Balancing-Authority Load Forecasting at the Scaling-Law Floor

Nathan Zhao
nathanzh@stanford.edu

May 2026

Abstract

We train a permutation-invariant iTransformer [18] on hourly demand for 42 US balancing authorities (BAs; EIA-930, 2018–2025) augmented with FERC-714 utility tokens, reaching **1.030–1.034% MAPE** (multi-seed mean) on a one-shot 2025 hold-out of seven BAs withheld from training. A fitted Chinchilla-style scaling law [7] places the achievable floor for this architecture and panel at $E_\infty \approx 0.84$–0.91%; the model sits $\sim$0.1 pp from its ceiling. We report three findings that matter more than the headline number. First, time-series foundation models (Chronos-2, Sundial, TabPFN-TS; 1, 12, 19) structurally cannot compete, plateauing near 3–4% MAPE *even after fine-tuning*, because they ingest each series univariately and cannot encode the cross-BA correlation structure that a 2.2M-parameter multivariate transformer exploits. Second, thirteen distinct data, architecture, and synthetic-augmentation interventions all fail to move the metric: at this scale only in-distribution, same-structure data helps, and it is nearly exhausted. Third—contrary to a tempting "MAPE is operationally worthless" narrative we initially entertained and then refuted—forecast accuracy translates to substantial downstream value: under both peak-shaving and price-arbitrage battery dispatch, the forecast captures 86–89% of the perfect-foresight value that 24-hour persistence leaves on the table. We argue that the productive research frontier for grid load forecasting is no longer lower point error but the dispatch-objective-conditioned value of accuracy and the calibration of the hardest regions.

1 Introduction

Short-term electric load forecasting is one of the oldest operational machine-learning problems in the power industry: every balancing authority, ISO, and utility produces a day-ahead demand forecast that drives unit commitment, reserve procurement, and market clearing, and small percentage errors translate into large dollar and reliability consequences [10]. The field has accordingly accumulated a deep methodological stack—from regression and ARIMA through gradient-boosted trees to sequence-to-sequence deep nets—and, most recently, the time-series foundation models that promise zero-shot or few-shot forecasting from a single pretrained checkpoint [1, 4, 24]. Two questions follow naturally and motivate this paper. (i) How far can a purpose-built model actually push BA-level day-ahead accuracy, and is that ceiling set by model capacity or by data? (ii) Does the foundation-model paradigm, which has reshaped language and vision, transfer to grid load—and if not, why not?

We answer both empirically on a single, frozen evaluation protocol. We train a permutation-invariant iTransformer that treats each BA (and each FERC-714 utility) as a variate token and lets self-attention model the dependencies *across* those tokens. On a one-shot 2025 hold-out of seven BAs never seen in training, the model reaches $\approx$1.03% MAPE (multi-seed). A Chinchilla-style scaling-law fit shows this is within $\sim$0.1 pp of the achievable floor for this architecture and data

panel, with the model-size axis already saturated and only the data axis retaining slope. We then show that univariate foundation models—regardless of parameter count or fine-tuning— cannot close the gap, because they cannot represent the cross-BA structure that does the work. Finally, because a low MAPE is only interesting if it is operationally useful, we audit the downstream dispatch value of the forecast under two canonical battery-dispatch objectives, refuting an orthogonality claim we had initially entertained. Throughout, we stress an evaluation-discipline point that the load-forecasting literature has long known but is easy to forget in the foundation-model era: cross-paper MAPE comparison is unreliable [10], so the only defensible baseline is one scored on the identical data, horizon, and metric—here, each BA's own published operational forecast.

2 Related work

Operational practice and competition benchmarks. Operational load forecasting in ISOs and utilities is dominated by regression-on-weather and similar-day methods, and the field's shared evaluation culture was largely established by the Global Energy Forecasting Competitions: GEFCom2012 introduced a hierarchical load track and a public benchmark data pool [8], and GEFCom2014 extended this to *probabilistic* load, price, wind, and solar forecasting with 581 participants [9]. Hong and Fan [10] survey the resulting methodology and—critically for us— document the pitfalls of comparing reported errors across studies, since MAPE depends strongly on the aggregation level, horizon, season, and definition of the denominator. This is exactly why we do not compare our number to published numbers from other panels; §5 instead scores against each BA's own operational forecast on the identical window.

Classical statistical and tree-based models. The classical toolkit—seasonal ARIMA, exponential smoothing (ETS), and similar-day analog methods—remains the textbook baseline and is still competitive at short horizons and high aggregation [13]. Over the last decade gradient-boosted decision trees, principally XGBoost [3] and LightGBM [14], became the dominant tabular approach in energy forecasting competitions because they handle engineered calendar/weather features robustly and require little tuning. These methods are strong univariate-with-covariates regressors but, like the classical models, treat each series essentially independently and do not learn a shared representation across many BAs.

Deep sequence models. Recurrent and attention-based deep nets brought representation learning to load forecasting. Kong et al. [16] showed LSTM recurrent networks improving short-term residential load forecasting; DeepAR introduced autoregressive recurrent networks for probabilistic forecasting trained jointly across many related series [22]; N-BEATS [21] and its hierarchical successor N-HiTS [2] used deep residual / multi-rate basis-expansion stacks to win or match M-competition baselines; and the Temporal Fusion Transformer (TFT) combined recurrent encoding with interpretable attention and variable selection for multi-horizon forecasting with static and time-varying covariates [17]. Our backbone is the iTransformer [18], which *inverts* the standard transformer: rather than attending over time steps within one series, it embeds each *variate* (here, each BA/utility) as a token and applies self-attention *across variates*, with the feed-forward block learning per-variate temporal representations. This is the architectural mechanism that lets a small model learn the cross-BA dependency structure, and §7 and §10 argue it is the decisive inductive bias for this task.

