

Are Day-Ahead Electricity Demand Forecasts Less Accurate at Peak? Evidence from Eight U.S. Balancing Authorities

Jose Aleman

University of Maryland, College Park

This version: July 31, 2026

Abstract

Grid operators schedule generation against a day-ahead forecast of hourly demand, and a forecast miss is most expensive in the hours when the system has the least spare capacity. This paper asks whether day-ahead forecast error is in fact larger in those hours. Using twenty-four months of hourly EIA-930 filings from the eight balancing authorities that serve most of the contiguous United States (139,055 region-hours after filtering), I define a stress hour as one in the top five percent of a region's own demand within season, and measure the stress penalty as the difference in mean absolute percentage error between stress and normal hours, with block-bootstrap confidence intervals that resample whole days. Pooled across regions the penalty is -0.05 percentage points: forecasts are no worse at peak. One region (New York ISO) is measurably less accurate at peak; two (ISO New England and PJM) are measurably more accurate; five are indistinguishable from zero. The result is not robust to the definition of stress: under a net-load ramp definition, PJM reverses sign and Southern Company becomes the largest penalty. The more consistent defect in these forecasts is directional bias, not peak-hour failure.

JEL classification: Q41, Q47, L94, C53

Keywords: electricity demand; day-ahead forecasting; balancing authorities; EIA-930; block bootstrap; net load; forecast bias

1 Introduction

Electricity is generated and consumed in the same instant, and the grid stores almost none of it. Every balancing authority in the United States therefore publishes, one day in advance, an hour-by-hour forecast of the next day's demand, and commits generating units against it. When the forecast is too low the shortfall is bought in the real-time market, typically above the day-ahead price; when it is too high, units are started whose fuel and operating costs are incurred regardless.

The asymmetry that motivates this paper is that the cost of a given percentage miss is not constant across hours. In the highest-demand hours of the year the system has the least headroom, the marginal unit is the most expensive one available, and real-time prices are at their most volatile. A reliability planner reading a single system-wide accuracy figure would like to know whether that figure is representative of the hours that matter, or whether it averages good performance in quiet hours together with poor performance at peak.

I test that proposition directly. The question is narrow and stated in advance: when a balancing authority is in the top five percent of its own demand, is its day-ahead forecast error higher than in its other hours, and by how much? The answer, pooled across the eight regions studied, is no — 3.88 percent mean absolute percentage error in stress hours against 3.93 percent elsewhere. The regional picture is more mixed, and the sign of the effect turns out to depend on how stress is defined, which is the more useful finding.

The paper contributes in three places. First, it separates the point estimate from the interval throughout: five of eight regions have penalties that a block bootstrap cannot distinguish from zero, and a ranking built on point estimates alone reorders the regions. Second, it shows that the conclusion is definition-dependent by re-running the entire analysis with stress defined by the speed of net-load change rather than the level of demand; the two definitions share only 5 to 12 percent of their flagged hours and PJM switches sign between them. Third, it documents two data-quality failures in the federal source that materially affect the headline numbers, and reports them rather than silently correcting them.

2 Institutional background

The entities that balance supply and demand on the U.S. grid are called balancing authorities. This study covers the eight largest by load: PJM Interconnection (mid-Atlantic), Midcontinent ISO, California ISO, ERCOT (Texas), ISO New England, New York ISO, Southwest Power Pool, and Southern Company. Together they serve the majority of load in the contiguous United States.

Each files Form EIA-930 with the U.S. Energy Information Administration on an hourly basis. The filing reports actual demand, the day-ahead forecast that was published for that hour, net generation, interchange with neighboring regions, and generation by fuel type. Because the forecast is recorded alongside the realization for the same hour, EIA-930 supports a clean ex post accuracy measurement without any reconstruction of what the operator knew at the time.

A second concept matters for the robustness analysis. Net load is demand minus wind and solar output. Wind and solar cannot be dispatched upward on request, so net load is the portion of demand that dispatchable plant must cover. A system can be strained because demand is high, or because net load is changing quickly — the steep evening ramp in California, known as the duck curve, is the canonical case. These two kinds of strain occur in different hours, and Section 5.4 measures both.

