GRIDCAST

A Reproducible Framework for California Electricity Demand and Price Forecasting

APPLIED ENERGY ANALYTICS / RESEARCH RELEASE 0.1

Real CAISO observations. Explicit forecast origins. Measured performance.

Demand: lightgbm achieves 24-hour held-out MAE 1191.45 MW, 31.2% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 93.2%.

Price: ridge_100 achieves 24-hour held-out MAE 6.03 USD/MWh, 7.1% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 90.7%.

Historical dataset: 2024-2025. Final held-out issues: July-December 2025. Local CPU experiments. No live deployment or financial savings claimed.

01 / Executive summary

Demand: lightgbm achieves 24-hour held-out MAE 1191.45 MW, 31.2% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 93.2%. Price: ridge_100 achieves 24-hour held-out MAE 6.03 USD/MWh, 7.1% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 90.7%. These are historical research results, not validated live operational performance.

02 / Business problem

The platform compares forecasting strategies for planning research, peak-load awareness and market risk analysis. It does not implement dispatch optimization, trading execution or procurement cost minimization. No dollar savings are claimed.

03 / California electricity market background

CAISO actual integrated load measures its balancing-area demand and excludes other California balancing authorities. NP15 total day-ahead LMP is a hub price in USD/MWh; it is neither a statewide retail rate nor real-time settlement price. OASIS labels the price value column MW; normalization corrects its unit using the market product definition [1-3].

04 / Research objectives

Evaluate direct predictions 1, 6, 24, 48 and 168 hours after an explicit hourly UTC issue timestamp. Compare simple baselines with statistical autoregression, linear regression, boosted trees and a small GRU under an expanding-origin protocol.

05 / Data sources

The primary inputs are CAISO SLD_FCST/ACTUAL for CA ISO-TAC and PRC_LMP/DAM, total LMP only, for TH_NP15_GEN-APND [1-3]. EIA-930 access was verified on a 48-hour CISO sample; it is not silently merged or substituted [5]. Open-Meteo Sacramento reanalysis supports explanatory plots only [6]. NOAA and CEC were researched as future complementary sources [7-8].

06 / Data acquisition

The executed window is 2024-01-01 00:00 UTC through 2026-01-01 exclusive. OASIS responses are downloaded in 28-day chunks, validated as ZIP/CSV, retained with URL, acquisition time and SHA-256, and skipped on repeat ingestion only when their checksum matches. HTTP retries handle rate limits and transient server failures. Longer OASIS windows are possible using historical bulk archives; this two-year window balances market comparability and laptop workloads.

