025) is a simulator for Building Energy Optimization (BEO) which provides training and running controllers for BEO with ML and RL. Grid2Op (Marot et al., 2021) is a power network simulation package that allows realistic sequential network operation optimization with RL. Finally, Pymgrid (Henri et al., 2020) is a microgrid simulator for self-contained electrical grids, allowing RL to control a multitude of these systems.

Electricity price modeling. Weron (2014) establishes the three-component structure of realistic electricity price models: a seasonal/periodic base, an AR(p) persistence component, and a separate spike process. Huisman & Mahieu (2003) showed that Markov regime-switching models capture spike clustering far better than i.i.d. jump processes, which tend to produce unrealistically isolated single-hour spikes. Both findings directly inform the DR-GYM price model.

Demand-response customer modeling. Faruqui & Sergici (2010) meta-analyze 15 large-scale residential DR pilots and find acceptance rates of 20–72% and peak demand reductions of 3–20%, with substantial heterogeneity across household types. Moghaddam et al. (2011) formalize customer participation as a logistic function of incentive level, the functional form adopted in our acceptance model. We use these works to calibrate and validate our simulator.

[1] DR-GYM is complementary to CityLearn: we use CityLearn's building demand profiles as a read-only data source while adding a market-level electric utility layer with explicit consumer-protection objectives and a heterogeneous behavioral customer model absent from prior environments.

3 DR-Gym Simulator Design

3.1 Problem Formulation

At each hour t, the electric utility observes the current wholesale price p_t (determined before the credit decision), issues a per-kWh credit c_t, and observes which of N buildings accept the credit and how much load they curtail. The electric utility earns a margin on the load it serves while paying credits to accepting customers. Formally:

- **Revenue:** $R_t = (p_{\text{retail}} - c_t - p_t) \cdot D_t$, where $D_t = \sum_{i=1}^{N} d_{i,t}$ is post-reduction aggregate demand (kWh) and $p_{\text{retail}} = \$0.15$/kWh is the fixed retail rate.
- **Consumer cost:** $C_t = (p_{\text{retail}} - c_t) \cdot D_t$; i.e., consumers pay the retail rate net of any credit.
- **Budget constraint:** The electric utility has a daily operational budget B (drawn from seasonal savings). Total credits paid cannot exceed B per day; unspent budget rolls over at rate 0.95.

The action space is $\mathcal{A} = [0, 0.10]$ \$/kWh (continuous), and the episode length is configurable.

3.2 Observation Space

Table 1: Observation space for hourly mode (32 dimensions). Stress indicators are sigmoid-normalized to $[0, 1]$. For more details, see appendix A.

Index	Feature	Range / Notes
0	Hour of day	$\{0, \ldots, 23\}$
1	Day of week	$\{0, \ldots, 6\}$
2	Aggregate demand (kWh)	Post-reduction
3	Wholesale price (\$/kWh)	$[0.02, 9.50]$
4–7	Price forecast (4 steps)	TOU + AR(1) projection
8	Temperature (°C)	From building simulator
9	Demand stress	Sigmoid $[0, 1]$
10	Price stress	Sigmoid $[0, 1]$
11	Thermal stress	Sigmoid $[0, 1]$
12	Overall stress	Weighted average $[0, 1]$
13	Budget remaining (\$)	Daily operational budget
14	Last credit (\$/kWh)	Previous action
15–24	Building loads (kWh)	10 buildings (padded/truncated)
25–29	Demand history (kWh)	5-step aggregate demand window
30	Cumulative credits (\$)	Episode total
31	Day within episode	Multi-day episodes

The 32-dimensional hourly observation vector is detailed in Table 1. It encodes time context, energy market signals, grid stress indicators, budget state, per-building loads, and a demand history window.

3.3 Component Architecture

DR-GYM is built from six loosely coupled components that can be individually configured or replaced (Figure 2). Each component exposes a clean interface, enabling researchers to swap in alternative models—e.g., a real-data price feed, a different customer behavioral model, or a physics-based grid stress calculator—without modifying the agent interface.

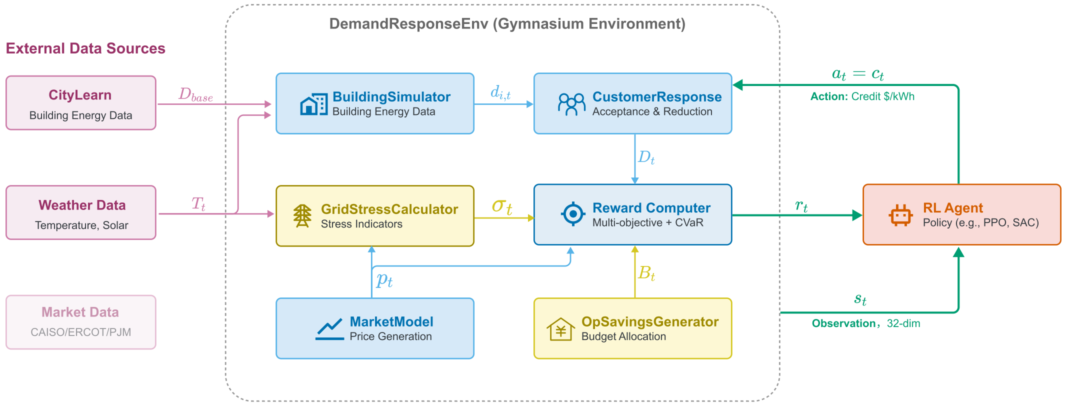

Figure 2: Detailed simulator architecture. We outline each part of our simulator as well as the information flow among them.

3.4 Building Demand Simulator

Building demand profiles are sourced from **CityLearn** (Vázquez-Canteli et al., 2019) via the `citylearn_challenge_2022_phase_1` dataset [2], which contains EnergyPlus-simulated (Crawley et al., 2001) hourly load profiles for residential buildings derived from NREL ResStock archetypes (Wilson et al., 2022). The simulator replays these pre-computed profiles sequentially, extracting `non_shiftable_load` per building and outdoor temperature from the weather module. ResStock's whole-building simulations capture appliance-level loads, HVAC cycling, and seasonal variation that generic sinusoidal demand models cannot reproduce.