3 Data

3.1 Source and coverage

The data are the raw EIA-930 series retrieved through the EIA API (v2), covering twenty-four months from September 2024 through September 2026 at hourly grain. The pull returns 559 thousand region-hour rows and 1.23 million fuel-hour rows, 1.79 million in total. After the filters described below, 139,055 balancing-authority-hours enter the analysis. The top-five-percent threshold yields 6,961 stress hours across the eight regions, and a check confirms that 5.01 percent of hours are flagged, which is what a correctly implemented within-region, within-season percentile should produce.

I use EIA's raw series rather than its adjusted series. The adjusted series smooths reporting errors, which would suppress exactly the data-quality problems documented in Section 6; the cost is that those errors remain in the headline numbers, and I flag them rather than correcting them.

3.2 Validation

A validation stage checks for missing hours, duplicate timestamps, impossible values, outliers, and an accounting identity: net generation minus interchange should equal demand. Two failures survive into the results and are reported in Section 6. The validation step writes a machine-generated data-quality report and

exits non-zero on a hard failure, so a broken pull cannot propagate silently into the warehouse.

4 Measurement

4.1 Accuracy

Forecast accuracy is measured as the mean absolute percentage error (MAPE) of the day-ahead forecast against realized demand, computed within region and hour class. MAPE is scale-free, which is necessary here because the regions differ by an order of magnitude in load, and it is the measure operators themselves report. Its known weakness — asymmetry between over- and under-forecasting — is addressed separately in Section 5.5, where the signed error is examined on its own.

4.2 Defining stress

A stress hour is an hour in the top five percent of demand for that balancing authority within the same season. Both conditioning choices are deliberate. Using each region's own percentile prevents a large region from being compared with a small one on raw megawatts. Conditioning on season treats summer-peaking and winter-peaking regions symmetrically. The top five percent was used rather than the top one percent so that every region retains enough stress hours for a stable regional average; the cost is that the definition reaches somewhat below the genuinely extreme hours.

The stress penalty is MAPE in stress hours minus MAPE in normal hours, expressed in percentage points. A positive penalty means the forecast is less accurate when demand is highest.

4.3 Inference

Confidence intervals come from a block bootstrap that resamples whole days rather than individual hours, run at 2,000 iterations per region. The block structure is not cosmetic: forecast errors within a day are driven by a shared weather forecast and are strongly correlated, so an hourly resample would understate the sampling variance and make every regional verdict look more certain than it is. An independent re-run at 4,000 iterations with a different seed moves no interval bound by more than 0.035 percentage points and changes no regional verdict.

5 Results

5.1 The aggregate result

Pooled across all eight regions, mean forecast error is 3.88 percent in stress hours and 3.93 percent in normal hours, a difference of -0.05 percentage points. There is no aggregate peak-hour degradation to explain. Whatever is true region by region does not add up to a system-wide effect.

Forecast error by region, normal hours and stress hours
Stress-hour bar colors: red = less accurate at peak, blue = more accurate, light gray = no measurable difference

