Scalable Framework for Delivering Demand Flexibility from Buildings Using Machine Learning

Matthew Brugman, Edo Energy
Easan Drury, Edo Energy

ABSTRACT

Electric utilities are experiencing increased congestion on their distribution grids from multiple sources, including population growth, load growth from electric vehicle adoption and data centers, an increasing number of extreme weather events, and aging infrastructure. Energy use in buildings can be modulated using advanced controls to provide cost-effective demand flexibility (DF) that is predictable, dispatchable, and scalable. Several past studies have demonstrated the ability of calibrated building-level simulations to predict baseline energy use, quantify the impact of control measures, and implement advanced controls using strategies like model predictive control (MPC). However, developing and calibrating sophisticated building-level simulations is expensive and requires a specialized labor force, making it costly and challenging to deploy at the scale of a utility service territory. In this paper, we outline a data-driven machine learning (ML) based methodology that learns the characteristics of energy use from near real-time energy and HVAC time-series data and learns the impacts of various control measures that are implemented by an onsite gateway. This framework is inherently cost-effective and scalable, utilizing cloud-based computational resources for ML forecasts and control strategies and an onsite gateway to implement and monitor control measures. We compare predicted versus actual DF from multiple building types (schools, office buildings, and fitness centers) and geographies and evaluate the accuracy of ML forecasts relative to common baseline methodologies like X-in-10. Lastly, we describe extensions to the framework for co-optimizing behind-the-meter distributed energy resources (DERs).

Introduction

Electric utilities are experiencing increased congestion from the combined impact of several factors, including population growth, increased electricity consumption from the buildout of data centers and the rapid adoption of electric vehicles (EV), increased frequency and intensity of extreme weather events creating new peaks in electricity demand, and aging utility infrastructures particularly on the distribution grid. To mitigate congestion, utilities are interested in demonstrating additional sources of load flexibility and operational dispatch strategies, in addition to upgrading capacity on the transmission and distribution grid.

At the same time, electric utility customers are seeing significant increases in their electricity rates, with rate increases outpacing inflation in many states. Cost drivers vary by geographic location and include maintaining an aging transmission and distribution (T&D) infrastructure, increased prevalence of wildfires and extreme storms, the rapid buildout of new generation and T&D capacity to support growing data center load, and other factors like natural gas price volatility (Wiser et al. 2025). Customers are also seeing an increase in rate options, including time-of-use (TOU) rates and rates tied to onsite DERs like solar photovoltaics and/or EV chargers.

Electricity use in commercial buildings accounts for approximately 35% of total electricity demand in the U.S., dominated by HVAC load (EIA 2026). Advanced HVAC controls can reduce energy use by up to 30% and shift the timing of energy use to reduce building peak load by 10–20% (Fernandez et al. 2017; Fernandez et al. 2018; Han et al. 2023). For utilities, this provides flexible grid capacity through load shifting and shedding. For building owners, it unlocks value through demand response revenues, demand charge reduction, time-of-use optimization, and carbon emissions reduction.

However, implementing advanced controls in buildings has historically required a significant investment of time and money. Past implementation of advanced control has often focused on developing detailed physics-based models with complex representations of a building's physical characteristics, HVAC systems, and control logic. The main challenges for enabling advanced controls to become an economically viable solution are reducing the time and effort to create predictive models and developing cost-effective and scalable methods for implementing supervisory control.

This paper presents a data-driven machine learning framework for delivering demand flexibility from commercial buildings. We describe a cloud-to-gateway architecture that replaces labor-intensive, physics-based modeling with automated ML pipelines, evaluate forecasting accuracy across multiple model types and forecast horizons at a single school site, and discuss how improved forecasts enable building-owner economic optimization including demand charge management and time-of-use arbitrage. We also describe how the same forecasting infrastructure that supports building-owner economics can simultaneously provide utilities with reliable, dispatchable demand flexibility, thereby aligning the interests of both parties through a shared technical foundation.

This work makes three contributions relative to the existing literature on ML-based building load forecasting. First, we demonstrate the critical role of building automation system (BAS) setpoint schedules as future covariates, which reduced day-ahead mean absolute error (MAE) by 54%; a finding not reported in prior studies that rely primarily on weather and calendar features (e.g., Wang et al. 2018; Fan et al. 2019). Second, we systematically compare training strategies (shared vs. per-horizon output windows, daily vs. per-chunk retraining) to quantify the accuracy and computation tradeoff across nine forecast horizons. Third, we connect forecast accuracy directly to economic outcomes through a Mixed Integer Linear Programming (MILP)-based dispatch optimization evaluated under two real tariff structures.

The Limitations of Traditional Building Modeling

To date, the building energy industry has largely relied on physics-based simulation tools such as EnergyPlus and DOE-2 to predict building behavior and optimize energy use. While these tools can be highly accurate when properly calibrated, they encounter significant friction when extended from isolated research projects to the scale of a utility service territory. This friction is primarily driven by three factors: the high setup effort for building simulation, the inherent constraints of model predictive control, and a persistent economic viability gap.

Simulation Setup Effort

The primary barrier to scaling simulation-supported demand flexibility is the labor-intensive nature of the modeling process. Creating a high-fidelity digital twin of a commercial building requires granular data entry covering the thermal envelope, lighting schedules, and detailed HVAC mechanical configurations. Even with recent advances in automated geometry extraction and equipment detection, the calibration phase remains a bottleneck. Typical calibration timelines range from 40 to 200 engineering hours per building depending on complexity (Coakley et al. 2014). The time required to build, calibrate, and maintain these models for hundreds or thousands of unique sites across a utility territory creates a deployment timeline that cannot keep pace with the needs of the grid.