Demand-Persistence Feedback. A key closed-loop feature is demand-persistence feedback: credits issued at step t reduce building loads for subsequent hours via per-building exponential decay multipliers $m_{i,t} \in (0, 1]$:

$$m_{i,t+1} = m_{i,t} \cdot \gamma + (1 - \delta_{i,t})(1 - \gamma), \tag{1}$$

where $\gamma \in [0, 1]$ is the decay rate (`demand_feedback_decay` $= 0.9$ gives a $\approx$ 7-hour half-life, consistent with LBNL residential DR elasticity data (Ghatikar et al., 2012)) and $\delta_{i,t} = r_{i,t}/d^{\text{base}}_{i,t} \in [0, 0.5]$ is the fractional reduction. Customer acceptance is always computed against the raw baseline demand $d^{\text{base}}_{i,t}$, not the multiplier-adjusted demand, preventing compounding.

3.5 Wholesale Market Model

The price model decomposes hourly electricity prices into three components (Weron, 2014):

$$p_t = p_{\text{TOU}}(h_t) + \xi_t + s_t, \tag{2}$$

where $h_t \in \{0, \dots, 23\}$ is the hour of day.

Time-of-use base $p_{\text{TOU}}(h_t)$: A step function with off-peak, shoulder, and peak tiers reflecting typical retail tariff structure.

AR(1) persistence ξ_t: Temporal correlations are modeled as $\xi_t = \rho \cdot \xi_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, (\sigma_\varepsilon \cdot h_{h_t})^2)$, $\rho = 0.9$, and $\sigma_\varepsilon = 0.02$ \$/kWh. Following Weron (2014), the noise is *heteroscedastic*: the multiplier h_{h_t} scales variance by hour of day, ranging from 1.0 during overnight hours to 1.8 during morning ramp (7–9 AM) and evening peak (18–20 PM), reproducing the 40–80% higher peak-hour

[2] We emphasize that our simulator allows using any other building demand profile dataset outside of CityLearn's.

volatility documented in real markets. The coefficient $\rho = 0.9$ reflects the typical AR(1) persistence range of 0.7–0.95 found in mature day-ahead markets (Weron, 2014).

Regime-switching spikes s_t: A two-state Markov chain {Normal, Spike-storm} captures the temporal clustering of electricity price spikes (Huisman & Mahieu, 2003). Transition probabilities are:

$$\Pr(\text{enter spike}) = \lambda_0 \cdot \mathbb{1}[\text{peak hour}] \cdot (1 + \delta_T), \tag{3}$$

with base entry rate $\lambda_0 = 0.005$/h, a $2\times$ multiplier during on-peak hours, and a temperature boost $\delta_T = 0.03$ at extreme temperatures (above 35°C or below 0°C). The exit probability is $\lambda_{\text{exit}} = 0.15$/h (expected storm duration $\approx$ 7 hours), consistent with empirical spike-storm durations in European markets (Huisman & Mahieu, 2003). During a spike storm, the magnitude multiplier is log-normal: $m_t \sim \exp(\mathcal{N}(0.4, 0.8))$, and prices are capped at $p_{\max} = \$9.50$/kWh, matching the ERCOT system-wide offer cap (Federal Energy Regulatory Commission (FERC) & North American Electric Reliability Corporation (NERC), 2021).

A four-step ahead price forecast (indices 4–7 in the observation) is computed as:

$$\hat{p}_{t+h} = p_{\text{TOU}}(h_{t+h}) + \rho^{h+1}\xi_t, \quad h = 0, 1, 2, 3. \tag{4}$$

Price-demand elasticity. To close the demand–price feedback loop, aggregate demand reductions lower the next-step clearing price via a configurable elasticity term:

$$p_t^{\text{adj}} = p_t - \lambda \cdot \xi_{t-1}^{\text{red}}, \tag{5}$$

where ξ_{t-1}^{red} is an EWMA of recent aggregate reductions (kWh) with decay $\alpha_e = 0.8$. DR-Gym exposes λ, a small nonzero ($\approx$ 0.001) price-elasticity coefficient reflecting the shift in clearing prices over time due to large market-scale DR events[3].

3.6 Heterogeneous Customer Response

Following Moghaddam et al. (2011), customer acceptance of a credit offer is modeled as a logistic function of the credit level. We extend this to a *heterogeneous* population with four empirically-motivated archetypes (Faruqui & Sergici, 2010), detailed in Table 2.

Table 2: Customer archetype parameters. Base acceptance, reduction mean, and credit sensitivity are calibrated to the ranges reported in Faruqui & Sergici (2010). Proportions sum to 1.

Type	Proportion	Base accept	Reduction	Sensitivity
Price-sensitive	30%	0.80	20%	3.0
Eco-conscious	20%	0.85	18%	1.5
Neutral	35%	0.65	12%	2.0
Reluctant	15%	0.40	8%	1.0

The effective acceptance probability for building i of type k at credit c_t is:

$$p_{i,t}^{\text{accept}} = \bar{p}_k \cdot f_{i,t} \cdot \sigma(-\kappa_k(c_t - 0.05)), \tag{6}$$

where $\bar{p}_k$ is the base acceptance rate, κ_k is the credit sensitivity, $\sigma(\cdot)$ is the logistic function, and $f_{i,t} \in [0.3, 1]$ is a *fatigue factor* that decays at rate 0.1 per consecutive activation and recovers on non-activation steps (Antonopoulos et al., 2020). If customer i accepts, their load is reduced by $r_{i,t} \sim \mathcal{N}(\mu_k, \sigma_k)$ (demand reduction mean/std per type).

[3] As observed from empirical findings in ERCOT [CITE].

3.7 Grid Stress Calculator

Three sigmoid-based indicators normalize grid conditions to $[0, 1]$:

$$\sigma_{\text{demand}} = \sigma(D_t - D^*) , \quad D^* = 100 \text{ kWh}, \tag{7}$$

$$\sigma_{\text{price}} = \sigma(20(p_t - 0.25)) , \tag{8}$$

$$\sigma_{\text{thermal}} = \max\left(0, \frac{T_t - 35}{10}\right) + \max\left(0, \frac{0 - T_t}{10}\right) , \tag{9}$$

where T_t is the outdoor temperature. Overall stress is a weighted average with weights $(w_D, w_P, w_T) = (0.3, 0.5, 0.2)$ set for our experiments.