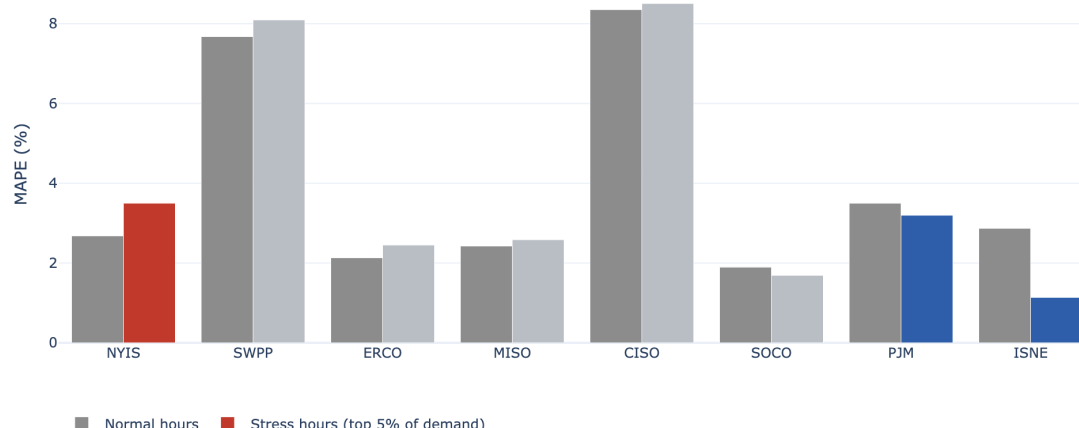


Figure 1.1. Forecast error in normal and stress hours, by balancing authority

Notes: Gray bars are mean absolute percentage error in normal hours, colored bars in stress hours. If forecasts degraded at peak everywhere, the colored bar would exceed the gray bar in every panel. It does so in one region.

5.2 Regional heterogeneity

Table 1.1 reports the region-level decomposition. Three of eight regions have a stress penalty whose 95 percent interval excludes zero. New York ISO is less accurate at peak by 0.81 percentage points; during its stress hours it imports an average of 2,399 MW from neighboring systems, which is consistent with those hours being the ones in which its own supply position is tightest. ISO New England and PJM run in the opposite direction, more accurate at peak by 1.73 and 0.31 points respectively. ISO New England's stress-hour error is roughly 0.40 times its normal-hour error, the largest improvement in the sample.

Table 1.1. Day-ahead forecast error and the stress penalty, by balancing authority

Balancing authority	MAPE, normal (%)	MAPE, stress (%)	Stress penalty (pp)	95% interval
New York ISO (NYIS)	2.68	3.49	+0.81	[+0.44, +1.19]
Southwest Power Pool (SWPP)	7.67	8.04	+0.40	[−0.98, +1.93]
ERCOT (ERCO)	2.13	2.44	+0.31	[−0.14, +0.85]
Midcontinent ISO (MISO)	2.42	2.58	+0.15	[−0.17, +0.49]
California ISO (CISO)	8.35	8.49	+0.14	[−0.84, +1.16]
Southern Company (SOCO)	1.89	1.69	−0.20	[−0.44, +0.03]
PJM Interconnection (PJM)	3.50	3.19	−0.31	[−0.56, −0.04]
ISO New England (ISNE)	2.87	1.14	−1.73	[−1.85, −1.60]
All eight regions pooled	3.93	3.88	−0.05	—

Notes: The stress penalty is MAPE in stress hours minus MAPE in normal hours. A stress hour is in the top 5 percent of the region's own demand within season. Point estimates and interval bounds are the median and the 2.5th and 97.5th percentiles of a day-block bootstrap with 2,000 resamples, from reports/bootstrap_intervals.csv. Regions are ordered by the penalty. Three intervals exclude zero: NYIS, PJM and ISNE. The pooled row is computed on the combined sample and is not bootstrapped.

The ordering in Table 1.1 makes a methodological point worth stating separately. Southwest Power Pool has the second-largest point estimate in the table, +0.40 points, and an interval running from -0.98 to $+1.93$. A ranking that reported point estimates alone would place SWPP among the worst performers; the interval says the data do not distinguish it from no effect at all. Figure 1.2 plots the penalties with their intervals, and grays the bars whose interval covers zero.

Stress penalty by region: 3 of 8 regions differ from zero
Lines show 95% confidence intervals. Gray bars: the interval includes zero.