The MPC Constraint

Advanced building control strategies range from enhanced rule-based approaches such as ASHRAE Guideline 36 high-performance sequences to model predictive control, which is often cited as the gold standard because it can account for weather forecasts, occupancy patterns, and thermal mass in a unified optimization framework (Drgona et al. 2020). However, its efficacy at scale is stymied by site-specific setup requirements (Cigler et al. 2013; Rockett and Hathaway 2017). In a laboratory setting or flagship smart building, MPC can perform exceptionally well. At commercial scale, the complexity of integrating MPC with legacy building automation systems introduces significant risk because each site requires custom mapping of control points and bespoke tuning of the predictive model. This bespoke requirement means that a successful pilot program rarely translates into a broad, cost-effective rollout. The control logic that optimizes a high school cannot be trivially transferred to a fitness center without another round of engineering effort.

The Economic Gap

Ultimately, the viability of any demand flexibility framework is governed by its return on investment for the building owner. Physics-based modeling and MPC methods frequently fall into an economic gap where the specialized labor costs for setup and maintenance exceed the value that the building owner can capture through reduced demand charges, time-of-use arbitrage, or utility incentive payments. Without subsidies to offset these soft costs, traditional approaches remain non-viable for the mass market. To achieve territory-wide scale, the industry requires a transition from labor-intensive simulation to data-driven, automated machine learning systems that can learn a building’s thermal and mechanical behaviors with minimal manual intervention.

Proposed Framework: Scalable ML and Cloud Architecture

To overcome the labor-to-savings hurdles of physics-based modeling, we developed a decoupled, cloud-native architecture. This framework shifts the intelligence of the system from site-specific engineering to centralized machine learning pipelines while maintaining a lightweight but robust physical presence at the building edge. The architecture is currently deployed across dozens of commercial buildings nationwide, spanning schools, office buildings, and fitness complexes. In this paper, we present the detailed forecasting methodology and results for a single site; the architectural description applies to the broader fleet.

Data Ingestion and System Architecture

The framework is built on a high-resolution data loop that captures building telemetry at 15-minute intervals, as shown in Figure 1. This frequency is sufficient to capture the thermal dynamics and cycling patterns of HVAC equipment without the storage and bandwidth overhead of second-by-second monitoring. Unlike white-box simulations that require manual entry of mechanical specifications and physical characteristics, this architecture treats the building as a gray-box entity, learning its thermal behaviors through historical and real-time data streams. The primary data sources are interval electric meter data (kilowatt-hours (kWh) at 15-minute resolution), zone-level HVAC setpoints from the BAS, and external weather observations including dry-bulb temperature and solar irradiance.

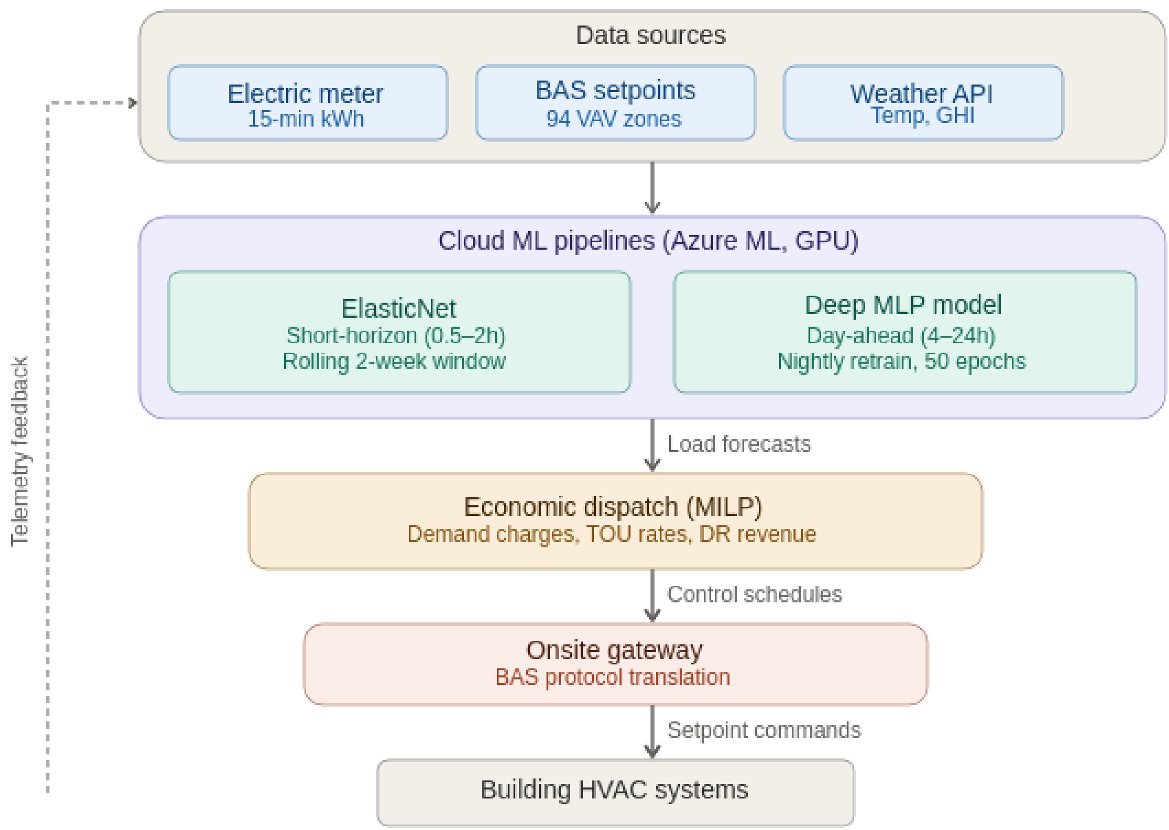

Figure 1. Simplified Architecture for Model Retraining and Data Collection

Cloud-Based ML Pipelines

The scalability of the framework is rooted in its centralized cloud-based ML pipelines. For this study we employ a hybrid approach combining a regularized linear model (ElasticNet) for short-horizon forecasts with a deep multilayer perceptron (MLP) forecast model for day-ahead forecasts. Both models are trained in the cloud on dedicated graphics processing unit (GPU) compute (NVIDIA V100 instances on Azure ML), enabling automated nightly retraining across the building fleet without requiring onsite engineering resources.

