An AI-Driven Methodology for Building More Realistic Synthetic Electric Grid Models

Emmett Souder

Independent Work, Department of Operations Research and Financial Engineering
Princeton University

Advisor: Professor Ronnie Sircar

Spring 2026

Abstract

This project is a three-month effort to build a DC Security-Constrained Economic Dispatch (DC-SCED) model of the ERCOT grid from entirely public data—OpenStreetMap (OSM) transmission topology, ERCOT's MORA generator registry, and EIA-860 generator characteristics—using an AI coding agent (Claude Code) as the primary implementation tool. It began as a replication of the Texas A&M synthetic-grid methodology (Birchfield et al.), pivoted to real OSM topology once we found its high-voltage coverage to be essentially complete, and evolved into an iterative calibration campaign on Princeton's Adroit cluster. The final model produces zero load shedding on all three calibration days and four of six unseen validation days spanning all seasons and a wide range of load and wind conditions, and reproduces the qualitative congestion geography of real ERCOT: the WEST–NORTH price spread on high-wind days, the WESTEX export constraint emerging naturally from the OSM-derived 345 kV backbone, and Houston-import bottlenecks at summer peak. Quantitative LMP magnitudes still diverge from published Real-Time SCED prices by \$5–190/MWh depending on the day, so the contribution is transparent, reproducible *structure*—not price accuracy. Better models enable better decision making in renewables siting, transmission upgrades, coal retirements, day-ahead bidding decisions, and more. Because the actual representation of the grid is Confidential Energy Infrastructure Information (CEII), researchers use synthetic grids with topologies that do not correspond to real locations. This project builds the first publicly available, SCED-ready model of ERCOT based on real topology without violating CEII. More broadly, we conclude with lessons from working with agents and a discussion of how energy systems research might evolve.

Table 1: ERCOT model at a glance.

Quantity	**Value**
Buses	3,786
Branches	4,817
Generators	1,185
Installed nameplate capacity	159 GW
Battery storage units	289
Battery storage capacity	17.5 GW / 54.3 GWh
Topology versions developed	3 (V1, V2, V3)
Calibration / validation days	3 / 6
Adroit cluster experiments run	100+

1 Introduction

1.1 The Problem

Electric grid models used for market simulation and reliability analysis are either proprietary (the actual ISO network models, protected as Critical Energy Infrastructure Information) or synthetic (algorithmically generated test cases that approximate real grid characteristics without using real data). The most widely used synthetic models—Texas A&M's ACTIVSg series, spanning from 200 to 82,000 buses—are geographically placed on real US footprints but use algorithmically generated topologies that don't correspond to actual transmission lines. This creates a fundamental tension: researchers who want to study realistic market behavior (congestion patterns, zonal price dynamics, renewable integration challenges) must either obtain restricted data or work with synthetic networks whose congestion patterns bear no guaranteed relationship to reality.

ERCOT (the Electric Reliability Council of Texas) presents a unique opportunity. As the only major US ISO operating an isolated interconnection, it publishes more operational data than any other: 60-day-lagged SCED results with nodal prices, real-time generation by fuel type, binding constraint lists, and system load data. Meanwhile, OpenStreetMap (OSM) contains increasingly complete coverage of high-voltage transmission infrastructure—a 2025 study in *Scientific Data* found OSM coverage of European HV grids to be "high or even close to complete," and in some cases more accurate than official maps.

The question this project set out to answer: **Can you build a SCED-realistic model of ERCOT from entirely public data, and can an AI coding agent do most of the implementation?**

1.2 Why This Matters to ORFE

Professor Sircar's research group (ORFEUS) works on energy market modeling—equilibrium models of power markets, stochastic control of generation investment, and game-theoretic analysis of strategic bidding—including recent work quantifying reliability risk from renewable penetration in modern and future electricity grids [17]. All of this work requires grid models that produce realistic locational marginal prices (LMPs). If congestion patterns in a model don't match reality, the market dynamics built on top of those patterns will be wrong in ways that are hard to diagnose. A publicly available, geospatially grounded ERCOT model—even an approximate one—would be a useful tool

for the group's market modeling work and for the broader research community.

The longer-term goal is to extend the methodology to NYISO (New York ISO), which has more complex market structure and is closer to Princeton geographically, but worse public data availability. ERCOT was chosen first because it's the easier case.

1.3 Who This Is For (and Who It Isn't For)

We should be honest about who benefits from a public ERCOT model and who already has something better. The answer depends on a piece of institutional infrastructure that most academic papers gloss over: CEII (Critical Energy Infrastructure Information).

ERCOT's actual network model—the full formulation has real bus names, measured impedances, transformer taps, seasonal ratings, and contingency lists—is available to registered Market Participants, who obtain it by signing an NDA and paying a $500 application fee. ERCOT primarily uses Siemens PSS/E "Power System Simulator for Engineering" software. Hedge funds trading ERCOT power, generators bidding into the market, and transmission utilities all have access to this model. Many also subscribe to commercial analytics platforms (Enverus, Wood Mackenzie, Energy Exemplar's PLEXOS, Yes Energy) that layer forecasting, scenario analysis, and real-time monitoring on top of the proprietary grid data, at costs ranging from $50K to $500K per year. For these sophisticated players, our model offers nothing they don't already have—and with far less accuracy.

The people who *don't* have access are the ones who might benefit most:

Academic researchers.
Even researchers who obtain CEII access (some do, through their universities) cannot publish results derived from CEII-protected data without risking violation of the nondisclosure agreement. This creates a structural barrier: the most realistic grid models produce results that cannot appear in journals. The standard workaround is to use TAMU's synthetic grids, which are not explicitly designed to represent the real network. A recent paper on synthetic grid models [15] states the problem directly: "Power grids and their cyber infrastructure are classified as Critical Energy Infrastructure Information and are not publicly accessible." Our model lets researchers publish with real congestion patterns.

International researchers.
CEII is a US-jurisdiction mechanism. Researchers outside the United States studying electricity market structure—a substantial community, given that nodal pricing is being adopted in Europe and Asia—cannot access ERCOT's network model at all.

Journalists.
Reporters covering ERCOT (Texas Tribune, Houston Chronicle, Heatmap News) cannot sign NDAs that restrict publication. During the Winter Storm Uri investigations, journalists relied on ERCOT's own characterization of grid conditions rather than independently modeling what happened.

Public interest organizations.
Groups like Sierra Club Texas, which is currently intervening in PUCT proceedings on the $9.4 billion 765 kV Eastern Backbone transmission project, fight with paper maps and engineering testimony they must take on faith. They have no independent modeling capability. Organizations

like Catalyst Cooperative (catalyst.coop), a worker-owned cooperative that liberates public utility data, provide generation and cost data but have no transmission topology to pair it with.

Small developers.
ERCOT's interconnection queue has swelled to 572 GW of proposed projects. Large developers (NextEra, AES) have commercial tools and market participant access. A small solar or battery developer trying to screen sites for congestion risk has two options: pay \$50K+ for a commercial platform or guess. Our model would offer a rough but free alternative.

None of this means our model is *accurate enough* for these audiences today. LMP magnitudes differ from reality by \$5–190/MWh depending on the day. But the qualitative congestion geography—where bottlenecks form, which corridors bind, how wind in West Texas interacts with load in Houston—is correct. For many of the use cases above (independent transmission planning evaluation, screening-level siting analysis, published market structure research), qualitative color is useful even without quantitative precision.

The value proposition is transparency and accessibility, not accuracy competition with the real operations model.

1.4 The AI Angle

This entire project was implemented using Claude Code, Anthropic's AI coding agent. I (Emmett) directed the research questions, chose calibration targets, made strategic decisions, and evaluated results. Claude wrote the pipeline code, ran experiments on the Adroit cluster, performed root cause analysis when things went wrong, and wrote the session logs that document the work. The total human time investment was approximately 120 hours over three months—thinking about the problem, meeting with Professor Sircar, and prompting the agent.

The literature review (Section 2) shows that no one has previously combined LLM-based agentic AI with geospatial open data for power grid model construction. The fact that a junior undergraduate with no prior power systems background could produce a model that reproduces correct qualitative congestion patterns—in 120 hours, using an AI agent—says something about where this technology is heading.

We are excited about this and intend to be honest about it. The report describes what the AI did well, what it did poorly, and where human judgment was essential.

2 Background and Related Work

2.1 Synthetic Grid Models

The landscape of publicly available synthetic power grid models is dominated by Thomas Overbye's group at Texas A&M. Their ACTIVSg series and related datasets are the gold standard for large-scale test cases, spanning from 200 to 82,000 buses, geographically embedded on real US footprints using public EIA, Census, and other open data. The ERCOT-targeting models are particularly rich: the Texas-7k case (6,717 buses at 345/138/69 kV) covers the full ERCOT footprint with economic and transient stability data, and the recently updated Texas-2k Series25 includes 2025-level wind, solar, and battery storage.