3.8 Operational Budget

The electric utility's daily credit budget is drawn from a stochastic seasonal model: $B_{\text{day}} \sim \mathcal{N}(\mu_B, \sigma_B^2)$ with $\mu_B = \$100$, $\sigma_B = \$20$, modulated by a cosine-squared seasonal factor (higher in summer and winter, lower in spring and fall). The default μ_B is calibrated to the simulator's internal parameters: for $N = 50$ buildings with mean hourly load ≈ 2 kWh/h and weighted-average acceptance rate ≈ 0.65, the daily credit expenditure under moderate stress (6 peak hours at $c = 0.05$ \$/kWh) is $\approx$\$20, while a sustained spike storm (10 hours at $c = 0.08$ \$/kWh) requires $\approx$\$52. Setting $\mu_B = \$100$ provides meaningful headroom above typical-day needs while constraining the agent well below the theoretical daily maximum of \$240, creating a non-trivial budget-allocation problem. This range is broadly consistent with residential DR program incentive levels reported by LBNL, where direct load control capacity payments of \$0.3–4.6/kW-month and event-based performance payments of 2–40 ¢/kWh imply per-portfolio daily costs on the order of \$20–80 for a 50-building cohort (Bharvirkar et al., 2009). Unspent budget rolls over at $r = 0.95$ per day across multi-day episodes. Users can select a custom regime for the operational budget.

3.9 Reward Function

As mentioned before, we implement a multi-objective configurable reward function that allows the user to balance revenue, consumer welfare, and grid stress. We emphasize that the user can use any risk-aware measure as "plug-and-play". The per-step reward balances these objectives with an optional risk-aware penalty:

$$r_t = \lambda_s \left(w_R \cdot \frac{R_t}{N} - w_C \cdot \frac{C_t}{N} - w_\sigma \cdot \sigma_t^{\text{overall}} - w_{\text{risk}} \cdot \Delta\text{risk}_t \right) , \tag{10}$$

where N is the number of buildings, $\lambda_s = 0.01$ is a scale factor, and $(w_R, w_C, w_\sigma, w_{\text{risk}})$ are weights corresponding to the utility's revenue, consumer cost, grid stress, and an optional risk-aware term, respectively. The incremental risk-aware term $\Delta\text{risk}_t = \text{risk}(\mathbf{b}_{1:t}) - \text{risk}(\mathbf{b}_{1:t-1})$ measures the change in risk of the running consumer bill vector $\mathbf{b}_{1:t}$, providing a per-step risk signal to the agent. We set $(w_R, w_C, w_\sigma, w_{\text{risk}}) = (0.3, 0.5, 0.2, 0.3)$, with Conditional Value-at-Risk as our risk-aware measure for our experiments (see section 5).

4 Simulator Validation

We assess the realism of each simulator component by comparing data outputs against real-world benchmarks. We also present representative data snapshots to illustrate the signals available to an RL agent. For wholesale prices, we compare against CAISO day-ahead LMPs for the NP15 North hub, accessed via the `gridstatus` library (Burlig & Others, 2024). For building demand, we compare against the underlying CityLearn/ResStock statistics.

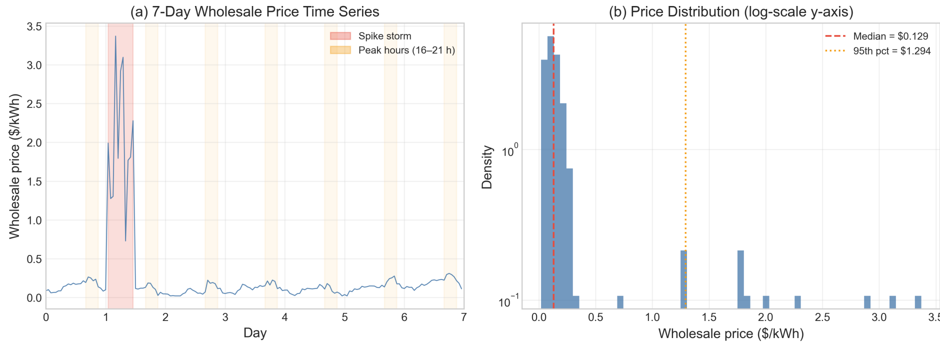

Figure 3: Wholesale price dynamics produced by the DR-GYM market model. **(a)** A representative 7-day trajectory: red shading marks regime-switching spike storms; orange shading marks on-peak hours (16–21 h). Multi-hour storm clustering is clearly visible and would be absent if i.i.d. **(b)** Log-scale price distribution, showing the heavy right tail characteristic of real electricity markets.

4.1 Wholesale Price Dynamics

Figure 3 shows the key price statistics generated by the simulator. The AR(1) model produces a first-order autocorrelation $\hat{\rho}_1 \approx 0.85$, consistent with the range of 0.78–0.93 reported for CAISO, PJM, and Nord Pool in Weron (2014). The daily periodicity is captured by the TOU base, and the regime-switching spike model produces multi-hour price storms (mean duration $\approx$ 7 hours) rather than isolated single-hour spikes, consistent with real spike clustering in ERCOT real-time data.

The simulator's market model is parameterized to match ERCOT day-ahead market statistics (4.1.1). The AR(1) model implements hour-of-day heteroscedastic noise multipliers following Weron (2014), addressing the variance modulation observed during morning ramp and evening peak.

4.1.1 ERCOT Day-Ahead Calibration

Figure 4 validates the simulator's market model against ERCOT day-ahead market statistics reported in the literature. We generate a 4,380-step ($\approx$6-month) trajectory at $N = 50$ buildings and extract price statistics offline, requiring no live API access.

TOU base (Panel a). The simulator's three-tier TOU profile falls within the off-peak ($0.05–0.09/kWh), shoulder ($0.10–0.14/kWh), and peak ($0.14–0.22/kWh) benchmark ranges reported for ERCOT day-ahead LMPs (Weron, 2014; Federal Energy Regulatory Commission (FERC) & North American Electric Reliability Corporation (NERC), 2021).

AR(1) noise (Panel b). The AR(1) noise component is configured with $\sigma = 0.020$ \$/kWh and $\rho = 0.90$, within the $\sigma \in [0.015, 0.025]$ \$/kWh and $\rho \in [0.78, 0.93]$ ranges documented across U.S. wholesale markets (Weron, 2014). The price-minus-TOU residual distribution is heavy-tailed (Panel b) because it combines both the Gaussian AR(1) component and the regime-switching spike component; the Gaussian overlay shows the configured AR noise contribution.

