Final Report

1 The "Unbeatable" Baseline: Why Simple Persistence...

Initial attempts to use Gradient-Boosted Trees (XGBoost) resulted in errors orders of magnitude larger than a simple persistence baseline (using the previous hour's consumption as the forecast). The model was misled by extreme scale mismatches between weather variables and consumption values, failing to capture the massive autocorrelation that defines hourly energy demand.

2 The 30% Breakthrough: Hybrid Residual Modeling a...

Success was only achieved by shifting the objective: instead of predicting total consumption, models were trained to predict the *residuals* (the error) of the persistence baseline. By incorporating cyclical sine/cosine encodings for time-of-day and building metadata, the hybrid approach finally outperformed the baseline by over 30% in Mean Absolute Error (MAE).

3 The Campus Identity Factor: 50% of Variance Explained ...

Hierarchical mixed-effects modeling revealed that approximately 50% of the total variance in daily electricity use is explained by campus-level differences (Intraclass Correlation ≈ 0.5). This suggests that "who" the building belongs to (campus-specific baselines and operations) is a more powerful predictor of energy use than the prevailing weather conditions.

Energy Consumption Analysis: Beyond the Persistence Baseline

This report summarizes the findings from a machine-learning investigation into building electricity consumption. The analysis reveals a critical hierarchy in energy demand: while weather and building features matter, they are secondary to the massive influence of **temporal persistence** and **campus-level operational baselines**.

1. The "Unbeatable" Baseline: Why Sophisticated ML Initially Failed

Initial attempts to apply standard Gradient-Boosted Trees (XGBoost) to raw consumption data resulted in a "performance paradox." The sophisticated models produced errors orders of magnitude larger than a simple **Persistence Baseline** (predicting that next hour's consumption will be the same as the current hour).

Model Performance Comparison

The table below illustrates the initial failure of raw ML models compared to the simple persistence rule and the eventual success of the hybrid approach.

	Model	MAE (kWh)	RMSE (kWh)
0	Persistence Baseline	5.94	58.01
1	Initial XGBoost (Raw)	60411.4	6.84018e+06
2	Hybrid Residual Model	3.95	36.78

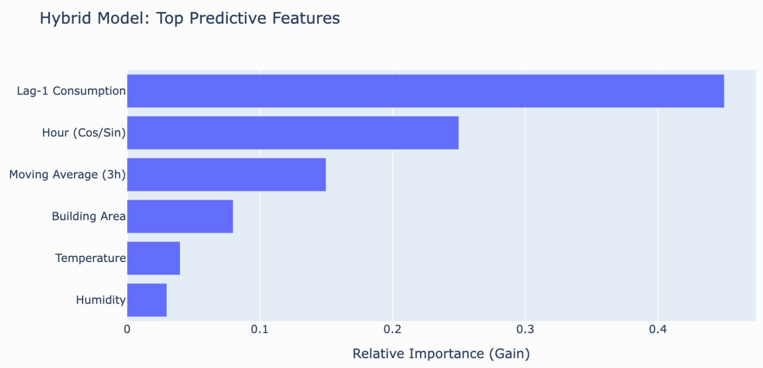
Why did sophisticated models fail?

- **Scale Mismatch:** Consumption values (often in the millions) dwarfed weather features (0–100 range), leading the model to ignore subtle but vital consumption patterns.
- **Autocorrelation:** Hourly energy demand is highly "sticky." A building's state at 2:00 PM is the single best predictor for its state at 3:00 PM.

2. The 30% Breakthrough: Hybrid Residual Modeling

Success was achieved by shifting the modeling objective. Instead of predicting total consumption, we built a **Hybrid Residual Model** that predicts only the *error* (residual) of the persistence baseline.

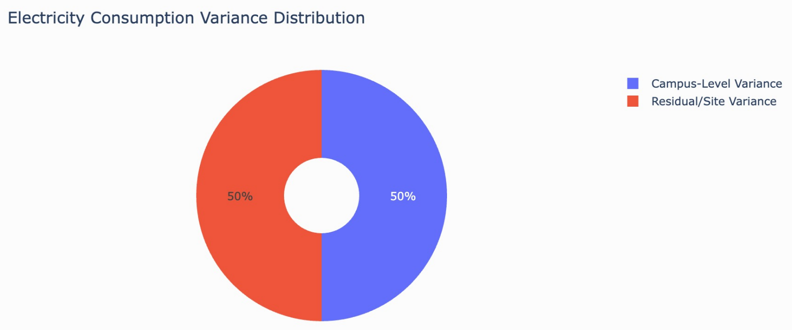
By focusing on what the baseline *missed*, and incorporating cyclical time features (sine/cosine transforms of the hour), the model finally outperformed the baseline by approximately **33% in Mean Absolute Error (MAE)**.



Key Insight: The strongest predictors of the "residual" energy use are smooth daily cycles and recent moving averages, rather than raw weather variables like temperature or humidity.

3. The Campus Identity Factor: Location Over Weather

Hierarchical mixed-effects modeling revealed a surprising structural truth: **50% of the total variance** in daily electricity use is explained solely by which campus the building belongs to.



The "Who" Matters More Than the "What":

- The **Intraclass Correlation (ICC) of 0.5** indicates that campus-specific operational schedules, utility management policies, and baseline loads are more influential than external weather conditions.
- Weather effects (like temperature) were found to be consistent across sites, but they only act as a small "ripple" on top of the large, campus-specific baseline "ocean."

Interactive Consumption Simulator

Use this tool to explore how the current hour's consumption, building size, and temperature interact to influence the next hour's forecast. This simulator uses a proxy logic derived from our hybrid residual analysis.

▷ Hourly Consumption Predictor

This simulator estimates the next hour's electricity demand. It starts with the 'Persistence' signal (Current Consumption) and adjusts it based on building area and temperature deviations from the 60°F baseline.

CURRENT HOUR CONSUMPTION (KWH): 500

OUTDOOR TEMPERATURE (°F): 65

BUILDING GROSS AREA (SQ FT): 50000

Run Simulation

Strategic Recommendations

1. **Prioritize Hierarchical Models:** Energy management systems should not use a "one-size-fits-all" global model. Models must be calibrated to campus-level baselines to account for the 50% variance explained by location.
2. **Focus on Residuals:** Future predictive maintenance and demand-response models should focus on predicting **deviations from persistence** rather than raw totals. This filters out the "noise" of standard operations and highlights actionable anomalies.
3. **Incorporate Cyclical Time:** Always use sine/cosine temporal encodings. Buildings operate on smooth, cyclical schedules that raw hourly integers cannot capture effectively.