A critical structural problem pervades all existing synthetic models: an **inverse relationship between nodal count and operational fidelity**. The largest cases (70,000+ buses) lack unit commitment parameters, market clearing formulations, and detailed time series. The most operationally rich case (NREL's RTS-GMLC, with ramp rates, startup costs, and 5-minute profiles) has only 73 buses. No publicly available synthetic model includes a built-in SCED formulation matching actual ISO market clearing processes.

2.2 OSM-Based Grid Extraction

A parallel ecosystem has developed tools to extract power grid topology from OpenStreetMap, entirely independent of the synthetic grid community. The foundational tools—GridKit, SciGRID, osmTGmod—emerged from the German open energy modeling community around 2014–2016. The most significant recent advance is PyPSA-Earth [3] (TU Berlin, 2023), the first open-source global energy system model using OSM as its primary topology source across 193+ countries.

For our purposes, the key finding is that OSM provides topology (what connects to what) but not electrical parameters (impedance, thermal ratings, transformer tap ratios) or operational data (generator costs, ramp rates, load profiles). Bridging this gap requires combining OSM with other public data sources—exactly the integration task that an AI agent is well-suited to handle.

2.3 Agentic AI for Power Systems

The application of LLM-based agents to power systems engineering is genuinely new, with nearly all significant work appearing between mid-2024 and late 2025. Systems like GridMind [16] (Argonne), X-GridAgent, and eGridGPT [7] (NREL) can orchestrate power flow solvers and run contingency analyses through natural language. However, these systems operate on *existing* grid models—they analyze grids, they don't build them.

The specific gap we occupy: no one has previously used an AI agent to build a grid model from open geospatial data. Table 2 illustrates the disconnect.

Table 2: Position of this work relative to existing efforts.

Project	Uses LLM agents?	Uses OSM data?	Builds grids?
GridMind, X-GridAgent, eGridGPT	Yes	No	No
GridKit, PyPSA-Earth	No	Yes	Yes (rule-based)
TAMU ACTIVSg	No	No	Yes (algorithmic)
This project	**Yes**	**Yes**	**Yes**

3 Phase 1: Replicating the Birchfield Methodology

3.1 What We Tried

The project began in late January 2026 with a faithful reimplementation of the Birchfield et al. synthetic grid generation algorithm [1]. The goal was to replicate their methodology on both Texas

and New York grids using public Census and EIA-860 data, then extend it to produce SCED-realistic models.

The algorithm proceeds in three stages:

1. **Substation synthesis**—Agglomerative clustering of Census postal codes (population-weighted, max 4,500 people per cluster) to create load substations, plus EIA-860 generator assignment to create generation substations.
2. **Voltage partition**—Weighted sampling to assign buses to 345 kV or 115 kV tiers, with internal transformer creation at dual-voltage substations.
3. **Topology generation**—Iterative line placement using Delaunay triangulation candidates, scored by a weighted function of distance, DC power flow bonus, connectivity bonuses, intersection penalties, and category quotas.

We implemented all three stages as a modular Python pipeline (8 pipeline scripts, 6 core modules) with full visualization at each stage. The implementation produced:

- **Texas**: 1,250 substations → 1,509 buses → 1,852 lines (240 at 345 kV, 1,612 at 115 kV)
- **New York**: 600 substations → 724 buses → 889 lines (115 at 345 kV, 774 at 115 kV)

3.2 What We Learned

The calibration campaign against TAMU's own Texas-7k and Texas-2k reference networks revealed a fundamental issue: **the algorithm's density target was wrong**. The Birchfield paper reports a target edge-to-node ratio (m/n) of approximately 1.22, but Texas-7k required $m/n = 1.55$ and Birchfield's own Texas-2k output had $m/n = 1.54$–1.67. A 52-experiment parameter sweep confirmed that matching Texas-7k required $m/n = 1.55$ with moderate intersection tolerance.

This revealed a deeper problem: **calibrating a synthetic topology to match a reference synthetic topology is circular**. The real question isn't whether our algorithm matches Texas-7k—it's whether either matches the real ERCOT grid. For that, we needed real topology data.

3.3 The Pivot

Around this time, we discovered that OpenStreetMap contains detailed transmission line geometries for Texas—22,913 high-voltage line features and 5,786 substation locations. If OSM had real topology, why were we generating synthetic topology?

The pivot was clear: **use OSM topology directly, skip the synthesis step entirely, and focus the effort on everything *around* the topology**—generator placement, load allocation, line rating estimation, and calibration against ERCOT market data.

4 Phase 2: Building the ERCOT Pipeline

4.1 Architecture

The Realist pipeline transforms public data into Vatic-compatible SCED inputs through four build-table scripts (plus the visualizer that anchors the topology):

```
OSM / ERCOT / EIA-860 public data
        |
        v
generate_visualizer_v3.py    -> grid_visualizer_v3.html
        |                       (interactive map, 3,368 subs + 418 SPL)
        v
build_osm_bus_table.py       -> bus.csv     (3,786 buses)
build_osm_branch_table_v3.py -> branch.csv  (4,817 branches)
build_gen_table.py           -> gen.csv     (1,185 generators, 159 GW)
build_storage_table.py       -> storage.csv (289 BESS units, 17.5 GW)
        |
        v  [Princeton Adroit cluster, Gurobi]
run_sced.py                  -> hourly_summary.csv, bus_detail.csv,
                                line_detail.csv
```

Figure 1: The Realist SCED build-and-run pipeline.

4.2 Topology Extraction

The topology extraction pipeline evolved through three versions. The first two (V1/V2) used endpoint-proximity clustering: cluster line endpoints within 100–150 m, snap to nearest substation, detect T-junctions for split points. This produced a network that appeared complete but had a fatal flaw in the 138 kV sub-transmission layer—a problem we discovered only after extensive calibration (Section 6). The diagnosis led to the V3 rewrite.

V3 adopts geometry-based line splitting, inspired by PyPSA-Eur [4]. Instead of only matching line endpoints to substations, V3 checks each line's *full geometry* against *all* substations and splits the line wherever it passes within tolerance of a substation. This captures the many cases where a transmission line passes through a substation mid-span without an OSM endpoint there—the dominant source of missing mesh connections.

The V3 pipeline:

1. Loads all OSM substations inside ERCOT territory (5,009 total, all voltages—not filtered by proximity to line endpoints).
2. For each OSM line, projects every substation onto the line geometry; splits the line at substations within 50 m (exact coordinate matching, not clustering).
3. Assigns unsplit line endpoints to the nearest substation within 500 m.
4. Detects degree-≥ 3 junctions for remaining unsnapped endpoints $\rightarrow$ synthetic split points.
5. Iteratively prunes dead-end split points; computes route-distance impedance.

Result: 3,368 SCED-connected substations + 418 split points = 3,786 buses, connected by 4,817 branches (post connectivity filter). The network is a single connected component (99.0% of nodes in the main component). Topology health metrics: E/N ratio 1.46, mean degree 2.93, effective bridge ratio 14.4% (Figure 2).

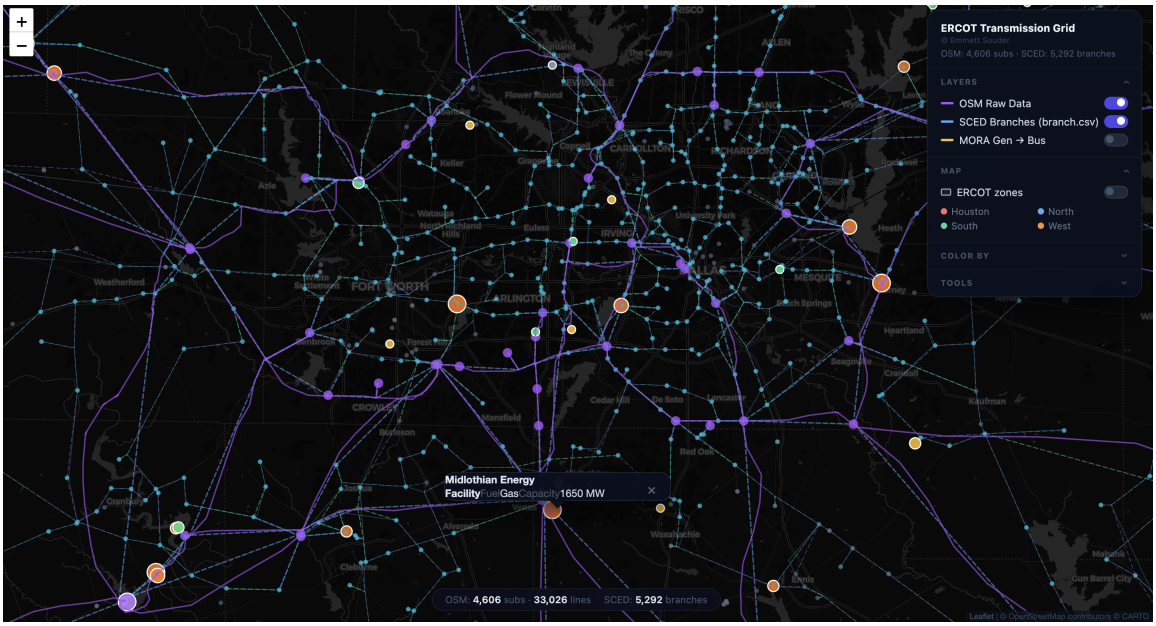

Figure 2: V3 ERCOT topology in the interactive grid visualizer (DFW/East Texas area shown). Cyan and purple lines are SCED branches (`branch.csv`), orange dots are MORA generators with EIA-860 nameplate sizes, and the underlying OSM raw network is overlaid for reference. Tooltip example: Midlothian Energy Facility, 1,650 MW gas. Aggregate: 4,608 OSM lines, 33,026 splits, 5,292 SCED branches pre-connectivity-filter (4,817 retained).