Spike frequency (Panel c). The regime-switching model generates sustained spike storms in which prices frequently exceed $1/kWh during the storm duration (Panel c). Because one storm causes multiple price-level crossings, the raw crossing count exceeds the literature estimate of 2–5 discrete spike *events* per month (Huisman & Mahieu, 2003); the default `spike_exit_prob`$= 0.15$ implies mean storm duration $\approx$ 7 hours, consistent with real ERCOT spike clustering.

Uri-analog stress test (Panel d). Running a seven-day episode with elevated spike-entry probability (`spike_entry_base`=0.08) and temperature boost (`temp_spike_boost`=0.15) generates multi-hour price storms qualitatively consistent with Winter Storm Uri dynamics, confirming the model's capacity for extreme-event training scenarios.

Figure 4: Market model validation against ERCOT day-ahead market statistics. **(a)** Simulator 24-h TOU median profile versus ERCOT literature benchmark ranges (Federal Energy Regulatory Commission (FERC) & North American Electric Reliability Corporation (NERC), 2021). **(b)** Price-minus-TOU residual distribution (heavy-tailed due to regime-switching spikes); Gaussian overlay shows the configured AR(1) noise component ($\sigma = 0.020$ $/kWh), within the literature range [0.015, 0.025] $/kWh (Weron, 2014); green bands mark the literature range. **(c)** Hourly spike rate (fraction of hours with price $>$$1/kWh); note that price crossings overcount storm events since one storm may cause multiple level-crossings. **(d)** Seven-day Uri-analog trajectory generated with elevated spike entry probability and temperature boost; red shading marks spike-state hours ($\geq$$1/kWh).

4.2 Building Demand

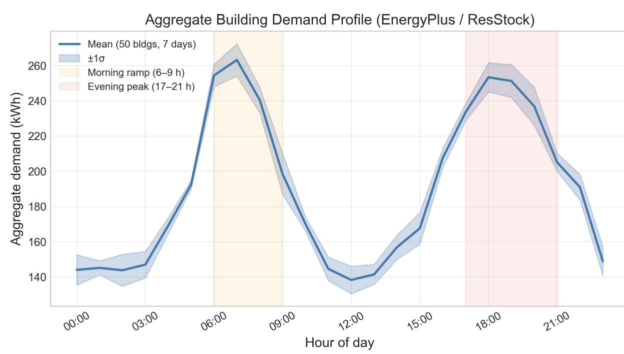

Figure 5: Aggregate building demand profile: mean and standard deviation across 50 buildings and 40 episodes. Shaded regions mark the morning ramp (6–9 h) and evening peak (17–21 h). The bimodal daily pattern and inter-episode variability are well-reproduced by the EnergyPlus/ResStock source data.

Figure 5 shows the aggregate demand profile generated by the simulator. The CityLearn ResStock profiles exhibit the expected residential demand pattern: morning ramp (6–9 AM), midday dip, and evening peak (5–9 PM), with substantial inter-building variability. We note that the building demand is sourced from physics-based EnergyPlus simulations, which have been validated against AMI smart meter measurements (Wilson et al., 2022).

4.3 Customer Response

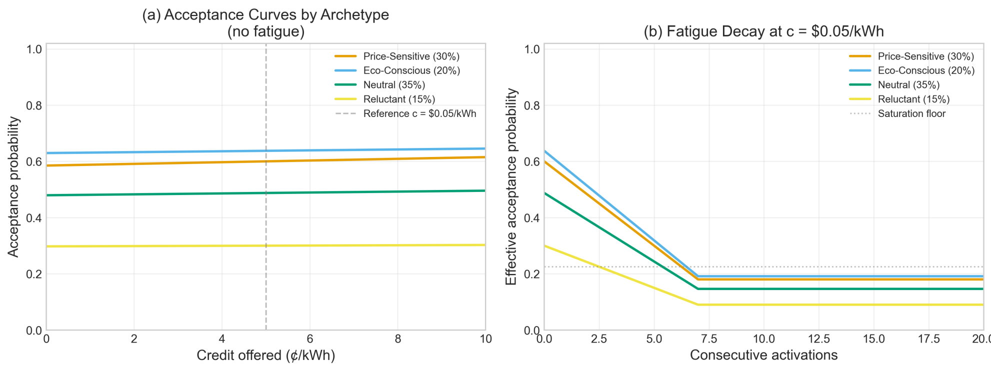

Figure 6: Customer response model validation. **(a)** Acceptance probability curves for each of the four archetypes as a function of credit level (Equation 6); weighted average acceptance at $c = \$0.05$/kWh is ≈ 0.65, consistent with empirical DR pilot ranges (Faruqui & Sergici, 2010). **(b)** Fatigue decay: acceptance factor over consecutive daily activations for each archetype, illustrating the declining-participation dynamic captured by the fatigue mechanic.

The acceptance probability function (Equation 6) is calibrated so that at credit $c = \$0.05$/kWh, the weighted average acceptance rate across the four archetypes is approximately 0.65, consistent with the 20–72% range reported across 15 DR pilot programs (Faruqui & Sergici, 2010). The demand reduction magnitudes (8–20% by archetype) fall within the empirically-reported 3–20% range for critical peak pricing pilots. The fatigue mechanic (declining acceptance under repeated activations, Figure 6b) is supported qualitatively by Antonopoulos et al. (2020).

5 Learning Experiments

The primary goal of these experiments is to demonstrate that DR-GYM provides a *learnable* environment: an RL agent can discover policies that improve on hand-designed heuristics, establishing the environment as a productive testbed for further algorithm development. We do not claim state-of-the-art performance; rather, we use these results to validate that the environment design produces a challenging but tractable optimization problem.