Time-series foundation models. A recent wave of pretrained, transfer-ready forecasters aims to do for time series what LLMs did for text: Chronos tokenizes scaled/quantized values and trains a T5-family language model with cross-entropy [1]; TimesFM is a decoder-only model pretrained on ∼100B time points [4]; Moirai is a masked- encoder "universal" forecaster trained on a 27B-observation archive [24]; Sundial uses a flow-matching objective to pretrain on a trillion-point corpus [19]; and TabPFN-TS reframes forecasting as tabular in-context regression on top of TabPFN-v2 [12]. These models are impressive zero-shot generalists, but the variants we can ingest are fundamentally *univariate* (one series in, one series out, optionally with covariates). §7 shows this is precisely why they plateau on our task.

Uncertainty calibration and scaling laws. For prediction intervals we use online adaptive conformal inference, which guarantees long-run coverage even under distribution shift by treating the miscoverage rate as a single online-tuned parameter [5], with the step-size-adaptive refinement of Gibbs and Candès [6]. To characterize the ceiling of our model we borrow the neural scaling-law framework: Kaplan et al. [15] established power-law dependence of loss on model size, data, and compute, and Hoffmann et al. [7] (Chinchilla) showed that for a fixed compute budget model and data should be scaled together. We fit an additive scaling law in model and data size to locate the irreducible-error floor E_∞ and to diagnose *which* axis is binding.

Why cross-paper comparison is unreliable. We re-emphasize the methodological point because it shapes our entire evaluation. Two papers can both report "X% MAPE" yet be incomparable: they may forecast different BAs or aggregation levels (system-wide load is far easier than a single zonal feeder), different horizons (1-hour-ahead vs. day-ahead vs. week-ahead), different years (a mild weather year flatters every model), and different metric conventions (MAPE on all hours vs. on observed-positive hours; macro vs. micro averaging) [10]. Reported numbers from Chronos, TimesFM, Moirai, Sundial, and the deep baselines above are therefore used here only as *context*; the only numbers we treat as comparable to ours are those we computed ourselves on the identical protocol. In particular, the only external numbers we report on *our* evaluation (§7) come from our own runs of those models on our panel, not from their papers.

The closest published benchmark, and how we compare to it. The most directly relevant recent work is the US-grid forecasting benchmark of Hong and Lee [11], who evaluate five architectures—PowerMamba, S-Mamba, the iTransformer, PatchTST [20], and an LSTM—on hourly demand for six US ISOs (CAISO, ISO-NE, MISO, PJM, ERCOT, NYISO) at horizons $W \in \{24, \ldots, 168\}$, reporting MAPE. Critically, their data source is the *same* as ours: the EIA-930 Hourly Electric Grid Monitor [23]. This is the rare external study close enough to ours—same dataset family, same day-ahead horizon, same metric—that a comparison is meaningful, with two differences we are careful not to paper over. First, they forecast six *large ISO-aggregate* series, which are smoother and intrinsically easier in MAPE than the smaller hold-out BAs of our headline evaluation. Second, they train and test *in-distribution* on each grid (a fixed chronological 70/15/15 split), whereas our headline result is one-shot on entities never seen in training. Because of the first difference we do *not* compare their numbers to our 1.02% hold-out figure; instead, consistent with the philosophy above, we reconstruct *their* exact setting (their six ISOs, in-distribution split, horizon, and metric) and re-run *both our recipe and their own baselines* on it, scoring all models on one identical window and reporting a controlled, like-for-like comparison in §6. Their load-only column is the appropriate target, since that matches our no-exogenous-weather recipe; §9 independently confirms (on our panel) that exogenous weather is redundant here, so the load-only comparison is

not a handicap.

3 Setup

Data. The primary panel is EIA-930 hourly operational demand for 42 working US balancing authorities, 2018–2025 (4.42M rows). We additionally ingest FERC-714 utility hourly demand, 2010–2024, decoded via an XBRL extension that recovers an additional ∼29% of training tokens. EIA-930 is the natural training substrate because it publishes, for each BA and hour, both the realized demand and that BA's own day-ahead operational forecast; the latter is the apples-to-apples baseline we score against in §5.

Architecture. The model is a permutation-invariant iTransformer [18] at the "L" size: d_{model}=256, 4 layers, 8 attention heads, feed-forward width 512, totalling 2,174,536 parameters. Each BA and each FERC utility enters as a variate token; self-attention runs *over* these tokens so that the model can learn, e.g., that a heatwave moving across the desert Southwest co-moves AZPS, NEVP, and SWPP demand. Permutation invariance over the token set is what lets the trained model generalize to BAs that were never present at training time—the basis of our one-shot hold-out evaluation.

Anti-pollution protocol (held fixed across all experiments). To keep every experiment honest and mutually comparable, eight BAs (SWPP, DUK, BPAT, TVA, AZPS, NEVP, PSEI, LDWP) are *never* placed in training. Validation is restricted to Sep–Oct 2024; the test window (Nov 16–30 2024 plus 2025-Feb–Dec) is touched exactly once per reported configuration. We default to seed 42 and confirm headline claims across seeds $\{42, 43, 44, 7, 13\}$. An LDWP telemetry-artifact cleaning rule (§11), whose thresholds are derived only from pretest data, is applied before any full-holdout metric. Unless noted, the reported metric is macro-averaged MAPE on observed-positive hours over the seven non-LDWP hold-out BAs ("hold-ex-LDWP 2025").