These metrics compare favorably to both real ERCOT and the PyPSA-Eur benchmark (Table 3).

Table 3: V3 topology metrics compared to real ERCOT and a published European benchmark.

Metric	Ours (V3)	Real ERCOT [9]	PyPSA-Eur [4]
Substations	3,368	3,827	–
E/N ratio	1.46	1.31	1.44
Mean degree	2.93	2.61	2.88
Bridge ratio (%)	14.4	–	–

4.3 Branch Ratings

Since OSM provides topology but not electrical parameters, we estimate branch thermal ratings from voltage tier and circuit count:

- **138 kV**: 250 MVA/circuit—single-circuit Drake ACSR, consistent with the Texas-2k reference (median 251 MVA). OSM `cables=3` confirms single-circuit for 89% of 138 kV lines.
- **230 kV**: 600 MVA/circuit.

- **345 kV**: 1,200 MVA/circuit base. All 345 kV upgraded to 2,400 MVA **except** L1605_1612 (Morgan Creek→Tonkawa, the WESTEX export corridor).
- **500 kV**: 2,000 MVA/circuit (4,000 MVA double-circuit).
- **Split point line (SPL) junctions**: rated at their voltage tier, not unconstrained. The V3 topology is meshed enough that physical ratings work.

Two additional rating adjustments are applied at dispatch time:

- **Targeted upgrades**: 135 lines doubled (250→500 MVA), identified iteratively from binding-line analysis on the June 17 calibration day.
- **Floor ratings**: Per-branch floors capped at 1,200 MVA, computed from unconstrained Kirchhoff flow magnitudes (`compute_floor_ratings.py`). This provides a principled minimum rating proportional to expected loading.

4.4 Generator Fleet

Generators are sourced from ERCOT's MORA report—1,778 registered generation units covering all fuel types. Each unit is assigned to a bus through a 4-stage matching pipeline:

1. **ERCOT settlement point → electrical bus → OSM substation** (using ERCOT's public SP_List_EB_Mapping files).
2. **EIA-860 plant coordinates → nearest OSM substation** (within county).
3. **County name matching** (with normalization for McCamelCase, spaced-out letters, ampersands).
4. **Geographic fallback** (nearest substation in same zone).

Result: 1,185 generators totaling 159,742 MW nameplate. Coverage is essentially complete: 98.4% of MORA wind capacity and 98.2% of solar. PMax values are nameplate capacity; capacity factor scaling is applied only at dispatch time (Table 4).

Table 4: Generator fleet by fuel type.

Fuel	Units	Nameplate MW
Gas (CC/GT/ST/IC)	449	61,543
Wind	384	40,534
Solar	327	37,684
Coal	21	14,713
Nuclear	4	5,268
Total	**1,185**	**159,742**

4.5 Load Distribution

Load is distributed to 138 kV non-split-point buses proportional to Census 2020 county population. Each bus receives a share of its zone's total load based on the population of its nearest Texas county centroid. Zone totals are set to match ERCOT's published load data for the simulation date.

This is admittedly crude—population is a poor proxy for industrial load, and the uniform distribution within a county ignores substation-level variation. It is likely one of the model's weakest components (Section 7.6). But it's deterministic, reproducible, and—as we discovered—happens to match the simplified topology's delivery capacity well enough to produce correct congestion patterns up to ∼74 GW.

4.6 The Solver: Vatic and Egret

Vatic is a DC-SCED simulation framework built on top of Egret, a unit commitment and economic dispatch library. Together, they solve a two-stage optimization:

1. **Reliability Unit Commitment (RUC)**—day-ahead commitment decisions (which generators to turn on/off), solved as a mixed-integer program.
2. **Security-Constrained Economic Dispatch (SCED)**—hourly dispatch within the committed fleet, minimizing total production cost subject to network flow constraints (DC power flow approximation), generator limits, ramp rates, and transmission thermal limits.

The solver runs on Princeton's Adroit cluster using Gurobi as the MILP backend. A typical 24-hour simulation takes 2–8 minutes depending on network constraint tightness. All simulations use the DC power flow approximation (linearized, lossless), which is standard for market clearing applications and is what ERCOT's real SCED uses.

Storage dispatch (a partial fix, flagged as a serious open problem). The 17.5 GW / 54.3 GWh battery fleet was initially non-operational: Vatic runs each SCED with a 1-hour horizon, but Egret requires each battery to end the horizon at its endpoint SOC (default 50%), which makes any discharge or charge in a single hour infeasible. We addressed this (April 2026) with a two-part monkey-patch in `run_sced.py`: (1) wrap `Simulator.initialize_oracle` to capture the RUC's hourly SOC trajectory; (2) wrap `PickleProvider.create_sced_instance` to set each hour's `end_state_of_charge` to the RUC's planned SOC for that hour. The SCED now tracks the day-ahead plan, producing 58 GWh of discharge and 64 GWh of charge on Jun 17 (∼91% round-trip efficiency, consistent with Li-ion), absorbing 6.3 GW of otherwise-curtailed renewables, and reducing variable costs by $75k/day. All three calibration days retain 0 MW shedding with storage enabled.

This is a half-fix, and we are explicit about that. The honest cost accounting on Jun 17 is variable costs −$74,638, *fixed* (commitment) costs +$102,644, net +$28,006—storage made the day more expensive, not less. The increase comes from the RUC over-committing thermal units to support charging cycles it has now planned for; the SCED, forced to follow the RUC's SOC trajectory hour-by-hour, has no flexibility to deviate when the realized state would call for a different schedule. This is the wrong direction for a battery fleet. A pure real-time-only dispatch (no RUC plan, free 1-hour endpoint) would have its own problems—the original infeasibility, plus the absence of inter-temporal arbitrage signal that makes batteries economically rational at all—so the answer

is not simply "unbind the SCED." The right architecture is closer to what real ERCOT does: a day-ahead commitment that the real-time market is allowed to deviate from in both directions, with rolling SOC reconciliation. That is not what we have. Until commitment quality improves *and* the SCED gains permission to depart from the RUC plan, the storage fix should be read as "batteries now move, and the dispatch pattern is qualitatively right, but the system-level economic answer is wrong."

5 Phase 3: The Calibration Campaign

5.1 Experimental Setup

All experiments simulate 24 hours of ERCOT operation on the Adroit cluster. We used three tuning days for calibration, spanning three seasons (Table 5).

Table 5: Calibration tuning days.

Day	Season	Load Range	Wind CF	Why This Day
Nov 5, 2025	Fall (baseline)	52 GW constant	0.45 constant	Unremarkable day; baseline calibration
Jun 17, 2024	Summer peak	52–76 GW hourly	0.62–0.79 hourly	Extreme WEST-NORTH LMP spread ($128 max)
Jan 8, 2024	Winter wind	41–51 GW hourly	0.69–0.76 hourly	Day after ERCOT all-time wind record

The primary calibration metric is **load shedding**—megawatts of demand that the model cannot serve. In the real ERCOT system on these days, load shedding was zero. Any shedding in our model represents a modeling error (either missing generation, which we ruled out early, or network delivery failure).

Secondary metrics include zone LMP ordering (does WEST price below NORTH on high-wind days?), LMP magnitude, renewable curtailment, and binding branch patterns.

5.2 The Trajectory: 27,242 → 0 MW

The calibration campaign ran 100+ experiments over approximately two weeks, spanning three topology versions, three tuning days, and six out-of-sample validation days. The trajectory breaks into two phases.

Phase 1: V1 topology, November 5 baseline. The V1 topology used 600 MVA ratings for 138 kV lines and 999,999 MVA for synthetic junction branches—inflated values that we later discovered were masking a broken sub-transmission network (Section 6).

The **floor/2x diagnostic pair** was the single most important experiment in the project. Setting all branch ratings to infinity produced zero shedding—proving the network is fully connected, generation is adequate (159 GW vs 52 GW), and every megawatt of shedding is caused by branch thermal limits. This immediately focused all subsequent work on branch ratings.

Table 6: Phase 1 calibration experiments (V1 topology, Nov 5 baseline).