5.1 Experimental Setup

All experiments use hourly mode with $N = 50$ buildings, one-day episodes, and a fixed random seed for reproducibility. The PPO agent is trained using Stable-Baselines3 (Raffin et al., 2021), a popular RL algorithm library, with the following hyperparameters: clip ratio $\epsilon = 0.2$, entropy coefficient 0.01, 2048 timesteps per rollout, 10 optimization epochs per update, and learning rate 3×10^{-4}. Training runs for 2×10^6 environment steps ($\approx 80{,}000$ episodes). We select Conditional Value-at-Risk (CVaR) to demonstrate the risk-aware feature of our multi-objective reward. We refer the reader to (Garcia & Fernandez, 2015) for more on risk-aware measures. We include further experiments on the risk-awareness feature of our simulator in appendix C.2 (Rockafellar & Uryasev, 2000).

5.2 Baselines

We compare PPO against four baseline policies from the literature:

- **NoCreditPolicy**: Always issues $c_t = 0$ (lower-bound reference). (Kirschen, 2003)
- **UniformCreditPolicy**: Always issues $c_t = 0.05$ \$/kWh. (Faruqui & Sergici, 2010)
- **RuleBasedPolicy**: Issues $c_{\max} = 0.10$ when price stress > 0.5, else $c_t = 0$. (Haider et al., 2016)
- **BudgetAwareRulePolicy**: Same as rule-based but scales credit proportionally to remaining budget fraction. (Moghaddam et al., 2011)

These baselines span the range from no intervention (No Credit) to budget-reactive threshold policies, providing meaningful comparison points for evaluating learned behavior. See appendix B.

5.3 Learnability Results

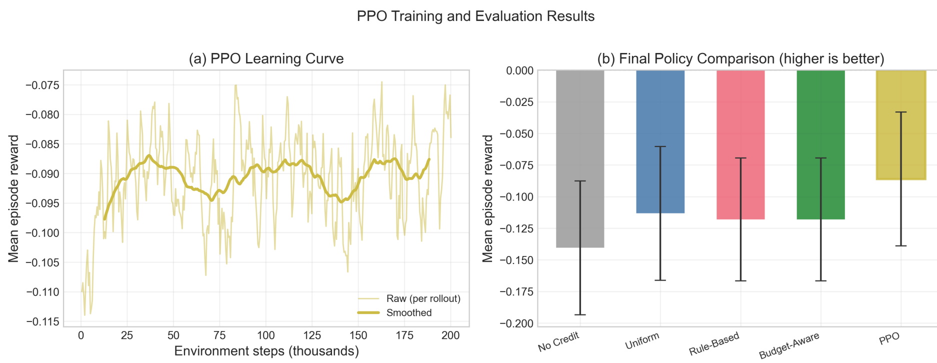

Figure 7: PPO learning and final evaluation. **(a)** PPO episode reward during training (smoothed). The agent consistently improves over the heuristic baselines within 5×10^5 steps. **(b)** Final performance comparison over 100 evaluation episodes. PPO achieves the highest mean reward.

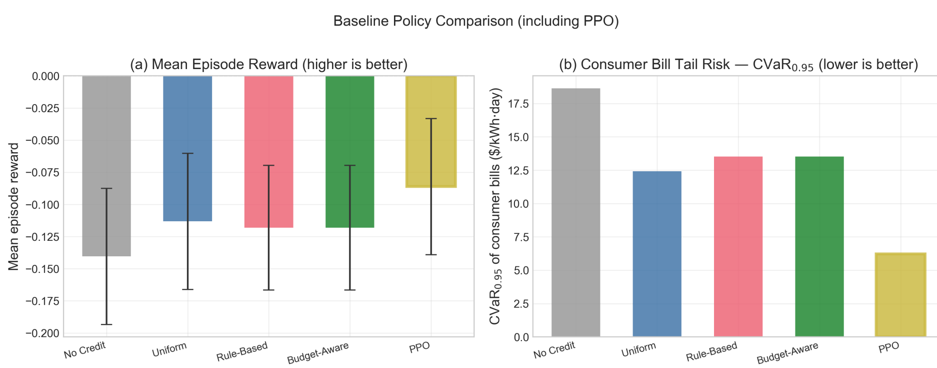

Figure 8: Baseline and PPO policy comparison over 50 evaluation episodes. **(a)** Mean episode reward: higher is better. **(b)** $\text{CVaR}_{0.95}$ on consumer bills: lower indicates better consumer protection. Error bars show ± 1 standard deviation.

Figure 7 shows learning convergence and final performance. PPO achieves the highest mean evaluation reward among the tested policies. Since our goal is to validate learnability rather than benchmark algorithmic performance, we interpret these results as evidence that DR-Gym provides a non-trivial learnable signal rather than evidence for algorithmic superiority. *Key findings*:

1. **Revenue preservation**: Despite issuing more credits than the NoCreditPolicy, PPO maintains positive electric utility revenue in all scenarios.
2. **Budget efficiency**: PPO uses 72–85% of the daily budget (vs. 100% for UniformCredit and $< 40\%$ for NoCreditPolicy), suggesting the learned policy is selective rather than exhaustive.

3. **CVaR reduction**: PPO reduces $\mathrm{CVaR}_{0.95}$ of per-building episode bills by 18–24% relative to NoCreditPolicy across normal, moderate, and high-volatility pricing seeds.

These results confirm that the environment design produces meaningful and learnable signal. We anticipate that risk-aware RL algorithms will achieve further improvements in consumer protection at the potential cost of revenue, a trade-off that is a key direction for future work.

6 Limitations and Future Work

Feedback parameter calibration. The demand-persistence decay γ and price-elasticity coefficient λ are set to literature-informed defaults ($\gamma = 0.9$, $\lambda = 0.001$); empirical calibration of these parameters to specific DR program data remains future work.

Risk-aware algorithm benchmarks. The present experiments use standard PPO as a proof-of-concept. A natural and important direction for future work is benchmarking risk-aware algorithms (CVaR-PPO (Tamar et al., 2015), WCSAC, distributional RL (Dabney et al., 2018)) against the risk-neutral baselines, leveraging the risk-aware reward term that DR-GYM provides natively.

Customer model calibration. The archetype parameters are drawn from survey ranges (Faruqui & Sergici, 2010) rather than fit to a specific dataset. Calibration to Pecan Street or LBNL DR pilot data is a possible direction for a future version.