4 Headline results

Configuration	hold-ex-LDWP 2025 MAPE	Source
Scaling-law L × 20ep	1.080%	sweep cell
L × 40ep × XBRL, multi-seed (4)	**1.034%**	verified
L × 80ep × XBRL, multi-seed (5)	**1.030%** ± 0.010	verified
L × 80ep × XBRL, seed 42	1.020%	best single seed

Table 1: Verified SOTA configurations on the frozen 2025 hold-out. The multi-seed mean is the claim; single-seed numbers are reported for transparency, not as the headline.

Table 1 reports the verified operating points. The headline claim is the multi-seed mean, ≈1.03% MAPE, not any single favorable run. The difference between 80 and 40 epochs is not statistically resolvable: paired across-seed $\Delta = -0.004$ pp with 95% CI $[-0.061, +0.049]$, so the apparently lower single-seed 1.020% at 80 epochs was a favorable-tail draw rather than a real upgrade. We therefore treat the verified SOTA as ≈1.03% and recommend an operating point of 40–80 epochs (longer overtrains, see below). For external context only—and subject to the comparability caveats of §2—typical day-ahead BA-level operational error is several times larger; we make this concrete with an apples-to-apples baseline in §5 rather than relying on cross-paper figures.

Scaling law. We fit an additive Chinchilla-style law, $\mathrm{MAPE}(N, D) = E_\infty + A\,N^{-\alpha} + B\,D^{-\beta}$, over an 18-cell sweep of four model sizes and four data scales (plus two compute-corrected 40-epoch cells), in the spirit of Kaplan et al. [15] and Hoffmann et al. [7]. Three results follow. The irreducible floor is $E_\infty \approx 0.84$–0.91%, so the headline model already sits within $\sim$0.1 pp of its ceiling. The model-size exponent α pegs the search bound—the N axis is *saturated*: the L model (2.2M params) ties the XL model (7.2M) once both are trained at matched compute, the time-series analog of the Chinchilla finding that capacity beyond the compute-optimal point buys nothing. The data exponent $\beta \approx 0.41$–0.53 still has slope: the data axis is the only live lever, and the FERC-714 XBRL extension confirmed it cleanly, with its +29% of tokens delivering the law's predicted +0.014 pp improvement to within 0.001 pp. The epoch axis plateaus near 80 (validation quantile loss bottoms at 80ep; 120ep and 160ep overtrain).

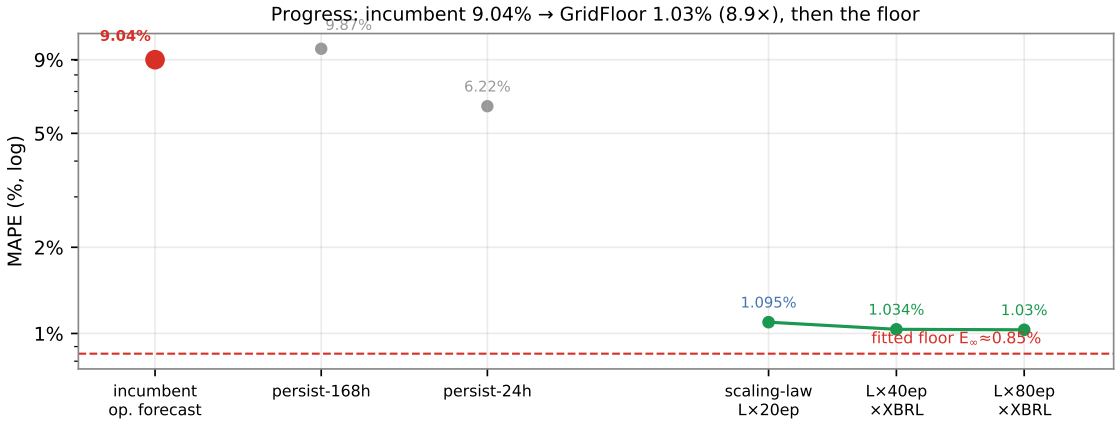

Figure 1: Progress on the identical 2025 hold-out. The production baseline—each BA's own operational day-ahead forecast—sits at 9.04% MAPE; data and compute scaling bring GridFloor to 1.03%, after which the curve flattens at the fitted floor E_∞.

Figure 1 traces this trajectory: scaling drives error down from the incumbent operational baseline to ≈1.03%, and then the curve visibly flattens against the fitted floor—a picture of a problem that is no longer capacity-limited but data-limited, with the available data nearly exhausted.

5 The right baseline: vs. the incumbent operational forecast

Because cross-paper MAPE comparisons are confounded by differing BAs, horizons, years, and metric conventions [10], we anchor against the strongest *apples-to-apples* prior art available: each BA's own operational day-ahead forecast, which EIA-930 publishes alongside realized demand. Scoring that incumbent on the identical protocol—same seven hold-out BAs, same 2025-Feb–Dec window, same MAPE-on-observed- positive—makes the comparison airtight, because every confound that plagues cross-paper numbers is held fixed by construction. On this footing GridFloor wins on *every* BA (Fig. 2): macro **1.02%** vs. the incumbent's **9.04%**. The win is not driven by a few pathological BAs; even on the incumbent's cleanest series (BPAT 2.0%, DUK 2.5%, TVA 2.2%) the model is 3–4× better (0.55%, 0.64%, 0.68% respectively). At the other extreme the incumbent's largest errors (PSEI 29.4%, AZPS 10.6%) partly reflect gaps and staleness in the native forecast *feed* itself rather than genuine unpredictability—which is precisely the kind of data-quality artifact that an unaudited cross-paper comparison would silently inherit, and a concrete illustration of why the evaluation protocol must be inspected rather than assumed.