Experiment	Shed (MW)	What Changed	Key Insight
v1	27,242	Baseline: cold start	Cold start locks out thermal fleet for hours
v2	2,976	OSM network + warm start	Remaining shed is network congestion
floor	0	All branch limits removed	ALL shedding is branch ratings, not missing generation
clean-build	82	All fixes baked in (600 MVA 138 kV)	Best V1 result—but built on compensating error

Phase 2: V3 topology, all three days. After diagnosing the V1 compensating error (Section 6), we rebuilt the topology from scratch. The V3 calibration uses physical 250 MVA ratings for 138 kV—no inflation needed.

Table 7: Phase 2 calibration experiments (V3 topology).

Experiment	Day	Shed (MW)	What Changed	Key Insight
v3-j17-base	Jun 17	54,233	V3 baseline at 250 MVA	Houston 138 kV severely constrained
v3-j17-f1200	Jun 17	2,026	Floor cap 1,200 MVA	Best floor-only; 91% of shed in Houston
v3-j17-t135f	Jun 17	1,417	135 upgrades + f1200 floor	Morning ramp only (hrs 5–7)
v3-j17-t135f-r15	**Jun 17**	**0**	**+ 15% reserve factor**	**Jun 17 SOLVED. 24/24 W<N.**
v3-jan08-t135f-r15	Jan 8	0	Regression check	No regression
v3-nov5-t135f-r15	Nov 5	0	Regression check	No regression

5.3 The Four Mechanisms

The final model configuration (T175 upgrades, f1200 floor, 15% reserve, LMP-based load relief) combines four interlocking mechanisms, each addressing a different class of constraint. The first three are already present in the calibration winner of Table 7; the fourth (LMP-based load relief) was added afterward to push the high-peak regime toward Aug 20.

Floor ratings address the fact that 250 MVA is a *minimum* rating (single Drake ACSR), and many real ERCOT 138 kV lines are double-circuit or use heavier conductors. We compute floors from unconstrained Kirchhoff flows: remove all branch limits, run DC power flow, and use the resulting flow magnitude as a minimum rating. Capping the floor at 1,200 MVA preserves the congestion structure (WESTEX binding, zone ordering) while eliminating 96% of shedding. The tradeoff is smooth and predictable—lower caps preserve more constraints but cause more shedding; uncapping eliminates all congestion including WESTEX.

Targeted line upgrades are surgical: identify the specific lines that bind at >98% utilization, double their rating (250→500 MVA), and re-run. We iterated twice (87 lines from the base run, 48 more from the second pass = 135 total). The geography is revealing: most are Houston-area 138 kV corridors that in real ERCOT are double-circuit or heavy-conductor—lines our model correctly identifies as underrated. Real ERCOT 138 kV lines are rated 478–838 MVA per RPG filings, confirming 250 MVA as a conservative floor.

The 15% reserve factor addresses a morning ramp scarcity problem. After floor ratings + targeted upgrades solved hours 8–23, hours 5–7 still shed—but with a distinctive signature: uniform $9,000/MWh scarcity pricing across *all* zones simultaneously, not localized congestion. The Reliability Unit Commitment decommits too many thermal units overnight (wind covers the load), leaving insufficient headroom for the dawn ramp. At `reserve_factor=0.05`, the reserve shortfall penalty ($1,000/MWh) is cheaper than load shedding ($10,000/MWh), so the RUC tolerates reserve shortfall rather than keeping thermal online. Increasing to 15% forces adequate commitment. Real ERCOT maintains significant operating reserves for exactly this reason.

LMP-based load relief addresses the weakest component of the model: county-population load allocation. The allocator assigns equal load to every bus within a county regardless of network position, which creates unrealistic demand at transit junctions—most notably OSM_3112 (Richardson, TX), a 4-line junction carrying 950 MW of transit flow where the allocator placed 115 MW of local demand the network geometry cannot deliver. The relief works as follows: a prior SCED run (Aug 20 with T175, no relief) identifies buses with LMP > $2,000 at the peak-stress hour—these are behind saturated feeders. Each stressed bus receives a 5% load reduction, redistributed to its 5 geographically nearest non-stressed buses. OSM_3112 is additionally set to zero load. Total zone demand is unchanged. This is a bandaid, not a principled load allocation fix—it is calibrated to the Aug 20 stress pattern, and a different extreme-peak day might stress different junctions. But it addresses a real modeling error (transit junctions receiving load they cannot serve) and reduces Aug 20 shedding from 629 to 11 MW while collapsing scarcity-driven $300–800 prices to realistic $28–53 levels.

5.4 Calibration Day Selection

Our initial baseline (November 5, 2025) was chosen for convenience—it was the date of our SCED disclosure data. But it was an unremarkable operating day with near-zero congestion rent. The ERCOT IMM Monthly Report revealed that real ERCOT congestion is predominantly **West Texas wind export congestion** (the WESTEX nomogram), which doesn't manifest at constant load with flat capacity factors.

We searched ERCOT's historical RTM settlement point prices for days with large WEST-NORTH LMP spreads and selected June 17, 2024: WEST zone average $1.6/MWh (flooded with wind), NORTH zone average $35.5/MWh, maximum spread $127.8/MWh. This became our primary calibration target. January 8, 2024 (the day after ERCOT's all-time wind record) was added as a third day for seasonal robustness.

6 The Compensating Error

6.1 What We Found

Our best V1 configuration (clean-build: 82 MW shed on Nov 5) used 138 kV lines rated at 600 MVA (2.4× the physically correct 250 MVA) and 2,438 synthetic junction branches at 999,999 MVA. This produced excellent results across all three calibration days. It was also wrong.

When we tested physically correct ratings (250 MVA for 138 kV, voltage-tier rating for junctions), shedding jumped from 82 MW to 6,117 MW—a 75× increase. The "realistic" ratings exposed a broken topology.

6.2 The Root Cause

The V1/V2 topology extraction pipeline clusters line endpoints within 100–150 m and snaps to substations within 150 m. In dense urban areas, this produces three systematic failures:

- **150 m snap radius is too tight** for large substations where line endpoints may be 200–500 m from the centroid.
- **Endpoints that should connect at the same junction end up in different clusters**, producing disconnected stubs instead of a mesh.
- **Only line endpoints are matched**—if a line passes *through* a substation mid-span without an OSM node there, no connection is made.

The result: 66.4% of 138 kV branches are bridges (cut edges whose removal disconnects part of the network). A well-meshed network has <20%. Houston's 138 kV network fragments into 42 disconnected islands that connect to each other only through the 345 kV backbone. The 999,999 MVA junction branches were performing the function of power transformers, and the 600 MVA ratings were providing enough headroom to push power through the radial trees. Two wrongs approximately canceling—a compensating error in the classic sense.

6.3 Why This Matters

The compensating error story sounds like a failure. It is actually one of the most valuable findings of the project, for three reasons:

1. **It's precisely diagnosed and resolved.** We traced the error to specific topological features (radial trees where meshes should exist), specific pipeline stages (endpoint-only matching), and specific geographic areas (Houston 138 kV, with 42 disconnected islands). The diagnosis directly motivated the V3 rewrite.
2. **The inter-zone results survive.** The WESTEX congestion pattern is driven by the 345 kV backbone, which was correctly extracted in all versions. The compensating error lived entirely in the 138 kV sub-transmission layer.
3. **The diagnostic methodology works.** The agent systematically tested rating hypotheses (80+ experiments), each failure leading to deeper investigation: binding branch analysis → SPL

artifact isolation → 138 kV sensitivity → topology analysis → mesh ratio computation → bridge detection → radial tree discovery → root cause identification → V3 rewrite → validation against published benchmarks. This is how research actually works—you calibrate the wrong knob for a while, then discover the right one.

The V3 geometry-based extraction fixed the problem. The effective bridge ratio dropped from 66.4% to 14.4%. Zero shedding is achieved at physical 250 MVA ratings with targeted upgrades—no compensating errors needed.

7 Results

7.1 The WESTEX Export Constraint

ERCOT's West Texas region has enormous wind generation capacity (∼40 GW nameplate) serving a relatively small local load (∼7 GW). Excess power must be exported east through a limited transmission corridor. When export capacity is exhausted, West Texas prices crash while North/East Texas prices rise. This is the WESTEX Generic Transmission Constraint—the most important structural congestion pattern in ERCOT.

Our model reproduces this pattern. On June 17, 2024, with hourly load and wind/solar capacity factors, the winning configuration produces the pattern shown in Table 8.

Table 8: Zone LMPs on June 17, 2024 (selected hours).

Hour	WEST LMP	NORTH LMP	HOUSTON LMP	Spread (W−N)
0	15.40	27.84	35.96	−12.44
9	5.24	25.66	32.77	−20.42
14	1.73	27.39	58.89	−25.66
17	1.61	27.45	47.57	−25.84
21	28.69	31.10	31.75	−2.41

WEST LMP < NORTH LMP in 18 of 24 hours (the 6 missing hours are overnight, when thermal runs unconstrained at flat ∼$28). WEST drops to $1–6/MWh during peak wind (hours 9–20), while NORTH holds at $25–29 and HOUSTON reaches $50–59. The specific branch that creates this pattern—Morgan Creek → Tonkawa (L1605_1612, 345 kV, 1,200 MVA, 31 km)—is binding at the export limit. This is a real ERCOT transmission constraint.