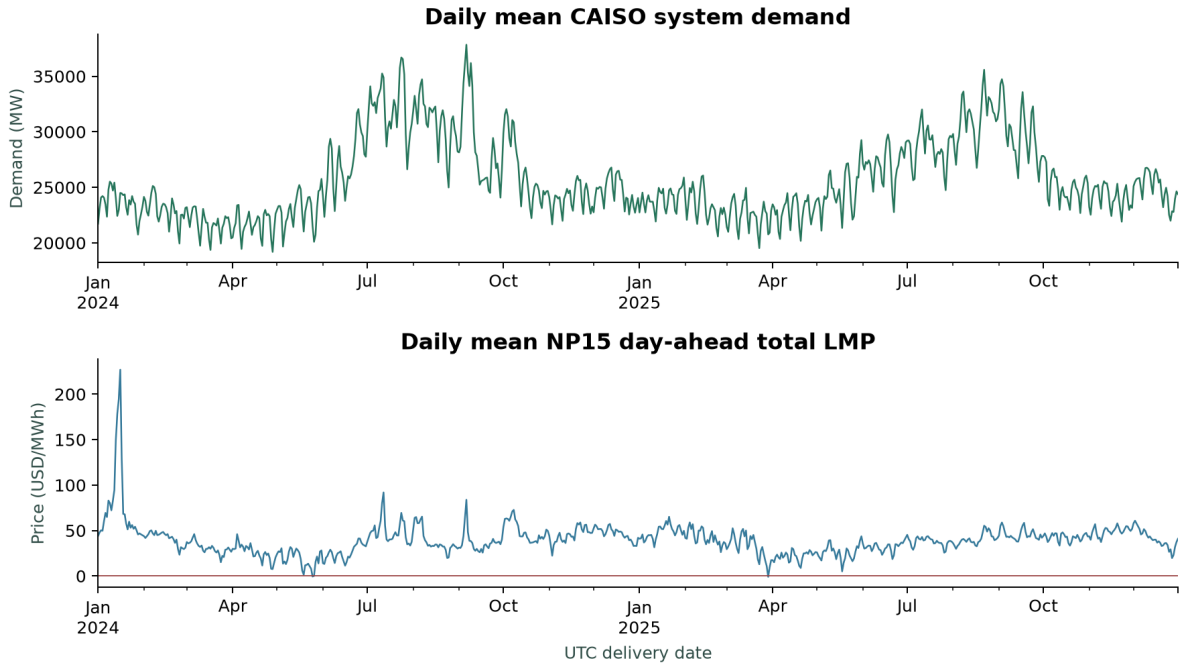
07 / Data quality assessment

Demand has 17,543 / 17,544 expected hours; the missing interval starts 2024-08-31 17:00 UTC. Price has 17,544 / 17,544 expected hours. Duplicate timestamps and invalid interval lengths cause errors. IQR flags are descriptive and do not delete extremes. Observed demand peaks at 47,759 MW; 559 hourly day-ahead prices are negative.

Target	Observed hours	Missing	Min	Max
demand	17,543	1	14,997.00	47,759.00
price	17,544	0	-38.47	582.64

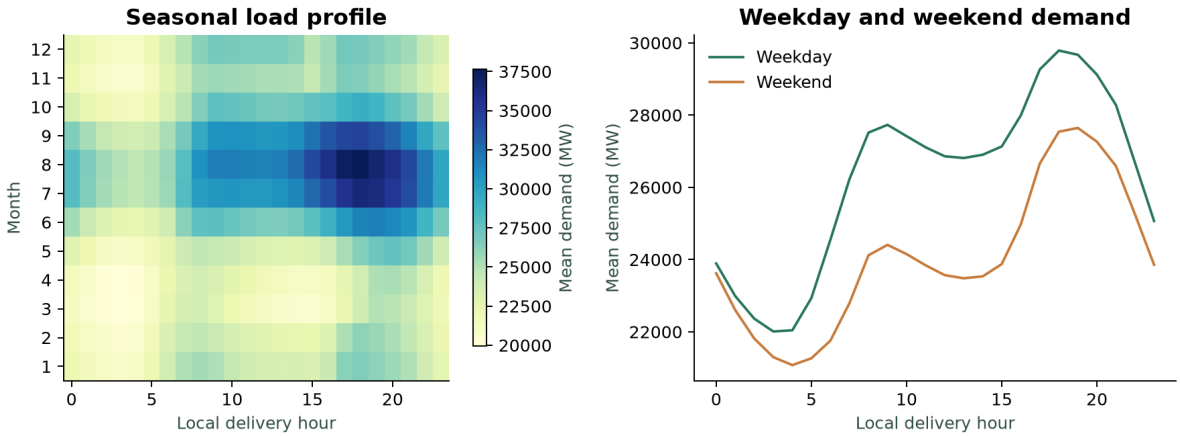
08 / Exploratory data analysis

The retrospective demand-temperature Pearson correlation is 0.669 across 17,543 matched hours. This unadjusted association mixes seasons, hours and local conditions; it is not causal. Seasonal and intraday summaries are computed using local market time. Price tails and rolling dispersion are inspected without clipping. Two years do not establish a long-run structural demand trend. Renewable generation, net load and multi-hub comparisons are outside the executed scope.



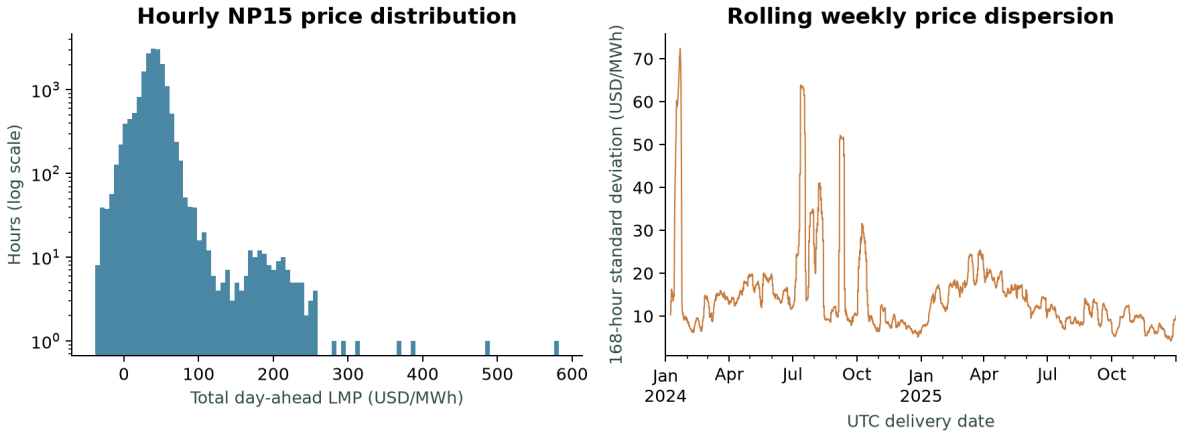
Source: CAISO OASIS; GridCast calculations

Demand and prices are different products and geographies. Daily means summarize hourly observations; one demand gap is not imputed.