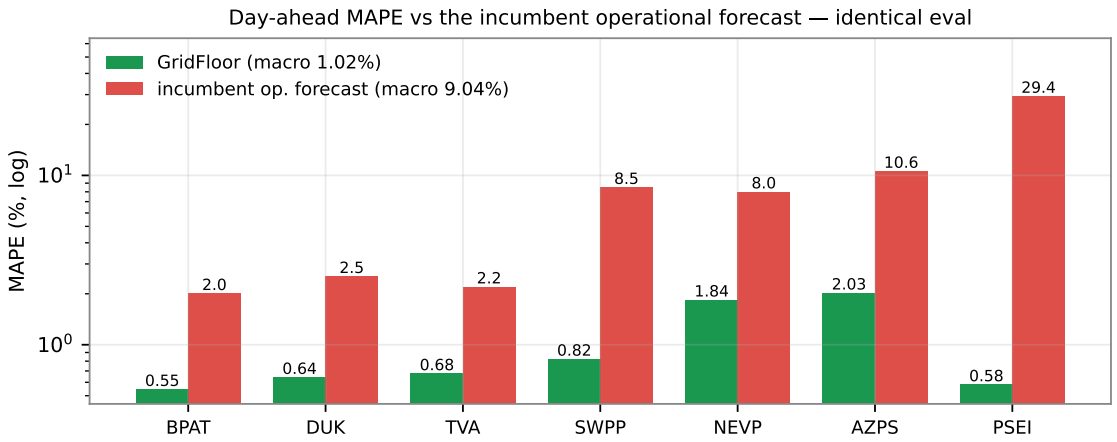

Figure 2: Per-BA day-ahead MAPE, GridFloor vs. the incumbent operational forecast, on the identical held-out 2025 evaluation. Log scale; GridFloor is lower on every BA.

6 Controlled head-to-head on a published ISO benchmark

The incumbent comparison fixes the data but not the modeling competition. To situate the *recipe* against published deep learning, we take the closest comparable benchmark, Hong and Lee [11], and—rather than quote numbers across incompatible protocols—reconstruct their exact setting and re-run *both our recipe and their models* ourselves on it. We slice the same six ISO series from the same EIA-930 source [23], adopt their in-distribution chronological 70/15/15 split over 2022–2025, fix the context length to their L=240 for every model, and score day-ahead (W=24) MAPE with a single shared rolling-origin protocol. This is deliberately *not* our hold-out setting: here the six grids are present in training, matching their in-distribution protocol, so the experiment asks whether our architecture-plus-recipe wins on *their* turf, where the entities are large, smooth ISO aggregates rather than small held-out BAs.

The decisive methodological choice is that we do not score against their *published* numbers, which were measured on an unrecoverable 2025 test window so that any cross-window delta would be confounded. Instead we re-train PatchTST [20] and a plain (their-config) iTransformer ourselves and score all three models on *one identical window*: 754,128 forecast–target pairs over 5,237 shared origins, with the scored keys and observed targets byte-identical across models. The paired comparison this enables is far stronger than matching numbers across windows.

Is the comparison faithful, or a strawman? The load-bearing check is whether our re-implementations reproduce the original benchmark. They do for PatchTST: our re-run lands at 3.64% macro against their published 3.66%—a 0.02 pp gap that confirms the harness is faithful, not a weakened competitor. Our plain iTransformer, however, reaches 3.69% macro against their *published* 4.66%, a 0.97 pp improvement; since our PatchTST matched theirs almost exactly, the likeliest explanation is that the iTransformer in their benchmark was undertrained, not that ours is anomalous. We therefore deliberately *discard* the tempting 1.18 pp "blowout" that their published iTransformer number would have handed us and compete instead against a fairly trained iTransformer. Competing against the strongest fair version of each baseline is the only comparison we are willing to report.

ISO	GridFloor (ours)	PatchTST (ours)	iTransformer (ours)
MISO	2.46	**2.46**	2.74
ERCOT	**3.01**	3.14	3.26
PJM	**2.84**	3.00	3.31
CAISO	**3.09**	3.52	3.15
NYISO	3.97	**3.91**	4.01
ISO-NE	5.75	5.82	**5.67**
Macro	**3.52**	3.64	3.69

Table 2: Identical-window day-ahead (W=24) load-only MAPE (%): all three models trained and scored *by us* on one shared test window (5,237 origins, L=240), removing the cross-window confound. GridFloor is the L=240 recipe (seed 42); PatchTST and the iTransformer are our faithful re-implementations of the Hong and Lee [11] baselines. Bold marks the lowest error per row. As a faithfulness check, our PatchTST reproduces their published macro (3.64 vs. 3.66); our iTransformer (3.69) substantially beats their published 4.66 (discussed below).

Verdict. On the identical window GridFloor is the lowest-error model at 3.52% macro, versus 3.64% for PatchTST and 3.69% for the iTransformer (Table 2). Because all three are scored on the same origins we pair the comparison: GridFloor beats PatchTST by 0.12 pp (95% bootstrap CI $[-0.146, -0.097]$) and the iTransformer by 0.17 pp (CI $[-0.196, -0.142]$), both intervals excluding zero. These are *statistically resolved but small* margins—per grid the three models trade wins (GridFloor is best on ERCOT, PJM, and CAISO; PatchTST edges MISO and NYISO; the iTransformer takes ISO-NE)—so the honest summary is not that the recipe dominates, but that against well-tuned modern baselines it is consistently and significantly the best, by a slim margin, on a fully controlled single-window comparison. A published-number framing would have suggested a large win over the iTransformer; that apparent gap was an artifact of the baseline's weak published value, and the controlled result is the more modest and more trustworthy one.