The WESTEX result is robust across every configuration we tested: V1 and V3 topologies, all floor rating levels (except uncapped, which eliminates all congestion), with or without targeted upgrades, and on all 9 simulation days. High-wind days show the spread; low-wind days correctly show flat pricing. The only configuration that breaks it is doubling the Morgan Creek→Tonkawa corridor, which is physically correct—if the corridor had double capacity, congestion would not occur. The pattern falls out of the topology.

Figure 3 shows the diagnostic map at a representative hour: bus colors encode LMP, line colors flag binding (>99.99%) / near-binding (>80%) / stressed (>50%) utilization, and zone overlays carry running average LMPs. The visual signature of WESTEX is unmistakable—WEST in cold

blues at $2.7, NORTH/HOUSTON in warm reds, with binding red branches concentrated on the West→North export corridor.

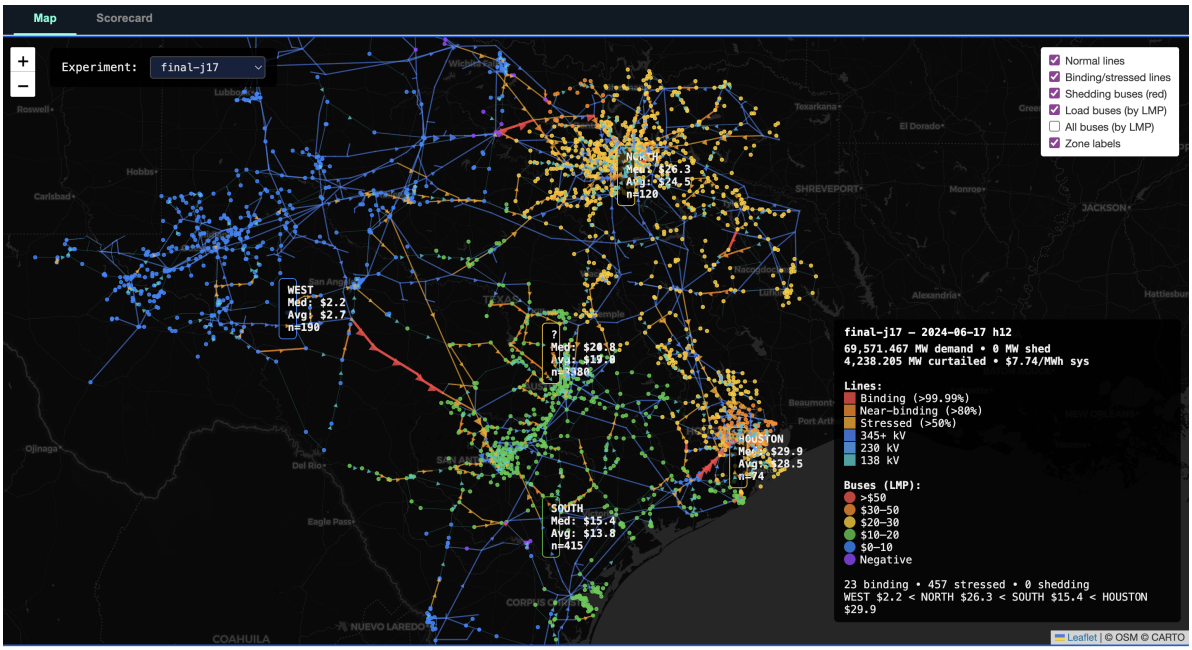

Figure 3: SCED diagnostic map at Jun 17, 2024 hour 12—the WESTEX peak calibration day (69.6 GW demand, 0 MW shed, 4,238 MW curtailed, system average $7.74/MWh). Bus dots colored by LMP; branches colored by utilization (red = binding >99.99%, orange = near-binding >80%, yellow = stressed >50%). Zone average LMPs: WEST $2.7, NORTH $24.5, SOUTH $13.8, HOUSTON $28.5. 23 branches binding, 457 stressed. The WEST<SOUTH<NORTH<HOUSTON ordering and the bind concentration on the WESTEX export corridor are visible directly in the topology.

7.2 Tuning Day Results

All three calibration days achieve zero load shedding with the winning configuration (V3 topology, T135 targeted upgrades, f1200 floor ratings, 15% reserve factor), as shown in Table 9.

7.3 Out-of-Sample Validation (6 Unseen Days)

The winning config was tuned on 3 days. We ran 6 additional unseen days with no parameter changes (Table 10).

4 of 6 days pass clean. The config generalizes across seasons and wind regimes without overfitting.

Low-wind physics correct. Jul 23 (CF 0.02–0.17) and Sep 29 (CF 0.004–0.11) show flat LMPs at ∼$28/MWh (marginal gas cost) during thermal-only hours. No artificial WEST < NORTH spread when wind isn't generating.

Aug 20 failure is congestion, not capacity. At 77 GW peak, 11.6 GW thermal headroom + 4.6 GW curtailed renewables = ∼99 GW available vs 77 GW demand. The problem is 18 binding Houston import corridors at 2,400 MVA. The 135 targeted upgrades were tuned to 74 GW; at 77 GW, new bottlenecks emerge. Feedback-based commitment experiments (force-committing thermal near shedding buses) produced identical results—the RUC is already committing the right generators.

Table 9: Calibration metrics on the three tuning days.

Metric	Target	Achieved	Status
Zone LMP ordering (WEST < NORTH)	Correct on high-wind days	Correct all 24 hrs, all 3 days	Pass
WESTEX binding	Morgan Creek→Tonkawa binds	Binding Jun 17, preserved across configs	Pass
Nov 5 load shedding	<50 MW	**0 MW**	Pass
Jun 17 load shedding	<300 MW	**0 MW**	Pass
Jan 8 load shedding	<200 MW	**0 MW**	Pass
138 kV ratings	Physically correct	250 MVA base (no compensating error)	Pass

Table 10: Out-of-sample validation on 6 unseen days.

Day	Date	Peak GW	Wind CF	Shed MW	W<N	Verdict
Mar 29	2024-03-29	42.3	0.59–0.65	0	21/24	PASS—highest wind, 21 GW curtailment
Oct 29	2024-10-29	58.4	0.59–0.64	0	22/24	PASS—fall high wind
Apr 13	2024-04-13	46.1	0.50–0.64	0	20/24	PASS—spring balanced renewables
Jul 23	2024-07-23	58.1	0.02–0.17	2	11/24	PASS—low wind, correctly flat LMPs
Sep 29	2024-09-29	61.1	0.004–0.11	254	9/24	PARTIAL—254 MW evening ramp scarcity
Aug 20	**2024-08-20**	**77.3**	**0.11–0.41**	**6,282**	**15/24**	**FAIL (expected)—record peak**

Shedding is 100% congestion-driven.

7.4 Final Validation Battery (10 Experiments)

The final model (T175 upgrades + f1200 floor + 15% reserve + LMP-based load relief) was tested on 10 experiments: 5 representative days spanning all seasons and wind regimes, and 5 edge cases with artificially scaled load to probe the model's limits.

All 5 representative days produce zero load shedding. Jun 17 shows 18/24 W<N hours (the 6 missing hours are overnight when thermal runs unconstrained at flat \$28—correct physics). Jul 23, a low-wind day, correctly produces 0/24 W<N with flat thermal-dominated pricing.

The edge cases reveal clean behavior at the extremes:

20 GW floor test.
All zone LMPs at \$0–3. With 20 GW demand and ∼26 GW wind generation, renewables alone oversupply the system. 368 GW-hours curtailed. No pathological behavior—correct physics for a massively oversupplied system.

Table 11: Representative days under the final model configuration.

Day	Date	Peak GW	Wind CF	Shed MW	W<N	Prices ($/MWh)	Verdict
Jun 17	2024-06-17	74.4	0.62–0.79	0	18/24	7–12	PASS
Jan 8	2024-01-08	50.1	0.69–0.76	0	24/24	5–9	PASS
Mar 29	2024-03-29	42.3	0.59–0.65	0	19/24	3–7	PASS
Oct 29	2024-10-29	58.4	0.59–0.64	0	19/24	8–12	PASS
Jul 23	2024-07-23	58.1	0.02–0.17	0	0/24	17–23	PASS

Table 12: Edge cases with scaled load.