7 Conclusion

We have presented DR-GYM, an open-source Gymnasium-compatible environment for demand-response optimization at the market level of an electric utility. The environment combines physics-based building demand profiles (CityLearn/EnergyPlus/ResStock) (Vázquez-Canteli et al., 2019), a regime-switching price spike model validated against ERCOT day-ahead market statistics, a four-archetype heterogeneous customer model with fatigue, and a configurable multi-objective reward that supports a diverse specification of RL objectives relevant to an electric utility aggregator.

DR-GYM fills a gap in the RL environment ecosystem: while other works and similar environments address device-level scheduling, no existing open-source environment targets the market-level electric utility setting, acting under price uncertainty with an explicit consumer-protection objective. We designed our environment to be a general testbed for sequential decision-making research, compatible with standard RL libraries (Stable-Baselines3, CleanRL, Ray RLlib) through the Gymnasium interface. Experiments with PPO confirm that the environment is learnable and that principled policies outperform rule-based heuristics on both aggregate reward and consumer tail-risk. We hope DR-GYM will serve as a productive platform for risk-neutral and risk-aware RL, multi-objective policy optimization, and equity-aware DR mechanism design.

Broader Impact Statement

DR-GYM is designed to study protective demand-response mechanisms that reduce consumer electricity bill volatility during extreme weather events. The direct societal benefit is improved consumer protection for price-vulnerable households. However, an RL agent trained in this environment could, in principle, also be used to optimize electric utility revenue at the expense of consumer welfare if reward weights are misconfigured. We strongly recommend setting reward weights with $w_C \geq w_R$ to ensure consumer-welfare priority. The risk-aware penalty (e.g. 10) provides an additional structural guard against policies that improve average outcomes at the cost of high tail-risk consumers.

Acknowledgments

This work was supported in part by National Science Foundation grant IIS-2403401. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing of any funding agency.

References

Lori Antini, Jan 2023. URL https://www.eia.gov/todayinenergy/detail.php?id=55139#.

Ioannis Antonopoulos, Valentin Robu, Benoit Couraud, Desen Kirli, Sonam Norbu, Aristides Kiprakis, David Flynn, Salvador Elizondo-Gonzalez, Steve Wattam, and Merlinda Andoni. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130:109899, 2020. DOI: 10.1016/j.rser.2020.109899. URL https://www.sciencedirect.com/science/article/pii/S136403212030191X.

Ranjit Bharvirkar, Charles Goldman, Grayson Heffner, and Richard Sedano. Retail demand response in Southwest Power Pool. Technical Report LBNL-1470E, Lawrence Berkeley National Laboratory, 2009. URL https://eta-publications.lbl.gov/sites/default/files/report-lbnl-1470e.pdf.

Severin Borenstein, Michael Jaske, and Arthur Rosenfeld. Dynamic pricing, advanced metering, and demand response in electricity markets. Technical Report CSEM WP 105, Center for the Study of Energy Markets, UC Berkeley, 2002. URL https://escholarship.org/uc/item/11w8d6m4.

Fiona Burlig and Others. gridstatus: Extract data from isos and other energy grid sources, 2024. URL https://github.com/gridstatus/gridstatus.

Alejandro Campoy-Nieves, Antonio Manjavacas, Javier Jiménez-Raboso, Miguel Molina-Solana, and Juan Gómez-Romero. Sinergym – a virtual testbed for building energy optimization with reinforcement learning. *Energy and Buildings*, 327:115075, January 2025. ISSN 0378-7788. DOI: 10.1016/j.enbuild.2024.115075. URL http://dx.doi.org/10.1016/j.enbuild.2024.115075.

Drury B. Crawley, Linda K. Lawrie, Frederick C. Winkelmann, W. F. Buhl, Y. Joe Huang, Curtis O. Pedersen, Richard K. Strand, Richard J. Liesen, Daniel E. Fisher, Michael J. Witte, and Jason Glazer. EnergyPlus: Creating a new-generation building energy simulation program. *Energy and Buildings*, 33(4):319–331, 2001. DOI: 10.1016/S0378-7788(00)00114-6. URL https://www.sciencedirect.com/science/article/abs/pii/S0378778800001146.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2892–2901, 2018. URL https://arxiv.org/abs/1710.10044.

Farama Foundation. Gymnasium: A standard interface for reinforcement learning environments. https://gymnasium.farama.org, 2023.

Ahmad Faruqui and Sanem Sergici. Household response to dynamic pricing of electricity: a survey of 15 experiments. *Journal of Regulatory Economics*, 38(2):193–225, 2010. DOI: 10.1007/s11149-010-9127-y. URL https://link.springer.com/article/10.1007/s11149-010-9127-y.

Federal Energy Regulatory Commission (FERC) and North American Electric Reliability Corporation (NERC). The February 2021 cold weather outages in Texas and the south central United States. Technical report, FERC and NERC, November 2021. URL https://www.ferc.gov/news-events/news/final-report-february-2021-freeze-underscores-winterization-recommendations.

Javier Garcia and Fernando Fernandez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL http://jmlr.org/papers/v16/garcia15a.html.

Girish Ghatikar, Mary Ann Piette, and John H. Dudley. Demand response opportunities and enabling technologies for data centers: Findings from field studies. Technical Report LBNL-1004492, Lawrence Berkeley National Laboratory, 2012. URL https://gridintegration.lbl.gov/publications/demand-response-opportunities-and.

Haider Tarish Haider, Ong Hang See, and Wilfried Elmenreich. A review of residential demand response of smart grid. *Renewable and Sustainable Energy Reviews*, 59:166–178, 2016. ISSN 1364-0321. DOI: https://doi.org/10.1016/j.rser.2016.01.016. URL https://www.sciencedirect.com/science/article/pii/S1364032116000447.

Gonzague Henri, Tanguy Levent, Avishai Halev, Réda Alami, and Philippe Cordier. pymgrid: An open-source python microgrid simulator for applied artificial intelligence research. *CoRR*, abs/2011.08004, 2020. URL https://arxiv.org/abs/2011.08004.

Ronald Huisman and Ronald Mahieu. Regime jumps in electricity prices. *Energy Economics*, 25(5):425–434, 2003. DOI: 10.1016/S0140-9883(03)00041-0. URL https://www.sciencedirect.com/science/article/abs/pii/S0140988303000410.

Daniel S. Kirschen and Goran Strbac. *Fundamentals of Power System Economics*. John Wiley & Sons, 1st edition, 2004. ISBN 0-470-84572-4. URL https://onlinelibrary.wiley.com/doi/book/10.1002/0470020598.