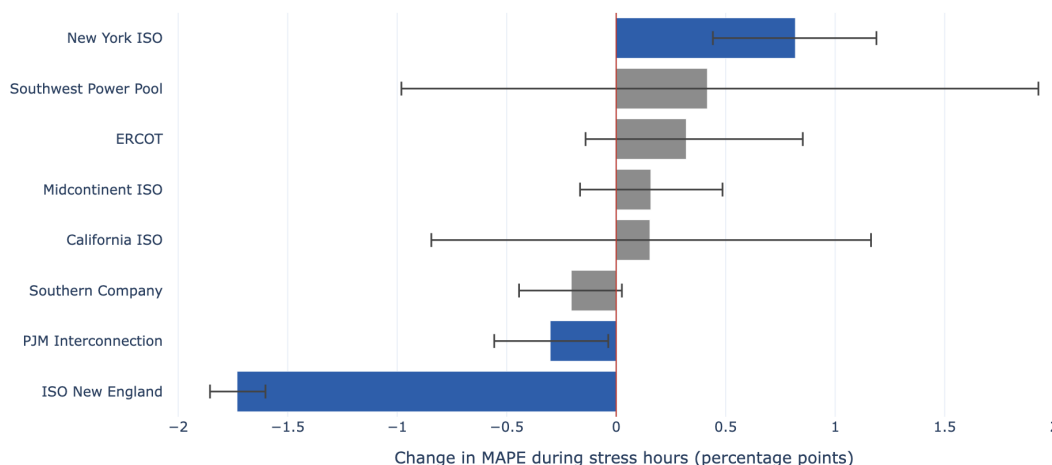


Figure 1.2. Stress penalty by region with 95 percent block-bootstrap intervals

Notes: Bars are the estimated stress penalty in percentage points; whiskers are 95 percent intervals from a day-block bootstrap. Bars whose interval crosses zero are gray. Five of eight regions are gray.

Figure 1.3 decomposes error by local hour of day. Both the stress-hour and normal-hour curves rise through the morning and peak near midday, and the gap between them is widest at midday. Overnight the two are close, and in some hours stress-hour error is the lower of the two. The pattern indicates that whatever drives the difference is tied to the daytime load shape rather than to the level of demand as such.

Forecast error by hour of day

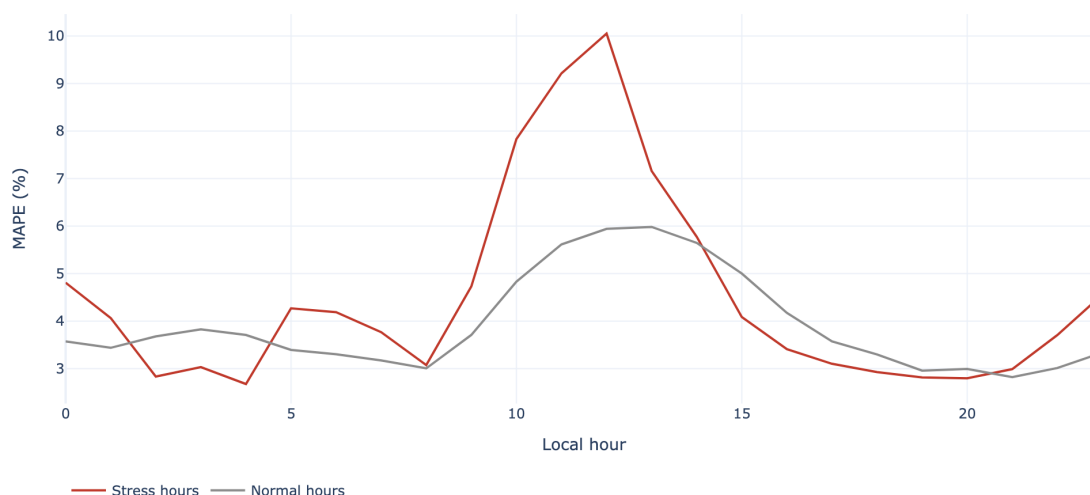


Figure 1.3. Forecast error by local hour of day, stress and normal hours

Notes: Local hour on the horizontal axis. A validation check confirms that daily peak demand lands at hours 17 to 19 local in every region, so the time-zone mapping is behaving correctly.

5.3 Skill against a persistence baseline

An accuracy level is uninformative without a benchmark. I score each forecast against persistence — yesterday’s demand at the same hour — and report the share of the baseline’s error that the published forecast removes. Across the sample, the published forecasts remove 26 percent of baseline error in normal hours and 32 percent in stress hours, so on this benchmark the forecasts are relatively better at peak, not worse. Two regions have negative skill scores: California ISO and Southwest Power Pool day-ahead forecasts are, on this measure, less accurate than simply repeating the previous day.