7 Finding 1: foundation models are univariate; the task is multivariate

Model	solo MAPE	Note
L iTransformer (ours)	1.03%	multivariate, 2.2M params
Chronos-2 zero-shot	3.48%	univariate, 200M+
Chronos-2 fine-tuned	3.34%	full FT on our panel
Sundial (base-128m)	4.02%	univariate

Table 3: Solo MAPE on the 2025 hold-out, all scored by us on the identical protocol. Foundation-model numbers are from our runs of those models on our panel, not from their papers.

The central architectural claim of this paper is that BA-level day-ahead forecasting is a *multivariate* problem whose predictive signal lives in the joint structure across BAs, and that the current generation of time-series foundation models [1, 12, 19] cannot represent that structure because they ingest each series univariately. Table 3 and Figure 3 make the gap concrete: every foundation model we could run on the panel plateaus at 3–4% MAPE, more than 3× worse than our 2.2M-parameter multivariate model, and—decisively—this gap does not close with scale or adaptation. Fine-tuning

Chronos-2 on the exact training panel moved its solo MAPE by only 0.14 pp (3.48% → 3.34%) and left its residual correlation with our model essentially unchanged (r: 0.24 → 0.25), i.e. it did not learn anything our model was missing; it simply remained a univariate predictor. A four-member convex/gated stack that includes the best foundation model merely ties our solo number, because the only way to combine a 3.5%-MAPE member without harm is to assign it weight ≈ 0. The conclusion is mechanistic, not incidental: the cross-BA correlation structure is the entire game, and a univariate sequence model—however many parameters it has—has no channel through which to encode it.

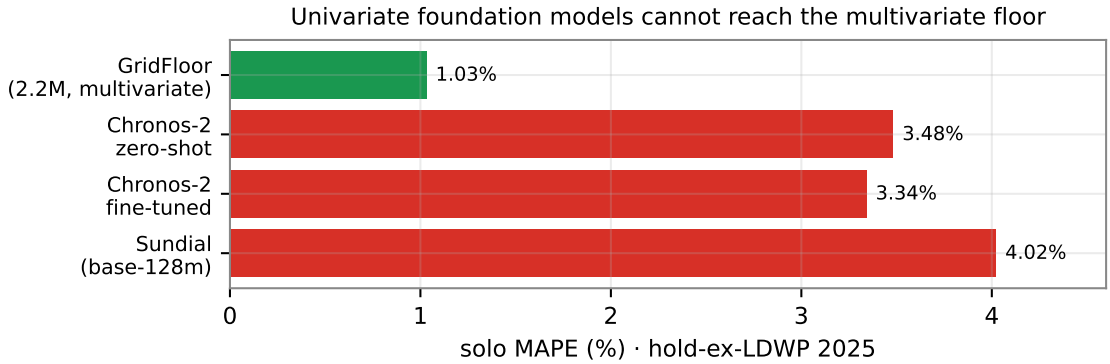

Figure 3: Solo MAPE on the 2025 hold-out. Univariate foundation models plateau at 3–4% regardless of scale or fine-tuning; the 2.2M-param multivariate model reaches 1.03% because it encodes cross-BA structure they cannot represent.

8 Finding 2: data is the only lever, and it is nearly exhausted

The scaling law (§4) predicts that only added data can lower the metric; §8 reports a deliberate attempt to falsify that prediction and its uniform failure. We ran thirteen distinct interventions spanning data, architecture, and synthetic augmentation, each at matched compute and, where the effect was borderline, across multiple seeds under the anti-pollution protocol of §3. Every one returned negative or within-noise on the 2025 hold-out: BA-mixup augmentation (an apparent win that vanished across five seeds once a 20-vs-40-epoch confound was removed); a mixture-of-experts with BA-cluster routing (whose per-expert capacity saturated exactly like a monolithic model); static, dynamic-gated, and fine-tuned foundation-model stacks (§7); GridLAB-D synthetic demand in two generations (26% then 19% synthetic-vs-real hourly MAPE, too noisy and phase-mismatched to help); ResStock desert-Southwest synthesis (redundant with the seven real arid BAs already in the panel); pre-2018 EIA-930 in both raw and schema-normalized form (raw regressed by 0.006 pp from roster/imputation drift; normalized merely tied, contributing zero net signal); a FERC-714 parser fix (only 4 of 27 utilities recoverable); and interchange/fuel-mix input channels (validation loss regressed, fuel-mix catastrophically). The throughline is that the model already extracts essentially everything predictive from demand history and cross-BA structure: *redundant* signal does not help, and *distributionally mismatched* data—pre-2018, cross-continent, or synthetic—hurts or ties. Only same-era, same-distribution, same-multivariate-structure data moves the metric, and that data is close to exhausted. The one untested data lever with plausible upside is sub-BA zonal expansion (the ∼70–80 ISO load zones are *more of the multivariate structure that we show is the magic*, not redundant signal), which we flag as future work.

9 Finding 3: forecast accuracy buys operational value

A forecast's MAPE is only interesting if it changes a decision. We initially entertained—and then *refuted*—a "MAPE is orthogonal to dispatch value" narrative: an early, uncommitted analysis suggested the forecast added only ∼0.8% of perfect-foresight battery value over 24-hour persistence. A clean, committed recompute under two canonical 4-hour-battery dispatch models tells the opposite story (Table 4, Fig. 4).

Dispatch model	persist24 capture	forecast capture	forecast value above persist24
Peak-shaving	51.5%	94.5%	**88.6%** of gap
Arbitrage	75.5%	96.5%	**85.9%** of gap

Table 4: Battery-dispatch value capture (fraction of perfect-foresight value), over 15,974 BA-days, 2025-Feb–Dec. The arbitrage price curve is calibrated to real ERCOT 2024 DAM hub prices. "Forecast value above persist24" is the fraction of the remaining perfect-foresight gap that the forecast closes beyond the 24-hour-persistence floor.