The MLP-based design allows for substantially faster training and inference compared to attention-based transformer architectures, making it practical to manage individual models for hundreds of building profiles simultaneously. The pipeline automatically ingests auxiliary features—weather forecasts, HVAC setpoint schedules, and temporal encodings—and mixes them with historical load data to produce building-specific demand forecasts. As building occupancy patterns shift or equipment characteristics change, the models are automatically retrained, ensuring accuracy without manual intervention.

Onsite Gateway

While the forecasting intelligence resides in the cloud, the onsite gateway that serves as the execution arm is a proprietary device that interfaces with standard BAS protocols to translate cloud-level optimization signals into local control commands such as HVAC setpoint adjustments. Control schedules covering multiple hours are sent with frequent updates, providing resilience against short-duration network outages.

Methodology: Forecasting and Baseline Comparison[1]

This section describes the forecasting methodology and evaluation framework used to compare model accuracy across forecast horizons. The objective is to determine which model architecture best supports two operational modes: real-time intraday corrections (horizons of 30 minutes to 2 hours) and day-ahead event planning (horizons of 8 to 24 hours). Accurate forecasts at both timescales are prerequisites for building-level economic optimization, whether for demand charge management, time-of-use arbitrage, or utility event participation.

Study Site and Data

The study building is an elementary school in the Federal Way Public Schools district, near Seattle, Washington. The school uses electric HVAC systems (heat pumps with variable air volume distribution) as its primary source of load flexibility. The building participated in 12 utility demand flexibility events during the evaluation period, including winter peaking events, summer peaking events, and emergency test events.

Interval electric meter data at 15-minute resolution was collected from the building automation system from December 2024 through March 2026. The raw data uses interval-end timestamps; these were shifted back by one interval to align with interval-start convention, so that each data point represents the energy consumed during the interval beginning at the labeled time. Weather data (dry-bulb temperature and global horizontal irradiance) was sourced from the

[1]Elements of the methodology described in this section are the subject of a provisional patent application filed by Edo Energy. Confidential - Patent pending.

Visual Crossing historical weather API (Visual Crossing 2025) for Federal Way, Washington, and aligned to the same 15-minute grid.

A key covariate in this study is the HVAC setpoint schedule. The building has 94 variable air volume (VAV) zones, each with independent heating and cooling setpoints. For modeling purposes, zone-level setpoints were averaged across all zones to produce two aggregate signals: mean effective heating setpoint and mean effective cooling setpoint. These setpoints encode when the building transitions from unoccupied setback temperatures to occupied comfort targets, which is the dominant driver of the morning load ramp and evening ramp-down. This aggregation assumes that zones are scheduled on a common occupancy program, which holds for this school site. Buildings with heterogeneous zone schedules (e.g., mixed-use facilities) may benefit from zone-cluster or zone-level setpoint features, which we leave to future work.

An important methodological note: for weather covariates, this study uses historical observations rather than weather forecasts. In an operational deployment, forecast weather would introduce additional uncertainty. However, for the purpose of isolating forecast model accuracy from weather forecast error, we treat weather as a known future covariate. HVAC setpoints are genuinely known in advance and therefore do not introduce information leakage.

Evaluation Framework

The evaluation period spans January 13, 2025, through December 19, 2025, a subset of the available meter data. This period was selected to represent a full school year at the site while avoiding earlier periods when onsite HVAC improvements were taking place. All metrics are computed over school days only (weekdays excluding district holidays, professional development days, and school breaks), yielding 178 evaluation days. This restriction is deliberate: weekends and breaks exhibit near-zero, highly predictable load profiles that would inflate apparent accuracy and obscure performance during the operationally relevant periods.

Within each school day, error metrics are further restricted to occupied hours: 06:00 to 18:00, corresponding to 15-minute time steps 25 through 73 in the day. Demand flexibility events from the utility provider occur exclusively during occupied hours, and it is during this window that forecast accuracy has economic consequences. The primary error metric is mean absolute error (MAE) in kilowatts, supplemented by root mean squared error (RMSE) and mean absolute percentage error (MAPE), consistent with ASHRAE Guideline 14 (ASHRAE 2014). Models are evaluated at nine forecast horizons: 0.5, 1, 2, 3, 4, 6, 8, 12, and 24 hours.

Industry Baseline: x-in-y Analogous Day Method

The x-in-y method is the industry-standard approach for estimating demand response baselines (Coughlin et al. 2009; NAESB 2009). To forecast a given day's load profile, the method identifies the x most similar or highest load days from the previous y eligible days and averages their load profiles. We evaluated five configurations: 4-in-5, 3-in-5, 5-in-10, 4-in-10, and 10-in-20. The 4-in-5 variant, which averages the four best-matching weekdays from the last five (excluding flexibility event days), is the utility-specified method for this site and performed best in our evaluation.