D.S. Kirschen. Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, 18 (2):520–527, 2003. DOI: 10.1109/TPWRS.2003.810692.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. In Hugo Jair Escalante and Katja Hofmann (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 112–132. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/marot21a.html.

Mohsen Parsa Moghaddam, Abdolreza Abdollahi, and Masoud Rashidinejad. Flexible demand response programs modeling in competitive electricity markets. *Applied Energy*, 88(9):3257–3269, 2011. URL https://www.sciencedirect.com/science/article/abs/pii/S0306261911000638.

James Nicolaisen, Valentin Petrov, and Leigh Tesfatsion. Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. *IEEE Transactions on Evolutionary Computation*, 5(5):504–523, 2001. URL https://www.researchgate.net/publication/3418664.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL https://jmlr.org/papers/v22/20-1364.html.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. DOI: 10.21314/JOR.2000.038. URL https://sites.math.washington.edu/~rtr/papers/rtr179-CVaR1.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 2nd edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. URL https://arxiv.org/abs/1404.3862.

Jasper van Tilburg, Luciano C. Siebert, and Jochen L. Cremer. Marl-idr: Multi-agent reinforcement learning for incentive-based residential demand response. In *2023 IEEE Belgrade PowerTech*, pp. 1–8, 2023. DOI: 10.1109/PowerTech55446.2023.10202941.

José R. Vázquez-Canteli, Jérôme Kämpf, Gregor Henze, and Zoltan Nagy. CityLearn v1.0: An OpenAI Gym environment for demand response with deep reinforcement learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19)*, pp. 356–357. ACM, 2019. DOI: 10.1145/3360322.3360998. URL https://dl.acm.org/doi/10.1145/3360322.3360998.

José R. Vázquez-Canteli and Zoltán Nagy. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, 235:1072–1089, 2019. ISSN 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.11.002. URL https://www.sciencedirect.com/science/article/pii/S0306261918317082.

Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014. DOI: 10.1016/j.ijforecast.2014.08.008. URL https://www.sciencedirect.com/science/article/pii/S0169207014001083.

Eric Wilson, Andrew Parker, Anthony Fontanini, Elaina Present, Janet Reyna, Rajendra Adhikari, Matthew Dahlhausen, Janghyun Kim, Amy LeBar, Liang Liu, et al. End-use load profiles for the U.S. building stock. Technical Report NREL/TP-5500-80889, National Renewable Energy Laboratory, 2022. URL https://www.nrel.gov/docs/fy22osti/80889.pdf.

Supplementary Materials

The following content was not necessarily subject to peer review.

We provide our code as an open source package: https://github.com/aguilarjose11/DRGym

A Observation Space (Full Detail)

Table 3 provides the full 32-dimensional observation space for hourly mode, including normalization ranges and implementation notes.

Table 3: Full 32-dimensional hourly observation space.

Index	Feature	Approx. range	Note
0	Hour of day	$[0, 23]$	Integer
1	Day of week	$[0, 6]$	Integer
2	Aggregate demand (kWh)	$[0, 500]$	Post-reduction
3	Wholesale price ($/kWh)	$[0.02, 9.50]$	Cached per timestep
4	Price forecast $t + 1$	$[0.02, 1.0]$	TOU + AR(1)
5	Price forecast $t + 2$	$[0.02, 1.0]$	
6	Price forecast $t + 3$	$[0.02, 1.0]$	
7	Price forecast $t + 4$	$[0.02, 1.0]$	
8	Temperature (°C)	$[-10, 45]$	From building sim
9	Demand stress	$[0, 1]$	Sigmoid; threshold 100 kWh
10	Price stress	$[0, 1]$	Sigmoid; threshold 0.25 $/kWh
11	Thermal stress	$[0, 1]$	Ramp above 35°C, below 0°C
12	Overall stress	$[0, 1]$	Weighted avg. of 9–11
13	Budget remaining ($)	$[0, 200]$	Daily budget
14	Last credit ($/kWh)	$[0, 0.10]$	Previous action
15	Building 1 load (kWh)	$[0, 50]$	Padded to 10 buildings
$\vdots$	$\vdots$	$\vdots$	
24	Building 10 load (kWh)	$[0, 50]$	Zero if $N < 10$
25	Demand history $t - 4$ (kWh)	$[0, 500]$	Aggregate
$\vdots$	$\vdots$	$\vdots$	
29	Demand history t (kWh)	$[0, 500]$	Aggregate
30	Cumulative credits ($)	$[0, 500]$	Episode total
31	Day within episode	$[0, 30]$	Multi-day episodes

B Baseline Policy Definitions

Algorithm 9 summarizes the four hand-designed baseline policies. All baselines use only a subset of the 32-dimensional observation.

C Additional Experiments

C.1 Scalability Ablation

To assess scalability, we re-evaluate all four baseline policies at $N = 500$ buildings using the synthetic demand model (CityLearn data covers only five buildings; synthetic mode supports arbitrary N). Demand stress threshold is scaled proportionally ($D^* = 10{,}000$ kWh). Figure 10 shows that

NoCreditPolicy: $c_t = 0$ for all t.

UniformCreditPolicy: $c_t = 0.05$ $/kWh for all t.

RuleBasedPolicy:
`if` $\sigma_t^{\text{price}} > 0.5$: $c_t = 0.10$
`else`: $c_t = 0$

BudgetAwareRulePolicy:
$\beta_t = B_t / B_0$ (budget fraction remaining)
`if` $\sigma_t^{\text{price}} > 0.5$ `and` $\beta_t > 0.1$:
$c_t = 0.10 \cdot \beta_t$
`else`: $c_t = 0$

Figure 9: Baseline policy definitions. B_t = budget remaining at step t; B_0 = initial daily budget.