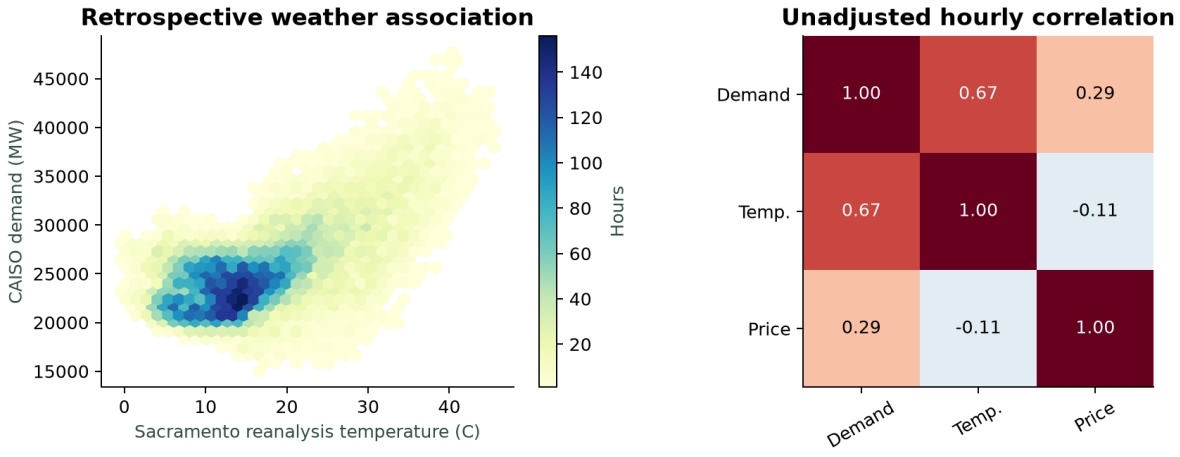
Source: CAISO OASIS; GridCast calculations

Calendar patterns use America/Los_Angeles time. Fall-back hours both contribute; spring-forward hours are absent.



Source: CAISO OASIS; GridCast calculations

Negative values and heavy tails remain in training and evaluation. Dispersion is descriptive, not an ex-ante volatility forecast.



Source: CAISO OASIS; Open-Meteo historical reanalysis (CC BY 4.0)

One Sacramento grid cell is an illustrative weather proxy, not load-weighted California weather. Reanalysis is excluded from forecasting; correlations do not identify causal effects.

09 / Forecasting problem formulation

For issue time t and horizon h, the target is y(t+h), where timestamps identify interval starts. Each horizon has its own fitted direct model. Demand y(s) is assumed usable at s+3h (two hours after interval end); price y(s) is deliberately withheld until s+48h. These are conservative assumptions, not verified publication vintages. For day-ahead LMP, some short-horizon targets may already be public at t; this is delivery-anchored statistical research, not a market gate-closure submission backtest.

10 / Feature engineering

Ten historical lags are offset by the target-specific availability delay: delay + {0,1,2,3,6,12,24,48,72,168} hours. Rolling means and standard deviations use 24/168 hours ending at the latest assumed available observation. Calendar sin/cos terms encode target local hour, weekday and day of year; weekend is binary. A weekly UTC anchor repeats the most recent fully available weekly profile. It does not correct seasonal-naive DST phase shifts. Holiday, renewable and weather-forecast predictors are deferred.

11 / Leakage prevention

Every fit label is assumed published before calibration starts; every calibration label is published before the evaluation quarter. Automated tests mutate all unavailable observations and verify that issue-time features stay identical for every horizon/delay pair. Rolling windows are shifted before aggregation. Realized future weather and outcomes never enter the predictor matrix. Retrospective corrected CAISO values may still differ from what was available historically; strict revision-vintage leakage cannot be ruled out.

12 / Experimental design

2024 initializes training. January-March and April-June 2025 are validation folds; July-September and October-December are final test folds. Models refit at each quarter using an expanding window, with stride three for CPU efficiency. The final quarter may incorporate previous-quarter labels once available, as a real expanding-origin evaluation would. It never retunes hyperparameters from those test results. Origins whose labels cross quarter boundaries are excluded. All candidate models share origins within a target/horizon.