On event days, the x-in-y forecast can be improved with a day-of adjustment (DOA): the mean difference between actual and predicted load during a pre-event window is applied as a uniform additive correction. We evaluated applying DOA on all school days (using a default 08:00 pre-period on non-event days) versus only on event days using each event's actual start time. The all-days mode unexpectedly increased error because the 06:00 to 07:00 pre-period coincides with the high-variance HVAC morning ramp. We adopted the events-only mode for all comparisons reported below.

Linear Model: ElasticNet

The first ML model is ElasticNet (Zou and Hastie 2005), a regularized linear regression combining L1 and L2 penalties ($\alpha = 0.1$, L1 ratio = 0.5), implemented via the Darts time-series library (Herzen et al. 2022). The model receives 96 lags of the target variable (24 hours of historical load), future covariates at lag zero (temperature, global horizontal irradiance, heating setpoint, cooling setpoint), and cyclic encodings of hour of day and day of week. A separate model is trained for each output time step within the forecast window, allowing the model to capture the different dynamics at each step in the prediction horizon.

The model is trained on a rolling two-week window: for each forecast chunk, a fresh ElasticNet is fit on the 1,344 timesteps (14 days at 96 steps per day) immediately preceding the forecast origin. This rolling window was chosen over an expanding window for three reasons. First, building load patterns drift seasonally, so data from six months prior is more likely to confuse than inform a short-term forecast. Second, a fixed window enables fair comparison with the deep MLP model, which also uses two weeks. Third, two weeks contain exactly ten school days, sufficient to capture the weekly schedule cycle.

Deep Learning Model: Deep MLP Forecast Model

The second ML model is a deep MLP forecast model that uses mixer blocks to exchange information across time steps and feature channels without the computational overhead of self-attention mechanisms. Drawing on the MLP-Mixer family of architectures, which has demonstrated strong performance in both computer vision (Tolstikhin et al. 2021) and time-series forecasting (Chen et al. 2023), the model alternates between time-mixing and feature-mixing MLP layers. The model receives 96 input time steps and produces forecasts via future covariates identical to those used by ElasticNet: temperature, global horizontal irradiance, heating and cooling setpoints, and cyclic temporal encodings. Each model is trained for 50 epochs on a rolling two-week window using GPU-accelerated compute.

We evaluate three deep MLP training strategies to understand the tradeoffs between computational cost and forecast accuracy:

Variant A: Daily retrain, shared output window. A single model is retrained each school day with an output chunk length of 96 steps (24 hours). All forecast horizons are obtained by truncating the same 96-step prediction. This is the most computationally efficient variant, requiring 178 model fits across the evaluation period regardless of how many horizons are evaluated.

Variant B: Daily retrain, per-horizon output window. A separate model is trained for each forecast horizon, with the output chunk length set to the horizon length. This requires $9 \times 178 =$

1,602 model fits. The hypothesis is that focusing each model's training loss on its specific horizon will improve short-horizon accuracy relative to Variant A.

Variant C: Per-chunk retrain, per-horizon output window. A fresh model is retrained immediately before each forecast window on the two-week window ending at that chunk's start time. This is the most computationally expensive variant but allows the model to incorporate the most recent observations of the building's thermal state. Results are available for horizons of 4, 6, 8, 12, and 24 hours; shorter horizons are computationally burdensome to calculate for the entire year.

Case Study: Demand Flexibility at a Public Elementary School

The building's load profile exhibits strong seasonality and diurnal patterns. Winter peak loads are driven by electric heat pump operation during morning warm-up, while summer peaks are driven by cooling during afternoon hours. The transition between unoccupied setback temperatures and occupied comfort targets produces a sharp morning ramp that is the single largest source of forecast difficulty, as its magnitude depends on overnight outdoor temperature, building thermal mass, and the exact timing of the setpoint transition. Six peaking events occurred across January and February 2026, with another four occurring across July and August 2026.

Results and Discussion

Industry Baseline Performance

Table 1 summarizes the average occupied-hours MAE for each x-in-y configuration with events-only day-of adjustment. The 4-in-5 configuration achieved the lowest MAE at 9.82 kW. Performance degraded with wider lookback windows: the 10-in-20 configuration, which draws from a broader pool of candidate days, reached 12.99 kW. This is expected, as wider windows increase the likelihood of including days with dissimilar weather or occupancy patterns.

Table 1. Average annual occupied-hours MAE for x-in-y baseline configurations

Configuration	Average Daily MAE (kW)
4-in-5, with DOA on events	9.82
3-in-5, with DOA on events	9.96
5-in-10, with DOA on events	11.06
4-in-10, with DOA on events	11.75
10-in-20, with DOA on events	12.99

ML Model Comparison Across Forecast Horizons

Table 2 presents the occupied-hours MAE for each model at each forecast horizon. Three findings stand out.

ElasticNet dominates at short horizons. At H = 0.5 hours, ElasticNet achieves 6.60 kW MAE, substantially lower than Deep MLP Variant A's approximately 9 kW. This advantage persists through H = 2 hours. The explanation is straightforward. At short horizons, the dominant predictive signal is thermal inertia—the building's load in the next 30 minutes is well-approximated by its recent load history. ElasticNet's 96-lag feature set captures this directly and efficiently. Deep MLP Variant A, by contrast, distributes its training loss across all 96 output steps, so it does not specialize for the first few timesteps.

Deep MLP outperforms at long horizons. At H = 24 hours, Deep MLP Variant A achieves 8.78 kW MAE versus ElasticNet's 16.21 kW, a 46% reduction. The deep MLP model also outperforms the 4-in-5 industry baseline (9.82 kW) at all horizons from 4 hours onward. The advantage of the deep MLP model at long horizons is driven by two factors: its ability to learn nonlinear interactions between HVAC setpoints and load (the setpoint transition does not produce a linear load response), and its capacity to mix dependencies across all input time steps jointly. ElasticNet treats each lag independently and cannot represent the complex interactions between weather, setpoints, and building thermal dynamics that unfold over a 24-hour forecast window.