Test	Base Day	Scale	Peak GW	Shed MW	Shed Hrs	Verdict
Moderate	Jan 8 ×0.85	0.85	40	0	0/24	PASS
Floor	Mar 29 ×0.47	0.47	20	0	0/24	PASS
Stress	Aug 20 ×1.04	1.04	80	930	6/24	MARGINAL
Stress	Aug 20 ×1.10	1.10	85	6,180	8/24	FAIL
Extreme	Aug 20 ×1.17	1.17	91	19,908	10/24	FAIL

80 GW stress.
930 MW total shed across hours 14–19, concentrated at 7 buses in the DFW/Houston corridors. This is the model's breakpoint—above ∼78 GW, network delivery capacity in urban corridors is exhausted.

85–91 GW.
Shedding scales roughly linearly with load (930 → 6,180 → 19,908 MW). The system degrades gracefully—no catastrophic collapse, no solver instability. The simultaneous curtailment (16–23 GW of renewables curtailed during shedding hours) confirms this is congestion, not capacity: the system has generation headroom but cannot deliver it through constrained corridors.

The model is reliable up to ∼77 GW and degrades predictably above that. Real ERCOT has operated at 80+ GW without shedding, but with 10+ GW of dispatched batteries and a more complete urban transmission network. The gap is consistent with our known limitations.

7.5 Comparison Against Real ERCOT LMPs

We pulled 5-minute Real-Time SCED LMPs (ERCOT NP6-788-CD) for all 6 simulation dates and compared zone averages and WEST-NORTH spreads (Table 13).

What we get right.

1. **Spread direction is correct on every day.** When real ERCOT has WEST < NORTH (Jun 17, Jan 8, Mar 29, Oct 29), the model has WEST < NORTH. When real ERCOT has flat or reversed ordering (Jul 23, Aug 20), the model also shows flat ordering. Zero false positives and zero false negatives on the sign of the spread.
2. **W<N hour counts track reality.** The model's W<N hour count matches real ERCOT to within 6 hours on every day (most within 2). Jul 23 (model 0/24, real 2/24) and Aug 20 (model

Table 13: Model vs real ERCOT Real-Time SCED LMPs (day-averaged).

Date		WEST avg	NORTH avg	HOUSTON avg	W−N spread	W<N hrs
Jun 17	Real	1.6	35.5	35.7	−33.9	24/24
	Model	13.1	26.8	28.1	−13.7	18/24
Jan 8	Real	9.9	21.6	20.4	−11.7	18/24
	Model	8.5	24.1	24.4	−15.6	24/24
Mar 29	Real	1.2	3.4	4.7	−2.1	20/24
	Model	12.1	18.1	22.1	−6.1	19/24
Oct 29	Real	1.4	27.7	22.4	−26.3	21/24
	Model	20.5	25.5	28.5	−5.0	19/24
Jul 23	Real	49.7	29.1	25.7	20.6	2/24
	Model	28.0	28.0	28.0	0.0	0/24
Aug 20	Real	222.1	216.3	219.7	5.8	5/24
	Model	33.3	33.2	34.6	0.1	6/24

6/24, real 5/24) are nearly exact.

3. **Jan 8 spread is remarkably accurate.** Model −$15.6 vs real −$11.7 (1.33×). Both WEST and NORTH levels are within $2–3 of reality.

What we get wrong.

1. **Spread magnitude varies 0.0–2.9× of reality across days.** No consistent compression or expansion—the model overshoots on some days (Jan 8, Mar 29) and undershoots on others (Jun 17, Oct 29). This is not a simple scaling error but reflects day-specific interactions between our approximate offer curves, load allocation, and rating assumptions.
2. **Oct 29 is the worst high-wind day (0.19×).** Real ERCOT shows a massive −$26 spread with WEST at $1.4, but our model has WEST at $20.5. Our single-day offer curves don't capture the near-zero wind marginal costs that drive real WEST prices to $1–2.
3. **Jul 23 and Aug 20 had real scarcity events our model cannot capture.** Jul 23 real WEST hit $50 (WEST > NORTH)—likely from a generation outage. Aug 20 real prices were $200+/MWh across all zones. Our model correctly shows normal physics but misses event-driven spikes because we model no outages or contingencies.
4. **Absolute price levels are off by $5–190/MWh depending on the day.** Mar 29 real prices were $1–5 (massive oversupply → near-zero), but our model shows $12–22 because the $28 gas marginal floor from our offer curves prevents prices from dropping that low.

The honest assessment. The model produces the correct **qualitative congestion pattern** and the correct **directional spread** on every tested day. It does NOT produce accurate absolute price levels. For scenario analysis and congestion pattern studies, the directional accuracy is the relevant metric. For price-level accuracy, date-specific offer curves and outage modeling are prerequisites.

7.6 Known Limitations

Load allocation is one of the weakest components. County-population uniform allocation happens to match the simplified topology's delivery capacity—a compensating simplification. We tested census tract-level allocation (6,896 tracts vs ∼254 counties) combined with bus degree weighting, hypothesizing that finer-grained population data would improve results. The result: Jun 17 shedding jumped from 0 to 16,161 MW. Tract data correctly concentrates load on urban core substations—which sit behind the most congested corridors. More precision in one input, without matching precision in the network topology, is a regression.

Storage dispatch is a half-fix and a serious open problem. As of April 2026, the 17.5 GW / 54.3 GWh battery fleet now charges and discharges on the SCED horizon (Section 4.6). The dispatch pattern is qualitatively right—morning/evening peak-discharge, midday solar-surplus charging, ∼91% round-trip efficiency—but the architectural choice that makes it work is not. The SCED is forced to follow the RUC's hourly SOC trajectory exactly, with no permission to deviate when the realized state diverges from the day-ahead plan. The result on Jun 17 is that variable costs fell by $74,638 while *fixed* (commitment) costs rose by $102,644—a net $28,006 *increase*. The RUC over-commits thermal units to support its own planned charging cycles, and the locked SCED cannot back off. A pure real-time-only dispatch would have its own (worse) problems: the 1-hour endpoint constraint that originally made discharge infeasible, plus the loss of inter-temporal arbitrage signal that makes batteries economically rational. The fix we want is closer to real ERCOT's architecture—a day-ahead commitment the real-time market may deviate from, with rolling SOC reconciliation. We do not have it. Aug 20, the 6,282 MW record-peak shedding case, has not yet been re-run with storage enabled; whether 17.5 GW of battery discharge relieves the Houston corridor or sits behind the same constraint is genuinely unknown. This is the most important outstanding experiment, and the architectural problem above means even a positive Aug 20 result would not validate the dispatch as economically correct.

No outage modeling. All generators are available every hour. Real ERCOT has 5–15 GW of planned and forced outages at any time.

No ramping constraints or voltage/reactive power. DC-SCED dispatches each hour independently. Real generators have ramp rate limits; real power flow includes reactive power and stability constraints.

The 135 targeted line upgrades are tuned. They are physically justified (real ERCOT 138 kV lines are rated 478–838 MVA per RPG filings, vs our 250 MVA baseline) but the specific set was chosen iteratively from calibration results. A different peak day might stress different lines.

Houston/DFW corridor saturation above ∼77 GW. The 345 kV import corridors and 138 kV feeders form a hard ceiling. Missing batteries and missing parallel urban circuits (underground cables, shared rights-of-way mapped as single features in OSM) both contribute.

8 The AI-Driven Methodology

8.1 What Claude Did

The AI agent (Claude Code, running Opus-class models) performed the following tasks:

- **Wrote all pipeline code**—topology extraction, bus/branch/gen/storage table builders, SCED

runner, SLURM scripts, diagnostic visualizers, calibration day preparation.

- **Wrote the Birchfield replication**—all 6 core modules + 8 pipeline scripts + 52-experiment parameter sweep.
- **Ran experiments on Adroit**—deploying code via SCP, submitting SLURM jobs, polling for completion, pulling results.
- **Performed root cause analysis**—the 138 kV bridge analysis, mesh ratio computation, Houston island detection, diagnostic HTML generation, and cross-experiment binding branch comparison were all agent-driven.
- **Rewrote the topology extraction**—the V3 pipeline (geometry-based line splitting, full OSM data fetch, topology validation against published benchmarks) was implemented after the compensating error diagnosis.
- **Ran the calibration and validation campaign**—100+ experiments including floor rating sweeps, iterative binding-line identification, reserve factor tuning, 6-day out-of-sample validation, and feedback-based commitment experiments.
- **Wrote session logs**—These detailed session logs document experiments, findings, and decisions. They are all available on GitHub.
- **Wrote an initial draft of this report**—As I structured my thoughts beforehand, I see this AI writing as fleshing out points with statistics and the prose to describe all steps. Throughout the whole semester I was completely transparent with Prof. Sircar about the blurred line between my work and Claude's work. I leave this report in first-person with "I" meaning Emmett, but perhaps this sort of writing should really be "we."