5.4 The result depends on how stress is defined

Peak demand and fast net-load change are different phenomena, and the second is arguably the harder operating problem. I therefore re-ran the entire pipeline with stress hours defined by the top five percent of net-load ramp rate rather than the top five percent of demand. Only 5 to 12 percent of the flagged hours are common to the two definitions, so this is close to an independent test on the same data.

Table 1.2. Stress penalty under a net-load ramp definition of stress

Balancing authority	MAPE, normal (%)	MAPE, stress (%)	Ramp penalty (pp)	95% interval
Southern Company (SOCO)	1.79	3.56	+1.78	[+0.38, +4.56]
PJM Interconnection (PJM)	3.39	5.17	+1.78	[+1.52, +2.04]
Southwest Power Pool (SWPP)	7.67	8.37	+0.72	[−0.03, +1.85]
Midcontinent ISO (MISO)	2.43	2.44	+0.01	[−0.12, +0.14]
ERCOT (ERCO)	2.15	2.13	−0.02	[−0.17, +0.15]
New York ISO (NYIS)	2.73	2.57	−0.15	[−0.28, −0.03]
ISO New England (ISNE)	2.85	1.45	−1.40	[−1.51, −1.27]

Balancing authority	MAPE, normal (%)	MAPE, stress (%)	Ramp penalty (pp)	95% interval
California ISO (CISO)	8.54	4.80	-3.74	[-4.16, -3.30]

Notes: Stress hours here are the top 5 percent of hours by net-load ramp rate within region and season, rather than by demand level. Intervals as in Table 1.1. SOCO has the largest point estimate and the widest interval in the study; PJM's estimate is essentially the same size and far more tightly bounded. PJM reverses sign between the two definitions: more accurate than usual at peak demand, less accurate than usual during fast net-load increases.

The comparison of Tables 1.1 and 1.2 is the paper's central caution. PJM is one of the three regions that looked measurably better at peak under the demand definition, and it is among the two worst under the ramp definition. Southern Company is indistinguishable from zero under the first and has the largest estimated penalty under the second. California ISO improves sharply under the ramp definition. A reliability assessment that reports only one of these two tables is reporting a choice of definition as though it were a property of the forecasts.

CISO: average demand and net load by hour
The midday dip and evening rise in net load is known as the duck curve.

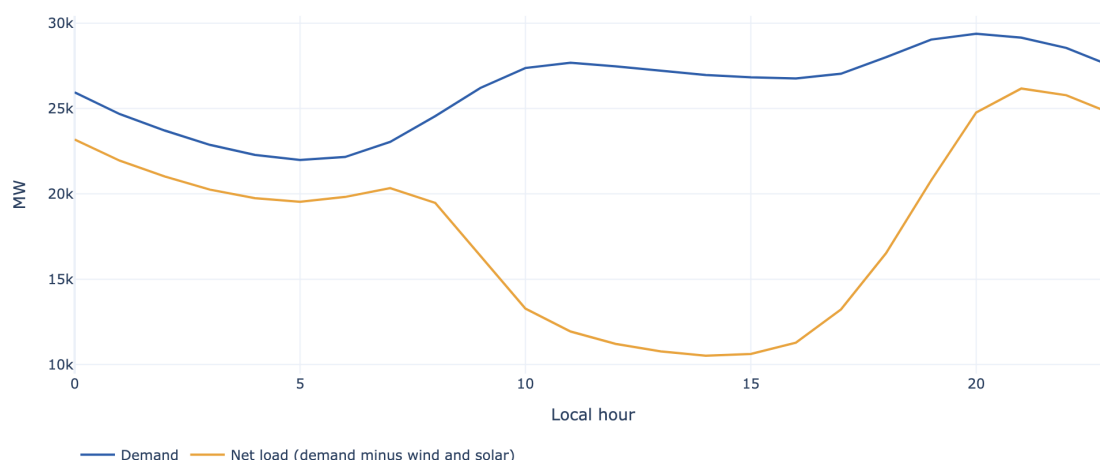


Figure 1.4. California ISO demand and net load, illustrating the ramp definition

Notes: Net load is demand minus wind and solar output. The gap between the two lines is renewable generation; the steep evening rise in net load is the duck curve. Hours of high demand and hours of fast net-load change do not coincide.