HVAC setpoints are critical covariates. Without setpoint covariates, the deep MLP model's H = 24-hour MAE was approximately 19 kW, worse than both ElasticNet and the 4-in-5 baseline. With setpoints included, it dropped to 8.78 kW, a 54% improvement. Weather alone is insufficient because it does not tell the model *when* the HVAC system will activate; the setpoint schedule carries this information explicitly. This finding has practical implications. Buildings participating in demand flexibility programs should expose their BAS setpoint schedules to the forecasting pipeline.

Table 2. Occupied-hours MAE (kW) by model and forecast horizon

Model	0.5h	1h	2h	3h	4h	6h	8h	12h	24h
4-in-5 + DOA	-	-	-	-	-	-	-	-	9.82
ElasticNet	6.60	7.60	8.81	9.60	10.21	11.13	11.63	14.03	16.21
Deep MLP A (daily, shared)	9.37	9.21	9.08	8.99	8.94	8.90	8.90	8.61	8.61
Deep MLP B (daily, per-horizon)	7.81	7.27	7.80	7.12	7.69	8.18	8.07	8.01	8.61
Deep MLP C (per-chunk)	-	-	-	-	7.46	8.06	7.74	7.76	8.78

Figure 2 visualizes the MAE trends from Table 3 across all forecast horizons, clearly illustrating the crossover point near H = 2 to 3 hours where the deep MLP model overtakes ElasticNet. Variant B (per-horizon output window) achieves the strongest performance across most horizons, confirming the hypothesis that focusing each model's training loss to its target horizon improves accuracy. Figure 3 shows actual versus predicted load profiles for a representative winter peaking event day (January 22, 2025), illustrating how the models capture (or miss) the morning HVAC ramp-up and the load reduction during the morning event window (08:00 to 10:00).