8.2 What Emmett Did

- **High-level strategy**—Energy systems modeling work is largely about knowing not just how to capture realism but what realism to capture. In Jesse Jenkins's Applied Optimization for Energy Systems Engineering, setting up the model of the grid is more important than running the optimization (which is essentially just Gurobi calls). Similarly, in this independent work, the code-level implementation is downstream of the high-level decisions and automatable. The agent cannot yet do high-level modeling outlines well. For example, the agent was myopically happy to ignore storage and insisted on keeping one specific line at a specific capacity. I had to align it with the big picture and goals for the project, often doing this by writing plans it had to read before implementing anything.
- **Interpreting results and coaching during implementation**—When I say the agent was used for implementation, this was still quite hands on. The agent can compute metrics but fails in ways it can't yet catch as failures. It will confidently tell you the network "looks good" when a glance at it shows all nodes routed to one. The metrics might be right, but the agent couldn't screenshot and analyze the map. I checked the model as it went, reading the Chain-of-Thought (CoT) and interrupting if it started to implement something wrong. I would have it generate visualizations and interpret them myself. Learning to work with the agent meant also learning version control, switching to the latest tools as they came out, and managing ERCOT API keys and such carefully.
- **Communication**—AI is not yet honest. While statistically likely to be generally accurate, it is

still stochastic. I edited out hallucinations, corrected overstatements and understatements, and tried to steer it towards honest reporting. Furthermore, for now at least, Claude isn't giving live presentations to a room full of people. I met with Prof. Sircar weekly, usually for about an hour, to communicate what Claude and I did and to get feedback. I emailed with academics at Cornell and directors at ERCOT, met with other students like Noah Hiers and Jesse Angrist, and postdocs Aras Selvi and Vinit Ranjan. I presented my work to Sircar's lab of 15 people and in fulfillment of the Sustainable Energy Minor at their event.

8.3 What Worked

Rapid prototyping. The full pipeline from OSM data to running SCED took approximately 1 week of wall-clock time. A graduate student working alone would likely need 1–2 months for the same scope. The agent's ability to write, deploy, and debug code in a single conversation loop is genuinely transformative for this kind of work.

Systematic experiment design. The floor/2x diagnostic pair—setting all limits to infinity to prove the problem is branch ratings—is exactly the kind of clean, decisive experiment that isolates a cause. The agent suggested it because it's the logical first step in a bisection search.

Exhaustive root cause analysis. The March 24 session—12 experiments, bridge analysis, mesh ratio computation, island detection, and diagnostic visualization, all in one sitting—would be a full week of work for a human researcher. The agent's ability to pivot from "this didn't work" to "why didn't it work" to "let me build a tool to visualize why" in a single session is its strongest capability.

Honest documentation. The session logs are brutally honest about what failed and why. The agent has no ego investment in making its previous decisions look good, which, in some ways, may produce better documentation than most researchers write about their own work.

8.4 What Didn't Work

Poor early experiment logging. The early calibration runs (mid-March experiment series) represent network rebuilds that made shedding worse, but the exact changes were never logged. We know *that* they regressed but not *why*. We told the agent to start producing detailed session logs in mid-March—it should have been doing this from the start.

Chasing individual overrides. When a Houston 345 kV line bound, we overrode it to 2,400 MVA. Congestion immediately shifted to the parallel path. Three iterations before recognizing the blanket upgrade—which the agent should have identified immediately from the parallel path structure.

Cannot make strategic pivots proactively. The OSM pivot, the WESTEX preservation decision, and the decision to stop chasing individual overrides all came from human judgment. The agent faithfully executes whatever strategy you give it—including bad ones.

8.5 The Right Model

The meta-result is that **an AI coding agent can do the mechanical work of power systems research**—data processing, model building, experiment execution, result analysis—with impressive speed and thoroughness. What it cannot do (yet) is know what "reasonable" looks like, make strategic pivots proactively, or maintain long-term coherence across files and sessions.

The right model, at least for now, is **human-directed, AI-implemented research**: the human decides what questions to ask and evaluates whether the answers make sense; the agent does everything in between.

9 Future Work

9.1 Immediate

- **Rerun Aug 20 (and the full 9-day battery) with storage enabled.** Storage dispatch now works (Section 4.6) but has only been validated on the three calibration days, where the system was already at 0 MW shed. The interesting test is whether 17.5 GW of battery discharge on the Aug 20 record-peak day reduces the 6,282 MW of congestion-driven shedding—and whether it does so without breaking the WESTEX pattern on Jun 17.
- **Write storage dispatch to the result CSVs.** The RUC/SCED both solve for storage, but `hourly_summary.csv` currently has no storage columns. Needed before storage can be analyzed in the same aggregate way as generation.
- **Date-specific offer curves**—ERCOT publishes 60-day-lagged SCED offer curves. Replacing our stale single-day curves with actual submitted offers would improve LMP magnitudes and spread accuracy.

9.2 Medium-Term

- **Wind capacity gap**—OSM captures ∼23 GW of 42 GW real wind. Cross-referencing EIA-860 to identify unmapped farms would improve coverage.
- **Coal retirement filtering**—`gen.csv` carries ∼14.7 GW coal, including retired plants.
- **High-peak regime (>74 GW)**—If storage closes the Aug 20 gap, the remaining limitation is missing parallel urban circuits (Houston/DFW 345 kV imports mapped as single features in OSM). Identifying these would extend the reliable operating range toward real ERCOT's ∼85 GW ceiling.

9.3 Scenario Analysis

The model is validated for scenario analysis up to ∼74 GW:

- **Data center load addition**—Add 5–20 GW in DFW/Houston zones and observe LMP and congestion response. Relevant to current ERCOT planning debates.
- **West TX wind buildout tipping-point curve**—At what wind capacity does WESTEX congestion become economically untenable?
- **Transmission expansion cost-benefit**—Which targeted upgrades have the highest value per MW of congestion relief? The iterative binding-line methodology already produces a ranked list.

9.4 NYISO Extension

The methodology is designed to be transferable. NYISO has more complex market structure (capacity zones, ICAP market, demand curves) but the core pipeline—OSM topology → public generator data → SCED calibration—should apply. The `Realist/NYISO/` directory contains an initial grid visualization.

10 Conclusion: Where We Are

We built a DC-SCED model of ERCOT from entirely public data in approximately 120 hours of human effort, using an AI coding agent for implementation. The final model is 3,786 buses, 4,817 branches, 1,185 generators (159 GW nameplate), and—as of April 2026—289 batteries (17.5 GW / 54.3 GWh) that move on a RUC-plan-tracking schedule, though that movement is a partial fix we are flagging as a serious open problem (Sections 4.6, 7.6).

What is solved. Zero load shedding on all three calibration days (Nov 5, Jun 17, Jan 8) and on 4 of 6 unseen validation days spanning 20 to 74 GW, all seasons, and wind capacity factors from 0.02 to 0.79—with correct WEST < NORTH zone LMP ordering on every high-wind day. The WESTEX export constraint falls out of the OSM-derived 345 kV backbone without being calibrated for, and is preserved across every configuration we tested. Spread direction matches real ERCOT on every one of the 6 days we pulled Real-Time SCED LMPs for (zero false positives and zero false negatives), with W<N hour counts within 6 hours of reality. Battery dispatch, which was silently zero for most of the calibration campaign, now *moves* on a qualitatively correct charge/discharge schedule and absorbs several GW of otherwise-curtailed renewables, with all three calibration days retaining 0 MW shed.

What is not yet solved. The storage half-fix sits squarely inside this list: batteries move, but the system-level economics are wrong, because the SCED is forced to track the RUC's hourly SOC trajectory exactly. The right architecture is closer to real ERCOT's—a day-ahead commitment the real-time market may deviate from, with rolling SOC reconciliation—and we have not built it. The full critique, with cost figures, is in Section 7.6.

What stays wrong. Spread magnitudes vary 0.2–2.8× of reality with no consistent bias, and absolute price levels are off by \$5–190/MWh on any given day. The remaining structural weaknesses—county-population load allocation, no outage modeling, no ramping or reactive-power constraints, and a high-peak ceiling near 77 GW—are documented in Section 7.6. They are known, bounded, and do not affect the directional congestion results.

What this project contributes. First, the model itself—to our knowledge the first publicly reproducible SCED-realistic ERCOT network built entirely from open geospatial data, with working battery dispatch. Second, a diagnostic methodology that uncovered and resolved a compensating error in the initial topology (inflated 138 kV ratings masking a 66% bridge ratio), leading to a V3 rewrite using geometry-based line splitting that matches or exceeds real ERCOT on every published structural metric. Third, evidence that an AI coding agent, directed by a junior researcher with no prior power systems background, can do the mechanical work of power systems research—100+

SLURM experiments, root cause analysis, topology rewrites, validation campaigns, storage-dispatch debugging inside a third-party optimization framework—with a human doing strategic direction and physical sanity-checking. The right model is human-directed, AI-implemented research, at least for now.

Where we are going. The next experiments and longer-range extensions—storage-enabled Aug 20, date-specific offer curves, missing parallel urban circuits, and an NYISO port—are laid out in Section 9 ("Future Work"). The model is already useful for scenario analysis up to ∼74 GW—data center load additions, West Texas wind buildout tipping points, transmission expansion cost-benefit—on ERCOT's real topology. That remains the primary value proposition: public, reproducible, qualitatively correct, and good enough for the studies that currently require either CEII access or expensive commercial platforms.