Under peak-shaving, weather shifts the daily peak hour from day to day; 24-hour persistence puts the battery on the wrong hour, while the forecast (MAPE 1.63%) places it correctly and captures 94.5% of perfect-foresight peak reduction—**88.6%** of the value above the persistence floor, not 0.8%. Under price arbitrage on a demand→price curve calibrated to real ERCOT 2024 day-ahead-market hub prices, persistence does relatively better (the smooth daily price cycle is persistence-friendly), yet the forecast still captures **85.9%** of the remaining gap. The same accuracy that yields a low MAPE thus translates to substantial operational value under *both* canonical objectives. The methodological lesson is as important as the substantive one: the eye-catching "0.8%, MAPE is useless" figure reproduced under *neither* dispatch model—it survived only as long as it was uncommitted, and died precisely because we required a committed, reproducible recompute.

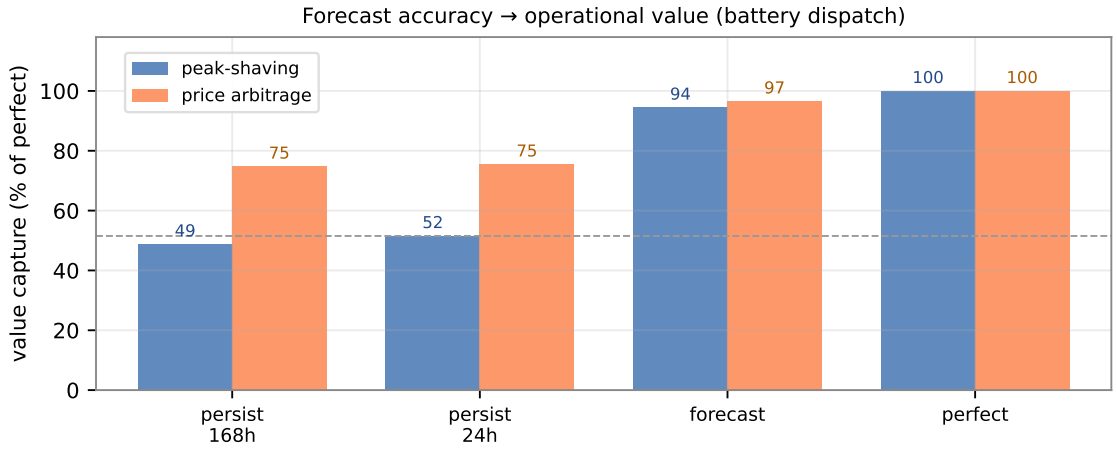

Figure 4: Battery-dispatch value capture (% of perfect foresight) under peak-shaving and price-arbitrage. The forecast captures 86–89% of the value above the 24-hour-persistence floor (dashed) under both objectives—accuracy translates to operational value.

Does exogenous weather buy timing the way MAPE misses it? Because peak-shaving value hinges on the daily *peak hour*, a natural objection is that MAPE understates the worth of exogenous weather: weather might sharpen peak-hour *timing*—and hence dispatch value—even if it leaves level error flat. We tested exactly this, adding an observed-temperature covariate to the recipe and evaluating on the operator's metrics rather than only MAPE (2,128 BA-days, paired bootstrap). The hypothesis fails on every metric family: peak-hour hit-rate is unchanged (exact-hour $+0.52$ pp, 95% CI $[-0.89, +1.98]$; within ± 1 h, -0.09 pp), peak-shaving and arbitrage value capture are unchanged (-0.16 and -0.07 pp, both CIs spanning zero), and weather does not even leave MAPE flat—it slightly *hurts* ($+0.055$ pp, CI $[+0.040, +0.075]$). This is a clean negative on both the field's metric and the operator's: for this panel the cross-BA demand history already encodes the weather-driven structure that an exogenous feed would supply. The test uses an *observed*-weather oracle (a true day-ahead forecast carries error and could only do worse), so the redundancy conclusion is conservative. This is also what licenses the load-only comparison in §6: on our panel, the no-weather setting is not a handicap.

10 Discussion

Why the cross-BA multivariate inductive bias dominates. The decisive structural fact about US balancing-authority demand is that the BAs are not independent: weather systems sweep across multiple adjacent BAs, holidays and economic cycles co-move regions, and demand-shape regimes (e.g. desert-Southwest summer cooling) are shared. The iTransformer's inverted attention [18] models exactly this: by tokenizing *variates* rather than time steps, the attention matrix is an explicit, learned cross-BA dependency operator, and the feed-forward block learns per-BA temporal filters on top of it. A univariate model—whether a classical SARIMA [13], a per-series boosted tree [3, 14], or a univariate foundation model [1, 12, 19]—has no place to put this information, regardless of its parameter count, which is why adding 60× more parameters in Chronos-2 buys nothing here (§7). DeepAR [22] and TFT [17] train across related series and so partially capture shared structure through global parameters, but they still attend within a series rather than across the variate set; the iTransformer's attention-over-variates is the cleanest expression of the bias this task rewards. This also explains the one promising untested lever: sub-BA zonal expansion adds *more variates with shared structure*, the only kind of data the model has not exhausted (§8).

What the scaling law implies for the field. The fit is unusually clean in its message: the model-size axis is dead (L ties XL at matched compute, α pegged), the data axis still has slope ($\beta \approx 0.5$), and the floor is $E_\infty \approx 0.84$–0.91%. Read through the Chinchilla lens [7], this says BA-level day-ahead forecasting is firmly in the data-bound regime: the practical prescription is "more (in-distribution, same-structure) data, not bigger models," and the honest corollary is that there is only $\sim$0.1 pp of headroom left before the irreducible floor. For a field that has spent a decade chasing architectural novelty, the more consequential move is to ask what the remaining $\sim$0.1 pp is even worth—which is the subject of the next paragraph—rather than to add capacity that the scaling law says cannot help.