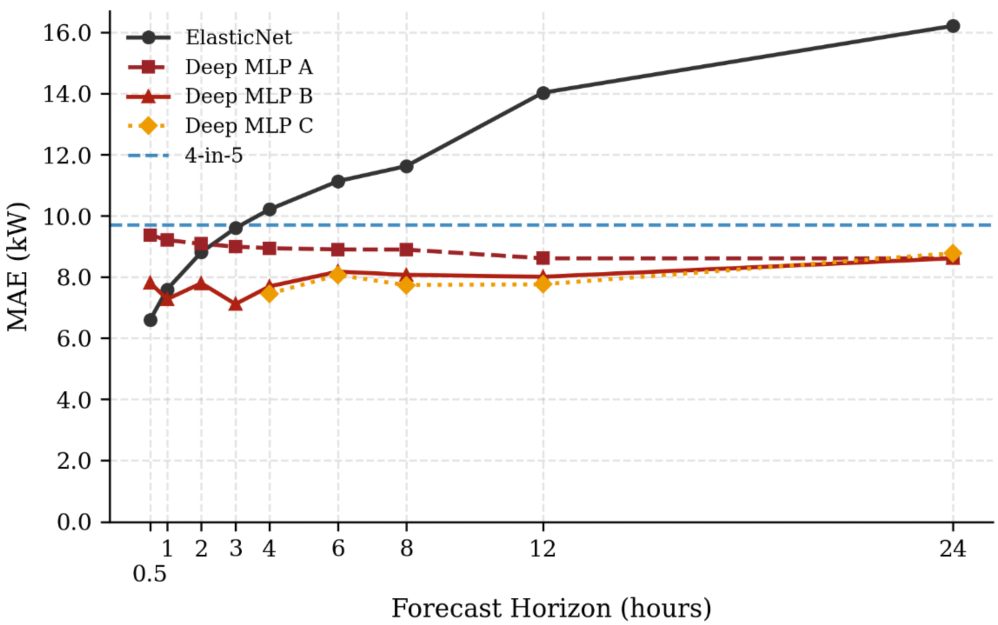

Figure 2. Average Daily MAE by Model and Forecast Horizon

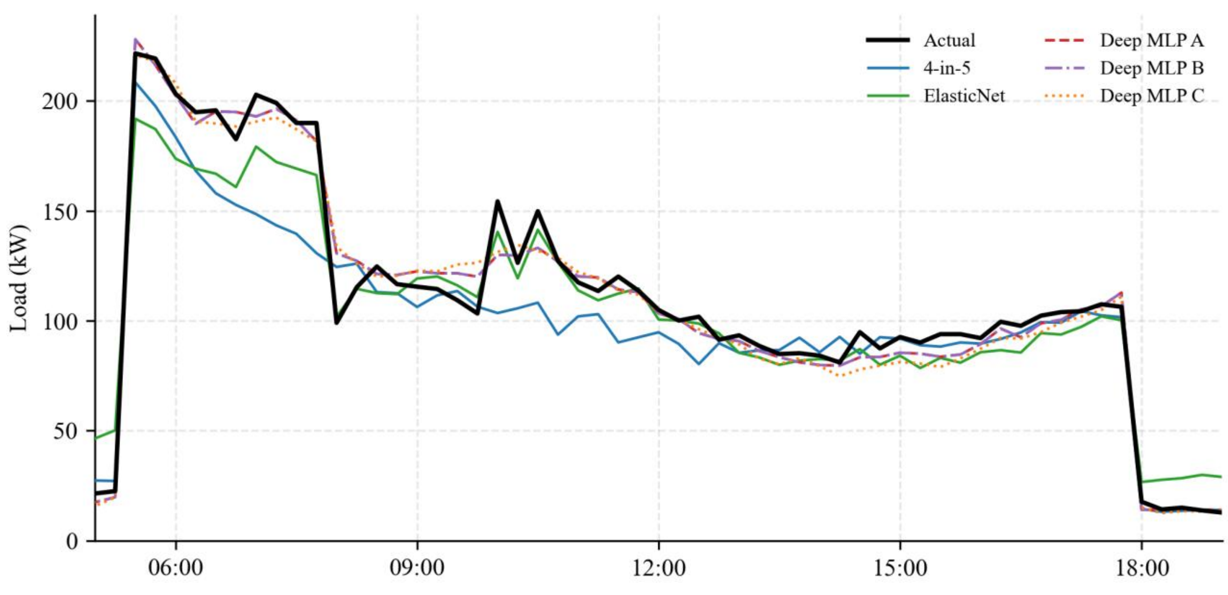

Figure 3. Actual Load versus Forecasts for January 22, 2025

The Value of Retraining Frequency

Deep MLP Variant C (per-chunk retraining) shows meaningful improvement over Variant A at H = 8 hours (7.74 vs 9.13 kW) and H = 12 hours (7.76 vs 8.82 kW). At H = 24 hours, the two variants are equivalent (8.78 kW), which is expected. With a 24-hour forecast window, there is only one forecast per day, making per-chunk and daily retraining identical. The improvement at medium horizons confirms that the most recent building observations carry information about the building's current thermal state that a model trained at midnight does not capture.

However, per-chunk retraining has practical limitations. At short horizons (H ≤ 2 hours), retraining a neural network every 30 minutes to 2 hours is computationally prohibitive. The daily retrain approach (Variant A) with its shared 24-hour output window is a practical compromise that captures most of the forecasting gain. A recommended operational strategy is to use ElasticNet for real-time intraday corrections at horizons below 2 hours, and a nightly-retrained deep MLP model for day-ahead planning at horizons of 4 hours and beyond.

Implications for Building-Owner Economics

The forecasting accuracy demonstrated here has direct implications for building-owner economic optimization. Demand charges, which are assessed based on the building's peak load during a billing period, can represent 30–50% of a commercial electricity bill. A day-ahead forecast with sub-10 kW MAE enables an optimization engine to pre-position HVAC setpoints to shave the anticipated peak with confidence that the predicted load shape is reliable. Similarly, under time-of-use rate structures, accurate hourly forecasts allow the optimizer to shift load from high-price to low-price periods by pre-conditioning the building's thermal mass.

Compared to the industry-standard 4-in-5 baseline (9.82 kW MAE), the deep MLP model's day-ahead accuracy of 8.78 kW represents a 10.6% improvement. More importantly, the ML approach provides a continuous, horizon-specific forecast rather than a single whole-day profile, enabling finer-grained optimization that the x-in-y method cannot support. The combination of day-ahead deep MLP forecasts with real-time ElasticNet corrections provides the optimization engine with continuously updated load predictions at multiple timescales.

Critically, the same forecasting infrastructure that enables the building-owner savings also supports the utility's need for dispatchable demand flexibility. Utilities require confidence that a dispatched curtailment event will deliver the promised load reduction; this confidence depends directly on the accuracy of forecasting. An inaccurate forecast creates settlement risk. If the forecast overstates what the building would have consumed, the apparent load reduction is inflated, and the utility pays for flexibility it did not receive. Conversely, an understated baseline forecast discourages building owners from participating because their actual curtailment effort is undervalued. The sub-10 kW day-ahead MAE demonstrated here narrows this uncertainty band, enabling utilities to treat building-level demand flexibility as a more reliable capacity resource in their distribution planning.

This creates a virtuous alignment between utility and building-owner interests. The building owner benefits from accurate forecasts through reduced demand charges and optimized time-of-use shifting. These value streams accrue regardless of whether a utility event is called. The utility benefits from the same forecasts through improved forecast accuracy during dispatched events, which in turn supports more aggressive procurement of building-level flexibility. Neither party requires a separate forecasting system; the machine learning approach described in this paper may serve both objectives simultaneously. The combination of these building-owner value streams with utility event participation revenues creates a stacked economic case that can justify the cost of the gateway hardware and cloud ML infrastructure at individual sites.

Economic Dispatch and Multi-Objective Optimization

Advanced control frameworks solve the problem of selecting the optimal control strategy from many control options in response to multiple value streams available to customers. For buildings, control measures typically include several equipment types and zones in the HVAC and lighting systems, as well as behind-the-meter DERs like EV chargers, battery energy storage systems (BESS), and thermal energy storage (TES) systems. The value streams often include utility demand response (DR) or demand flexibility programs, tariff optimization which includes reducing peak demand charges and/or shifting load from high to low-cost TOU periods and responding to other signals like marginal carbon emission rates or wholesale electricity prices.