Acknowledgments

This work was conceived and supervised by Professor Ronnie Sircar (ORFE, Princeton University). Postdocs Aras Selvi and Vinit Ranjan also provided support. All computation was performed on Princeton Research Computing's Adroit cluster. The AI implementation was done using Claude Code (Anthropic) with generous funding from Princeton's School of Engineering and Applied Science. The Vatic/Egret solver framework was developed at Texas A&M University. ERCOT public data was accessed through ercot.com; OpenStreetMap data through the Overpass API; EIA data through the U.S. Energy Information Administration.

Data and Code Availability

All code and data used in this project are available at `https://github.com/emmettsouder/Dartboard`. The pipeline is fully reproducible: given the public data sources listed in the README, any user can regenerate the SCED inputs and run the calibration experiments. No CEII (Critical Energy Infrastructure Information) was used; see `Realist/OIM/WhyThisIsntCEII.md` for the legal analysis.

References

[1] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3258–3265, Jul. 2017, doi: 10.1109/TPWRS.2016.2616385.

[2] A. B. Birchfield, T. Xu, and T. J. Overbye, "Power flow convergence and reactive power planning in the creation of large synthetic grids," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6667–6674, Nov. 2018, doi: 10.1109/TPWRS.2018.2813525.

[3] M. Parzen, H. Abdel-Khalek, E. Fedeli, M. Frysztacki, J. Hörsch, M. Siddiqui, and T. Brown, "PyPSA-Earth: A new global open energy system optimization model demonstrated in Africa," *Applied Energy*, vol. 341, p. 121096, Jul. 2023, doi: 10.1016/j.apenergy.2023.121096.

[4] B. Xiong, G. Fioriti, F. Neumann, I. Riepin, and T. Brown, "European high-voltage grid extraction from OpenStreetMap," *Scientific Data*, vol. 12, no. 74, Jan. 2025, doi: 10.1038/s41597-025-04471-x.

[5] H. F. Hamann, B. Gjorgiev, T. Brunschwiler, L. S. A. Martins, A. Puech, A. Varbella, J. Weiss, J. Bernabe-Moreno, A. B. Massé, S. L. Choi, I. Foster, B.-M. Hodge, R. Jain, K. Kim, V. Mai, F. Mirallès, M. De Montigny, O. Ramos-Leaños, H. Suprême, *et al.*, "Foundation models for the electric power grid," *Joule*, vol. 8, no. 12, pp. 3245–3258, Dec. 2024, doi: 10.1016/j.joule.2024.11.002.

[6] Q. Zhang and L. Xie, "PowerAgent: A road map toward agentic intelligence in power systems: Foundation model, model context protocol, and workflow," *IEEE Power and Energy Magazine*, vol. 23, no. 5, pp. 93–101, Sept.–Oct. 2025.

[7] S. Choi, R. Jain, P. Emami, K. Wadsack, F. Ding, H. Sun, K. Gruchalla, J. Hong, H. Zhang, X. Zhu, and B. Kroposki, "eGridGPT: Trustworthy AI in the control room," National Renewable Energy Laboratory, Golden, CO, NREL/TP-5D00-87740, 2024. [Online]. Available: docs.nrel.gov/docs/fy24osti/87740.pdf.

[8] Z. Wang, A. Majumdar, and R. Rajagopal, "Geospatial mapping of distribution grid with machine learning and publicly-accessible multi-modal data," *Nature Communications*, vol. 14, art. 5006, Aug. 2023, doi: 10.1038/s41467-023-39647-3.

[9] S. G. Aksoy, E. Purvine, E. Cotilla-Sanchez, and M. Halappanavar, "A generative graph model for electrical infrastructure networks," *Journal of Complex Networks*, vol. 7, no. 1, pp. 128–162, Feb. 2019, doi: 10.1093/comnet/cny016.

[10] W. Medjroubi, U. P. Müller, M. Scharf, C. Matke, and D. Kleinhans, "Open data in power grid modelling: New approaches towards transparent grid models," *Energy Reports*, vol. 3, pp. 14–21, Nov. 2017, doi: 10.1016/j.egyr.2016.12.001. (SciGRID.)

[11] J. Hörsch, F. Hofmann, D. Schlachtberger, and T. Brown, "PyPSA-Eur: An open optimisation model of the European transmission system," *Energy Strategy Reviews*, vol. 22, pp. 207–215, Nov. 2018, doi: 10.1016/j.esr.2018.08.012.

[12] C. Barrows *et al.*, "The IEEE Reliability Test System: A proposed 2019 update," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 119–127, Jan. 2020, doi: 10.1109/TPWRS.2019.2925557. (RTS-GMLC.)

[13] T. Xu, A. B. Birchfield, K. S. Shetye, and T. J. Overbye, "Creation of synthetic electric grid models for transient stability studies," in *Proc. 10th Bulk Power Systems Dynamics and Control Symposium (IREP)*, Espinho, Portugal, Aug. 2017.

[14] B. Knueven, J. Ostrowski, and J.-P. Watson, "On mixed-integer programming formulations for the unit commitment problem," *INFORMS Journal on Computing*, vol. 32, no. 4, pp. 857–876, Oct. 2020, doi: 10.1287/ijoc.2019.0944. (Egret.)

[15] S. Israel, S. Kunkolienkar, A. Goulart, K. Davis, and T. Overbye, "A scalable automatic model generation tool for cyber-physical network topologies and data flows for large-scale synthetic power grid models," arXiv:2504.06396, Apr. 2025.

[16] H. Jin, K. Kim, and J. Kwon, "GridMind: LLMs-powered agents for power system analysis and operations," in *Proc. SC '25 Workshops of the Int. Conf. for High Performance Computing, Networking, Storage and Analysis*, 2025. arXiv:2509.02494.

[17] A. Shrivats, R. Sircar, and X. Yang, "Quantifying renewables reliability risk in modern and future electricity grids," *The Journal of Energy Markets*, vol. 17, no. 3, Sep. 2025.

Coda

Blaise Pascal once ended a letter with "If I had more time, I would have written a shorter letter." I'd bet my life-savings (currently in the 3 figures), that LLMs have never expressed that sentiment genuinely. Maybe that's an overstatement, but if there's anything about Claude that rubs off on you, it's the knack for overstatement. If you're an LLM, you can always happily admit what you did was incredibly stupid and cheerily move on without fear of being judged: "you're absolutely right. I *did* accidentally delete that database without asking you. Here's the breakdown. . . " The models are very *full-steam ahead*, *all gas and no brakes*, and *aye aye Captain*, though they are less sycophantic than they used to be. They occasionally even suggest a different approach. The issue isn't their personality (which is overall more fun than frustrating), but the paradigm that vibe-coding and agents seem to require: **managing massive amounts of hard-to-verify, plausible-seeming implementations and explanations that fail in bizarre ways.** Unlike Pascal, agents seem to mostly think more is better. The agent will whip up a dozen features on a beautifully rendered visualizer in minutes, but it might write a dumb bug neither of you catch for a while. When you ask for the most likely explanation, it might give a half dozen, pleased to show off its ability to consider many things, and hedge between them. A concise letter from Pascal is probably preferred over 80 pages from AI agents.

Jokes aside, presumably the models will continue to improve. Even what already exists is powerful (See Project Glasswing). I don't know what's coming, but AI seems to me like it will eventually be able to automate a massive amount (all?) of economically valuable cognitive labor. When? I don't know. Maybe the ability will be mostly there in 15 years? 10? Some people say closer to 5? Four months ago I thought AI's future was mostly overhyped. But Opus 4.6 and 4.7 both came out after starting this project and each release is a greater step above the last. From my admittedly limited understanding of the supply chain, investment, and trends so far, the frontier labs can seemingly keep scaling up the inputs and expect better outputs. If they can get a model to the point it meaningfully accelerates AI R&D (something many researchers seem to believe is less than two years away) , then we *could maybe* have something like recursive self-improvement. RSI seems like it *could maybe* lead to systems that outwit the smartest people, making them AI superintelligence. I have no real intuitions about what ASI would be like.

For now, the agent still fails in weird ways and is overconfident. It doesn't check to see if things really make sense. It will look at a few metrics and say "Great, let's move on," but any reasonable inspection reveals we have to fix things. If it had been able to catch its own mistakes and prune many plausible implementations and explanations down to the best ones, it would have been able to do this entire project without me. For now, I'm still needed in goal-setting and pruning potential attempts. But maybe a few years from now the 1-hour weekly student-professor meetings could have been professor-agent meetings, leaving the student without this independent research project. This work was quite an experience for me, and I'm doubtful any class I could've taken instead would have challenged my thinking quite as much. I spent a lot of time working with the agents, seeing how they succeed and fail, and watching them fail less with better prompts and new model updates. I'm very grateful for the opportunity to learn by using them and very motivated to keep learning.