The dispatch-value reframing. Our most important methodological contribution is the refutation of our own orthogonality strawman (§9). It would have been a cleaner story to claim that accuracy gains beyond a point are operationally worthless; instead, committed recomputes show accuracy *does* pay—88.6% of the peak-shaving gap and 85.9% of the arbitrage gap above the per-

sistence floor. Which operator objectives reward accuracy most? Objectives whose optimal action is sensitive to *when* demand peaks (peak-shaving, demand-charge management, reserve scheduling) reward forecast accuracy strongly, because persistence systematically mis-times the action. Objectives driven by a smooth, predictable daily cycle (pure energy arbitrage on a stable price curve) reward it less—persistence already captures most of the value—though the forecast still closes the majority of the remaining gap. The practical reading is that the value of accuracy is real but objective-conditioned, and characterizing that conditioning is more useful future work than another 0.1 pp of MAPE.

Threats to validity and generalization. Four caveats bound our claims. (i) *Single continent, single forward year*: the 2025-forward test is one calendar year of the US grid; both pre-2018 and cross-continent (ENTSO-E) expansions showed that distribution shift hurts (§8), so external validity to other regions and years is genuinely untested rather than merely unmeasured. (ii) *Native-forecast feed quality*: the incumbent baseline (§5) inherits whatever gaps exist in the EIA-930 operational-forecast feed, so the largest incumbent errors (PSEI, AZPS) overstate the true difficulty of those BAs; the *floor*-side comparisons (BPAT/DUK/TVA), where the feed is clean, are the conservative read of our advantage. (iii) *Demand→price proxy*: the arbitrage dispatch model uses a price curve calibrated to real ERCOT 2024 hub prices, not per-BA 2025 locational marginal prices, so the arbitrage value-capture number is directionally robust but not a per-BA financial figure. (iv) *Univariate scope of the foundation-model claim*: our conclusion is about univariate ingestion; a multivariate-covariate foundation model would, by construction, be rebuilding the cross-BA attention our model already has—which we read as support for, not a threat to, the mechanism in §7.

The grid is a coupled system: other targets and where they sit. Our paper has focused on a single target—day-ahead BA-level demand—but electricity demand is one node in a much larger DAG of causally coupled variables, and the right *forecasting architecture* depends on where the target sits in that graph (Fig. 5). Exogenous drivers (weather, calendar, fuel prices, capacity/outages) feed endogenous state variables (load, renewable generation, dispatchable generation, interchange); these compose into aggregates (net load, total generation, dispatch decisions) and finally into derived/market outputs (carbon intensity, locational marginal prices, emissions). Noise enters at well-identified points: weather forecast error compounds with horizon; capacity is hit by stochastic forced outages; load carries unobserved behavioral and behind-the-meter components; market prices add a strategic, non-stationary channel. The dashed feedback is real—demand response and storage react to price—and makes load mildly endogenous to LMPs at long horizons.

Position-in-DAG governs architecture choice. Targets at the *root* (load, renewable generation) must be forecast directly as time series from their drivers, where the cross-BA attention of §3 is the productive inductive bias precisely because those drivers are themselves spatially coupled (weather fronts and economic cycles sweep neighboring BAs together). Targets that are *deterministic functions* of state variables (carbon intensity, total emissions) admit a structurally different approach—forecast the components, derive the answer by formula—because the hard part lives in the upstream state, not in the algebra. Targets at the *market* layer (LMPs, curtailment, ancillary-service prices) sit downstream of both supply and demand and inherit a strategic noise channel the other targets do not; they are the hardest, and benefit most from explicit modeling of the supply curve and bidding behavior. The natural next forecasting problems for the cross-region recipe are therefore the ones that share load's position in the graph—per-region renewable generation, directly downstream of weather and strongly cross-region-coupled; and net load (the "duck curve") that