Mixed Integer Linear Programming is an effective tool for identifying optimal dispatch strategies for many control options subject to multiple value streams. Within a MILP, the characteristics of the building systems and control measures are represented by a series of constraints, such as the depth of load flexibility from building HVAC control representing shifting and shedding load while remaining within configurable zone temperature deadbands, ventilation requirements, lighting levels, and/or battery state of charge (SOC) and max charge/discharge rates. The multiple value streams from DR/DF programs, tariff optimization, and dynamic energy prices are parameterized in the cost function. Equations 1-4 below describe the cost function for one customer billing period:

$$Cost = fixed\ charge + \sum_i energy_i * energy\ rate_i + \sum_j max(energy)_j * demand\ charge_j - DR_{revenue} \quad \textit{[Eq 1]}$$
$$DR_{revenue} = DR_{revenue,kW} + DR_{revenue,kWh} \quad \textit{[Eq 2]}$$
$$DR_{revenue,kW} = DR_{kW,k} * mean(energy_k) \quad \textit{[Eq 3]}$$
$$DR_{revenue,kWh} = DR_{kWh,k} * \sum_k energy_k \quad \textit{[Eq 4]}$$

Equation 1 describes a customer's energy cost per billing period, which is frequently a combination of a fixed monthly charge, the sum of energy used for hours i times the energy rate ($/kWh) for hours i, and the peak customer demand during the hours j times the demand charge ($/kW) for hours j, minus the revenue from participating in utility DR/DF programs. Equations 2-4 describe common revenue streams for DR/DF programs that include a capacity-based payment ($/kW) scaled by the average amount of customer load reduced (kW) during the k hours of the DR/DF event (*Eq 3*), and/or an energy-based payment ($/kWh) for reduction in customer load (kWh) during the k hours of the event (*Eq 4*). As described above, utilities frequently use an x-in-y methodology to define customer baselines to quantify the amount of customer demand or energy that is shifted or shed during the k event hours.

The MILP is solved for one customer billing period at a time using the open-source HiGHs solver (Huangfu and Hall 2018). While open-source solvers are not as performant as commercial solvers, they can horizontally scale in a cloud-based environment to determine optimal control measures for hundreds to thousands of buildings and DERs.

Economic Dispatch Model Results

To understand the impact of forecast error on optimal control strategies, we ran the MILP economic dispatch model under several scenarios: perfect foresight (in which future load is known exactly), and forecasted load from the algorithms described above. For each forecast scenario, the forecast was used as the building load input to the economic dispatch model, which determined optimal setpoint adjustments to minimize electricity cost. The difference between the forecast and the optimized net load relative to the forecast defined the shifted and shed load for each interval. This shifted and shed load was then applied to the measured load for every day of the billing period, and the resulting net load was used to calculate electricity costs. It is important to note that the economic dispatch results presented here are simulated. The MILP determines optimal control schedules retrospectively using measured load data and forecasts, but these schedules were not deployed as real-time control actions during the study period. We evaluated two tariff structures representative of common U.S. commercial rate designs. The first represents a commercial rate with a flat seasonal energy and demand charge that is applicable for all hours of the day, modeled after Puget Sound Energy's Schedule 025 for medium sized commercial customers. The second represents a commercial rate with TOU energy and demand charges, modeled after Southern California Edison's TOU-GS-3-E tariff. TOU rate structures have become increasingly common over the past decade as utilities deploy Advanced Metering Infrastructure (AMI) that monitors customer load at 5- to 15-minute intervals.

Figure 4 shows electric bill savings from simulated advanced controls under both tariffs. For the TOU tariff, the perfect foresight scenario shows a 15.6% electric bill savings, with 12.0% from reducing demand charges and 3.6% from reducing energy charges. For the flat rate tariff, the perfect foresight scenario shows a 17.4% electric bill savings, with 12.8% from reducing demand charges and 4.6% from reducing energy charges. The lower overall savings for the TOU tariff is driven by a few factors: first, the study period focused on non-holiday periods for a school which removed summer months with high daytime TOU rates; second, the school load profile frequently peaks during the early morning warmup period; and third, the early morning TOU rate for non-summer months is more expensive than later morning rates. These factors combine into a tradeoff for the economic dispatch model where it can choose to precondition more with higher-cost electricity before the peak, to offset more of the morning warmup peak to minimize demand charges. For the flat rate, there is no tradeoff since electricity costs are the same for the preconditioning period as the event and rebound periods. Energy savings for both scenarios were primarily driven by allowing the economic dispatch model to widen temperature deadbands to shift and shed load for flex events.

Figure 4. Savings percentage from the economic dispatch model for an electricity tariff with TOU rates (top), an electricity tariff with flat rates (middle), and the combination of savings from both rate structures representing a portfolio of buildings with customers on both tariffs (bottom). *Perfect foresight* represents results from the economic dispatch model with full information about future building load, and all other bars represent the results of the economic dispatch models optimized to different forecasts.

Figure 4 shows that dispatching to the forecasted load led to similar energy savings across forecast algorithms, but significant differences in demand charge reduction. For the TOU rate, we see electric bill savings ranging from 7.3% to -1.1%. Only the *xiny_4in5* and *mlpA_6hr* scenarios led to decreased demand charges, while nine of the forecast algorithms led to increased demand charges. However, all but one forecast scenario created positive savings based on the tradeoff between charging at higher energy costs to reduce demand charges described above. For the flat rate scenario, savings ranged from 8.7% to 4.6%, and all but one forecast reduced demand charges.

The best performing forecast for the economic dispatch model was the *xiny_4in5* algorithm. This was surprising given that it did not have the lowest error of the forecasts explored in this study (Figure 2). We find that the error in the *xiny_4in5* forecast was driven by frequent positive bias, which forced the economic dispatch model to time events more frequently during the peak morning warmup period, which led to it creating more demand charge reduction than other forecasts with lower overall error and bias. This suggests areas for improving future approaches by developing forecasts with strategically high bias or developing custom error metrics for forecasts that asymmetrically penalize underestimates more than overestimates of building load profiles.

Conclusions

This paper presented a scalable, cloud-to-gateway framework for delivering demand flexibility from commercial buildings using machine learning. We evaluated three forecasting approaches at a public elementary school and tested the impact of forecast accuracy on economic dispatch under two representative tariff structures.