5.5 Directional bias

The most consistent feature of these forecasts is not that they fail at peak but that they miss in the same direction. In seven of the eight regions the forecast errs in one direction in more than 65 percent of hours. California ISO's mean signed error is -6.95 percentage points against a MAPE of 8.36, so roughly 83 percent of its average absolute error is a standing offset rather than noise. Southwest Power Pool is the mirror image, over-forecasting by +6.14 points on average. Midcontinent ISO runs high in 84 percent of hours and New York ISO runs low in 84 percent. ERCOT, at 58 percent of hours above actual, is the only region near balanced.

The distinction matters for what a remedy would look like. A noisy forecast requires a better model. A forecast that misses in the same direction almost every hour requires recalibration, which is considerably cheaper, and the direction of the bias identifies the exposure: an under-forecasting region buys short in the

real-time market, while an over-forecasting region commits capacity it does not need.

5.6 Magnitude

Under-forecast energy during stress hours totals 7.34 million MWh over the study period. Valued at \$100 per MWh this is \$733.9 million, with a band of \$293.6 million to \$1,467.8 million across a \$40 to \$200 per MWh range. This is an upper bound in two senses: the price is an assumption taken from published sources rather than measured here, and the calculation assumes every under-forecast megawatt-hour is purchased at the real-time price, whereas in practice reserves and demand response absorb part of it.

6 Data quality

Two failures in the federal source affect the headline numbers. Both were detected by the validation stage, documented, and left in place rather than corrected, on the view that a project measuring forecast accuracy should not quietly repair the data it is measuring.

First, California ISO's reported values do not balance. Net generation minus interchange should equal demand; for CISO the discrepancy exceeds five percent in 6,396 of the 17,304 hours in which the identity can be tested. Restricting to hours that pass the check lowers CISO's MAPE from 8.36 to 6.95 percent and moves its stress penalty from +0.14 to -1.67 points. A substantial share of CISO's measured forecast error is therefore an artifact of its reporting, not a property of its forecast.

Second, Southwest Power Pool has one day of corrupted forecast data. From 01:00 on 16 April 2026 through midnight, its reported day-ahead forecast runs between 5 and 34 percent of actual demand for twenty-four consecutive hours, while net generation tracks demand normally — the signature of an error in the forecast field rather than a genuine operating event. Removing that single day cuts SWPP's stress penalty from +0.42 to +0.12 points. Its negative skill score is essentially unchanged (-0.77 to -0.74), so the finding that SWPP's forecast underperforms persistence does not rest on that day. The original validation script tested demand for implausible values but had no equivalent test on the forecast field; that check was added after the fact.

SWPP: demand and day-ahead forecast, January 23 to January 29, 2026
The dotted line marks the hour described above.