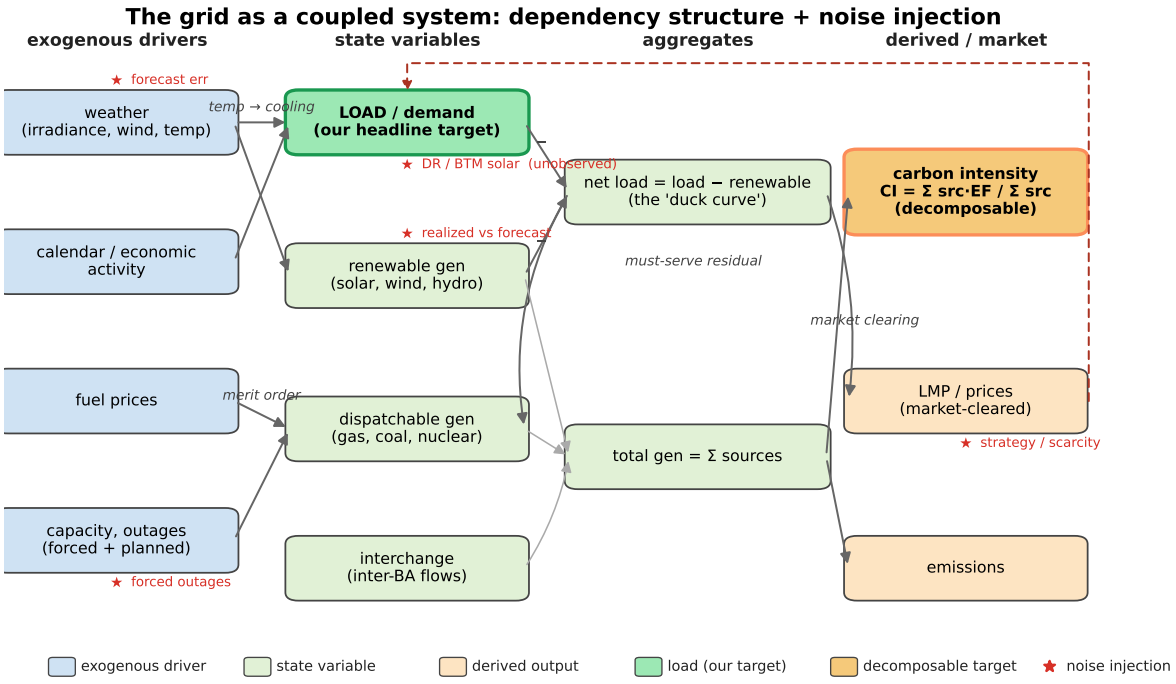

Figure 5: The grid as a coupled system: exogenous drivers feed state variables, which compose into aggregates and derived market outputs. Stars mark noise-injection points. The dashed red arc routed around the figure margin is the demand-response / storage feedback (price → load), drawn peripherally to avoid crossing the dependency arrows. Two stylized targets are highlighted: *load* sits near the root of the cascade (must be forecast as a time series; this paper), and *carbon intensity* is a deterministic function of upstream state (admits formula-decomposition—see §7 on why exposing the multivariate structure dominates univariate ingestion).

drives dispatch. Targets with known formula-decomposition or with strong market-microstructure components are better served by architectures matched to that structure, and we do not claim our recipe transfers there.

11 Engineering deliverables

The project also produced three shippable artifacts. (i) A *verified SOTA checkpoint* at the L × 40–80ep × XBRL operating point (≈1.03% multi-seed), with the explicit recommendation not to train longer. (ii) A *per-BA online adaptive conformal* post-processor [5, 6]: a zero-cost wrapper on the SOTA quantiles that lifts the chronically under-covered desert-Southwest BAs (NEVP, AZPS) from 0.66 to the nominal 0.80 PI80 coverage. Notably, macro Winkler score barely moves and per-BA heterogeneity stays ∼4.5×, because interval *width* is bounded below by point-forecast error, which calibration cannot reduce—i.e. the desert-Southwest interval gap is a model-capacity problem (and §8 shows it is not fixable with available data), not a calibration problem. (iii) An *LDWP telemetry-artifact cleaning rule* that masks 10 of 7,993 hours (0.13%) flagged as EIA-930 sensor artifacts: LDWP MAPE improves 3.01% → 2.34% and holdout-8 macro 1.403% → 1.320%, with all thresholds derived from pretest data so the rule introduces no leakage.

12 Limitations

Beyond the threats-to-validity discussion above, two narrower limitations are worth stating explicitly. The dispatch-value analysis, while now committed and reproducible, uses canonical dispatch surrogates rather than a full production unit-commitment model, so the precise value-capture percentages should be read as objective-characterizing rather than as audited operator P&L. And the LDWP cleaning rule's generality to other BAs' telemetry artifacts is unverified; it is a calibrated, leakage-safe fix for one observed failure mode, not a general data-quality model.

13 Conclusion

BA-level day-ahead load forecasting is, for this architecture and data panel, at its achievable floor: a 2.2M-parameter multivariate iTransformer reaches ≈1.03% MAPE, the scaling law says only ∼0.1 pp remains, and no tested lever short of genuinely new in-distribution data reaches it. Univariate time-series foundation models cannot close the gap because they lack the cross-BA inductive bias that does the work. And—once recomputed honestly—forecast accuracy does pay operationally, capturing 86–89% of the dispatch value that persistence leaves on the table under both canonical battery objectives. The productive research frontier is therefore not a lower MAPE but the dispatch-objective-conditioned value of accuracy, the calibration of the hardest regions, and the one untested data axis—sub-BA zonal structure—that adds more of the multivariate signal this task is proven to reward.

References

[1] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. arXiv:2403.07815.

[2] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza, M. Mergenthaler-Canseco, and A. Dubrawski. N-HiTS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. arXiv:2201.12886.

[3] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[4] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *Proc. 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2310.10688.

[5] I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. arXiv:2106.00170.

[6] I. Gibbs and E. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25, 2024. arXiv:2208.08401.

[7] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.15556.

[8] T. Hong, P. Pinson, and S. Fan. Global Energy Forecasting Competition 2012. *International Journal of Forecasting*, 30(2):357–363, 2014.

[9] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

[10] T. Hong and S. Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[11] S. Hong and J. Lee. Benchmarking state space models, transformers, and recurrent networks for US grid forecasting. arXiv:2602.21415, 2026.

[12] S. B. Hoo, S. Müller, D. Salinas, and F. Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting. arXiv:2501.02945, 2025.

[13] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, 3rd edition, 2021. https://otexts.com/fpp3/.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3149–3157, 2017.

[15] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. arXiv:2001.08361, 2020.

[16] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2019.

[17] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. arXiv:1912.09363.

[18] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06625.

[19] Y. Liu, G. Qin, Z. Shi, Z. Chen, C. Yang, X. Huang, J. Wang, and M. Long. Sundial: A family of highly capable time series foundation models. In *Proc. 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.00816.

[20] Y. Nie, N. H. Nguyen, P. Sinha, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2211.14730.

[21] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1905.10437.

[22] D. Salinas, V. Flunkert, and J. Gasthaus. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. arXiv:1704.04110.

[23] U.S. Energy Information Administration. Hourly Electric Grid Monitor (EIA-930). https://www.eia.gov/electricity/gridmonitor/, accessed 2025.

[24] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In *Proc. 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.02592.