A regularized linear model (ElasticNet) outperforms the industry-standard 4-in-5 baseline at short horizons, while a deep MLP model with HVAC setpoint covariates reduces day-ahead MAE by 46% relative to ElasticNet and outperforms the industry baseline at all horizons beyond 4 hours. HVAC setpoints proved critical because the deep MLP model performed worse than all alternatives at 24 hours without them. The economic dispatch analysis showed that forecast accuracy has the greatest impact under tariffs with significant demand charges and time-of-use differentiation, where the best forecast model captured 11% savings versus 17.4% under perfect foresight. Under flat-rate tariffs, all models converged near the optimization tolerance.

These results suggest a practical strategy: x-in-y or ElasticNet for intraday corrections at horizons of 2 hours or less, and a nightly-retrained deep MLP model for day-ahead planning. The same forecasting infrastructure that drives building-owner savings simultaneously supports utility needs for dispatchable demand flexibility, aligning both parties through a shared technical foundation.

This study has several limitations. The forecasting evaluation is conducted at a single site with a relatively consistent occupancy schedule; performance may differ for buildings with more variable occupancy patterns or mixed-use configurations. Weather covariates use historical observations rather than forecasts, which may overstate operational accuracy. The economic dispatch results are simulated rather than deployed. Future work will extend the evaluation to additional building types in the deployed fleet, incorporate weather forecast uncertainty, validate closed-loop dispatch performance, and explore global model architectures that can transfer learned representations across buildings to reduce cold-start training requirements.

References

ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers). 2014. *ASHRAE Guideline 14-2014: Measurement of Energy, Demand, and Water Savings.* Atlanta: ASHRAE.

Chen, S., C. Li, N. Yoder, S. Arik, and T. Pfister. 2023. "TSMIXER: An All-MLP Architecture for Time Series Forecasting." *Transactions on Machine Learning Research* (September). https://arxiv.org/abs/2303.06053.

Cigler, J., J. Siroky, M. Korda, and C. Jones. 2013. "On the Selection of the Most Appropriate MPC Problem Formulation for Buildings." In *Proceedings of the 11th REHVA World Congress CLIMA.* Prague.

Coakley, D., P. Raftery, and M. Keane. 2014. "A Review of Methods to Match Building Energy Simulation Models to Measured Data." *Renewable and Sustainable Energy Reviews* 37: 123–141.

Coughlin, K., M. Piette, C. Goldman, and S. Kiliccote. 2009. *Statistical Analysis of Baseline Load Models for Non-Residential Buildings.* Berkeley, CA: Lawrence Berkeley National Laboratory. LBNL-2074E. eta-publications.lbl.gov/sites/default/files/lbnl-4984e.pdf.

Drgona, J., J. Arroyo, I. Figueroa, D. Blum, K. Arendt, D. Kim, E. Olle, J. Oravec, M. Wetter, D. Vrabie, and L. Helsen. 2020. "All You Need to Know about Model Predictive Control for Buildings." *Annual Reviews in Control* 50: 190–232.

EIA (U.S. Energy Information Administration). 2026. "Electricity Explained: Use of Electricity in the United States." eia.gov/energyexplained/electricity/.

Fan, C., et al. 2019. "Assessment of Deep Recurrent Neural Network-Based Strategies for Short-Term Building Energy Predictions." Applied Energy 236: 700–710.

Fernandez, N., Y. Xie, S. Katipamula, M. Zhao, W. Wang, and C. Corbin. 2017. *Impacts of Commercial Building Controls on Energy Savings and Peak Load Reduction.* Richland, WA: Pacific Northwest National Laboratory. PNNL-25985.

Fernandez, N., S. Katipamula, W. Wang, and Y. Xie. 2018. "Energy Savings Potential from Improved Building Controls for the US Commercial Building Sector." *Energy Efficiency* 11 (10): 393–412.

Han, S., D. Kim, T. Barham, and K. Ramsey. 2023. "The First Field Application of a Low-Cost MPC for Grid-Interactive K-12 Schools: Lessons-Learned and Savings Assessment." *Energy & Buildings* 296: 113351.

Huangfu, Q., and J.A.J. Hall. 2018. “Parallelizing the dual revised simplex method.” *Mathematical Programming Computation, 10*(1), 119-142. DOI: 10.1007/s12532-017-0130-5.

Herzen, J., F. Lässig, S. Piazzetta, T. Neuer, L. Tafti, G. Raille, T. Van Pottelbergh, M. Pasieka, A. Skrodzki, N. Huguenin, M. Dumonal, J. Kościsz, D. Bader, F. Gusset, M. Benheddi, C. Williamson, M. Kosinski, M. Petrik, and G. Grosch. 2022. "Darts: User-Friendly Modern Machine Learning for Time Series." *Journal of Machine Learning Research* 23 (124): 1–6.

NAESB (North American Energy Standards Board). 2009. *Measurement and Verification of Wholesale Electricity Demand Response.* Houston: NAESB.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Rockett, P., and E. Hathaway. 2017. "Model-Predictive Control for Non-Domestic Buildings: A Critical Review and Prospects." *Building Research and Information* 45 (5): 556–571.

Tolstikhin, I., N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy. 2021. "MLP-Mixer: An All-MLP Architecture for Vision." *Advances in Neural Information Processing Systems* 34: 24261–24272.

Visual Crossing. 2025. "Visual Crossing Weather API." visualcrossing.com.

Wang, Z., et al. 2018. "Random Forest Based Hourly Building Energy Prediction." Energy and Buildings 171: 11–25.

Wiser, R., E. Shaughnessy, G. Barbose, P. Cappers, and W. Gorman. 2025. "Factors Influencing Recent Trends in Retail Electricity Prices in the United States." *The Electricity Journal* 38.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B* 67 (2): 301–320.