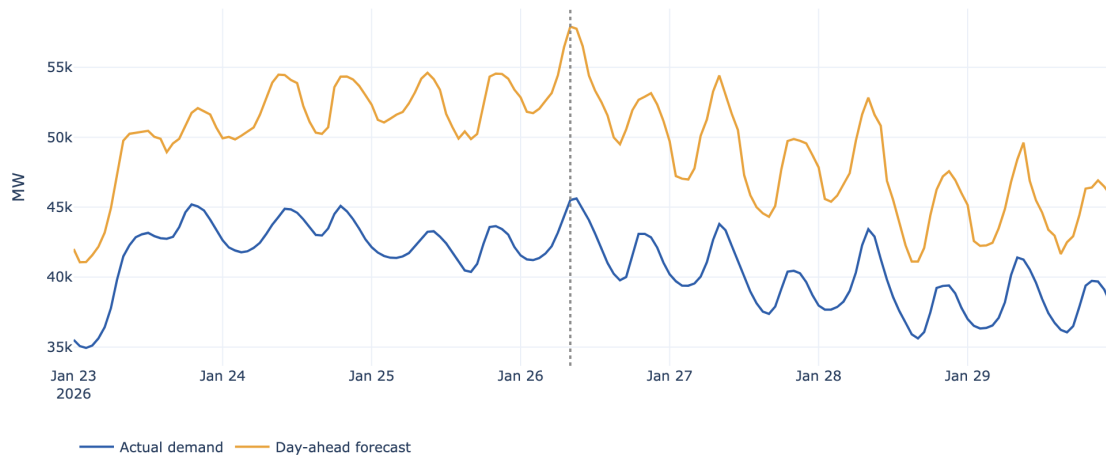


Figure 1.5. Southwest Power Pool demand against day-ahead forecast, 23–29 January 2026

Notes: This week contains the largest forecast error among extreme-demand hours that pass every data check: 08:00 on 26 January, demand 45,510 MW against a forecast of 57,910 MW, an error of 27.2 percent. The forecast tracks the daily shape but sits above actual demand in all 168 hours of the week, by at least 5,206 MW.

7 Limitations

Midcontinent ISO and Southwest Power Pool each span more than one time zone and are assigned a single local zone, which displaces some hours relative to true local time and blurs the hour-of-day decomposition for those two regions. The sample is twenty-four months, which is short relative to the weather variation that drives demand; an unusually hot or unusually mild year could reorder the regional rankings. The dollar figures in Section 5.6 rest on an assumed price, reported with a sensitivity band rather than measured. And the study compares forecast error across hour classes; it does not observe the operator’s loss function, so it cannot say whether the observed error pattern is costly relative to what the operator was optimizing.

8 Conclusion

For most of the U.S. grid, the concern that day-ahead demand forecasts fail in the highest-demand hours is not supported by two years of EIA-930 data. Pooled error is 3.88 percent at peak against 3.93 percent elsewhere, and five of eight regions show no difference that survives a day-block bootstrap. New York ISO is the clear exception, less accurate at peak by 0.81 points while relying on an average of 2,399 MW of imports in those hours.

The two results most likely to be useful operationally are secondary to the original question. The peak-hour verdict depends on how peak is defined: under a net-load ramp definition PJM and Southern Company become the problem regions and PJM reverses sign outright. And the dominant regularity in the data is directional bias rather than peak-hour failure, most starkly in California ISO and Southwest Power Pool, which is a recalibration problem rather than a modeling one. Part of California ISO’s measured error, finally, originates in its own reported data rather than in its forecast.

References

- Hong, Tao, and Shu Fan. 2016. “Probabilistic Electric Load Forecasting: A Tutorial Review.” *International Journal of Forecasting* 32 (3): 914–938.
- Hyndman, Rob J., and Shu Fan. 2010. “Density Forecasting for Long-Term Peak Electricity Demand.” *IEEE Transactions on Power Systems* 25 (2): 1142–1153.
- Künsch, Hans R. 1989. “The Jackknife and the Bootstrap for General Stationary Observations.” *Annals of Statistics* 17 (3): 1217–1241.
- Politis, Dimitris N., and Joseph P. Romano. 1994. “The Stationary Bootstrap.” *Journal of the American Statistical Association* 89 (428): 1303–1313.
- U.S. Energy Information Administration. 2026. Hourly Electric Grid Monitor, Form EIA-930. Washington, DC.

Data and code availability

All data are public and require no license. The source is the U.S. Energy Information Administration, Form EIA-930, retrieved through API v2. The full pipeline — pull, validation, DuckDB star schema, block bootstrap, Power BI semantic model and HTML dashboard — together with the raw and processed parquet files, the data-quality report and an independent audit of every headline number, is available at github.com/josealemanm/united-states-energy-grid-stress-analysis. An interactive version of the report runs in the browser at josealemanm.github.io/united-states-energy-grid-stress-analysis.