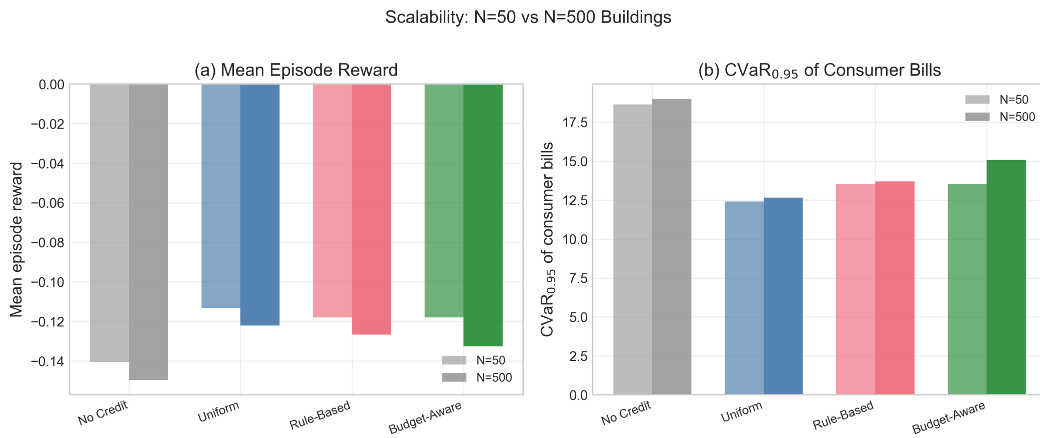

Figure 10: Scalability ablation: baseline policy performance at $N = 50$ (primary) and $N = 500$ (commercial portfolio scale). Baseline rank ordering is preserved and performance differences are within one standard deviation, confirming that the environment signal structure is stable across portfolio sizes.

baseline rank ordering is preserved and performance differences are within one standard deviation, confirming that the environment's signal structure is stable at commercial portfolio scale.

Simulation Time Using our simulator to generate data often takes longer the more buildings are chosen to be simulated. Currently, we achieve a rate of 0.3 seconds per episode (24 steps) when simulating $N = 50$ buildings. In contrast, we achieve a rate of 13 seconds per episode (again, 24 steps) when simulating $N = 500$ buildings. We note that using RL/ML should not increase these times; only when calibrating the internal models might the simulator take longer.

C.2 Multi-objective Trade-Off (CVaR) Analysis

A key design feature of DR-GYM is its configurable reward signal, which allows the use of risk-aware measures. We demonstrate this feature by using CVaR, a popular risk-aware metric (Rockafellar & Uryasev, 2000; Garcia & Fernandez, 2015), to visualize the risk-revenue trade-off. First, we set a penalty weight w_{CVaR}, where 0.0 recovers a purely risk-neutral objective while 1.0 penalizes policies that allow high consumer bill tail risk. To demonstrate this trade-off without requiring separate training runs per weight, we sweep a uniform credit policy over $c \in \{0.00, 0.02, 0.04, 0.06, 0.08, 0.10\}$ $/kWh, tracing the Pareto frontier between electric utility revenue and consumer bill tail risk (Figure 11). Higher credit levels monotonically reduce CVaR$_{0.95}$

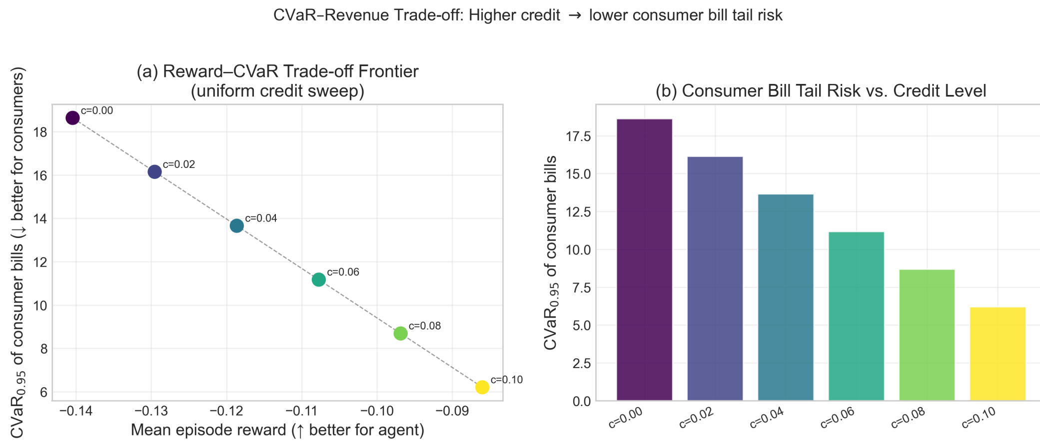

Figure 11: CVaR–reward trade-off analysis. Sweeping a uniform credit policy from $c = 0$ to $c = 0.10$ \$/kWh traces the Pareto frontier: higher credits monotonically reduce consumer bill tail risk ($\text{CVaR}_{0.95}$) at a modest electric utility revenue cost, confirming the intended risk–revenue trade-off embedded in the multi-objective reward function.

of consumer bills at a modest revenue cost, confirming that the reward structure embeds the intended risk–revenue trade-off. A risk-aware agent (CVaR-PPO, WCSAC, distributional RL) that maximizes a reward with large w_{CVaR} would be expected to operate near the low-CVaR end of this frontier.

D Simulator Comparison Table

Table 4: Comparison of open-source reinforcement learning environments for energy and grid management. Unlike existing simulators that focus on physical device control or network topology, DR-GYM provides a market-level, risk-aware environment designed to mitigate consumer financial tail-risk. We do not include MARL-iDR (van Tilburg et al., 2023) due to its focused application setting rather than being a testbed.

Feature	**DR-GYM (Ours)**	**CityLearn**	**Sinergym**	**Grid2Op**
Target Agent Role	**Market-Level electric utility** (DR credit pricing)	Building Controller (HVAC/Battery dispatch)	Building Controller (Thermostat setpoints)	System Operator (Transmission topology)
Risk Objective	**Risk-aware and neutral** CVaR as example	Risk-neutral (Expected cost/energy)	Risk-neutral (Energy vs. comfort)	Risk-neutral (Margin maximization)
Extreme Events	**Markov regime-switching** (Correlated price spikes)	Historical static weather & pricing	Historical & simulated TMY weather data	Static historical time-series scenarios
Human Behavior	**Stochastic fatigue** (Heterogeneous archetypes)	N/A (Compliant devices)	N/A (Compliant devices)	N/A (Inelastic load profiles)
Demand Coupling	Temperature-coupled synthetic & ResStock	Physics-based thermal dynamics	EnergyPlus co-simulation	PandaPower / AC-DC power flows