13 / Baseline models

Persistence repeats the latest assumed available observation. Seasonal naive uses the latest available matching UTC weekly hour, repeating a week when horizon plus delay exceeds 168 hours. Historical mean and weekday/hour calendar means use fit data only. A baseline can win model selection; the application never forces a complex model to appear superior.

14 / Statistical models

The executed statistical comparator is a regularized direct seasonal autoregression over lag and weekly-anchor terms, without calendar predictors. It is not a SARIMA model. ARIMA/SARIMA, SARIMAX and exponential smoothing were considered but not executed in this release; state-space specification and repeated refits would require a separately budgeted comparison. No results are attributed to unrun models.

15 / Machine learning models

Ridge standardizes all features using fit data and compares alpha 1 vs 100. XGBoost uses 140 trees, depth four, learning rate .05, squared-error loss and two CPU threads. LightGBM uses 140 trees, depth five, 24 leaves, learning rate .05 and two CPU threads. Random forests were not added because two distinct boosted-tree implementations already provide nonlinear comparisons. The registry records the exact fitted estimator parameters.

16 / Deep learning models

The executed PyTorch GRU reads ten irregular lags ordered oldest to newest. Its 24-dimensional final hidden state is concatenated with standardized rolling/calendar context and passed through a linear output head. This is a sparse lag sequence, not a regularly sampled 168-step sequence. Five epochs use Adam (.003), batches of 256, MSE loss, gradient clipping at one, seed 42 and CPU execution. Standardization and target scaling fit on training only. LSTM, TCN and TFT are deferred, not claimed.

17 / Hyperparameter optimization

The bounded validation search compares two ridge penalties and fixed predeclared structural choices for all other candidates. Selection minimizes mean quarterly validation MAE independently by target/horizon and is persisted before that combination's first test fold. This is intentionally modest tuning, not an exhaustive optimization study. The full configuration and checksum are stored with results.

18 / Walk-forward validation

At each quarterly refit, a contiguous pre-fold calibration block of 720 eligible hourly issues is held out. Its end is offset from the fold by horizon + publication delay. Fit labels are purged at calibration start, preventing training targets from overlapping calibration availability. Quarterly retraining holds the fitted mapping fixed within each fold while features update hourly. These are repeated historical direct forecasts, not recursive trajectories.

19 / Evaluation metrics

MAE = mean absolute error. RMSE = square root of mean squared error. Bias = mean(prediction - actual). sMAPE = mean(200 times absolute error / (absolute actual + absolute prediction)), with both-zero pairs set to zero. MASE divides MAE by pre-evaluation weekly seasonal absolute differences [9]. Fold-level MASE scales expand; summary MASE uses the fixed pre-validation or pre-test history. Negative prices are supported. Near-zero prices make sMAPE secondary; 95th-percentile absolute error and event segments assess tails.

20 / Experimental results

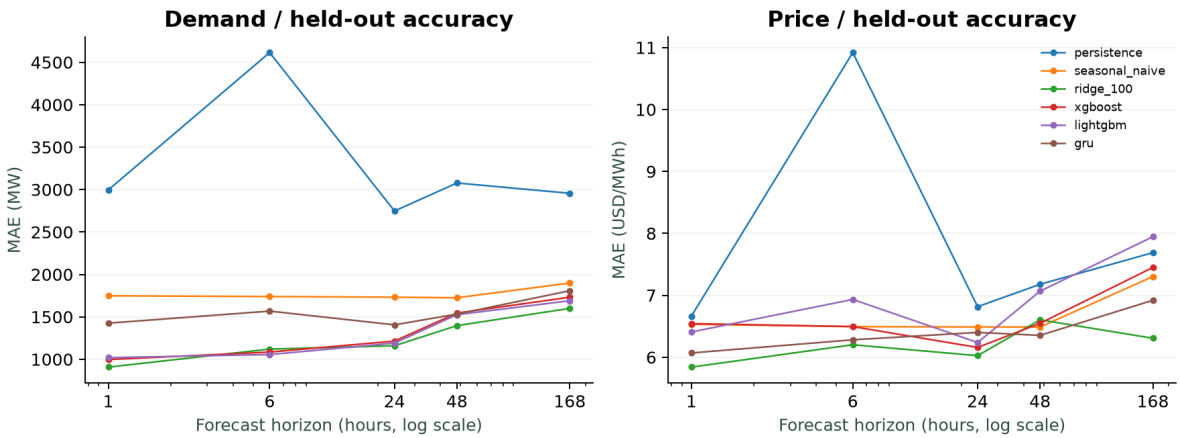
Demand: lightgbm achieves 24-hour held-out MAE 1191.45 MW, 31.2% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 93.2%. Price: ridge_100 achieves 24-hour held-out MAE 6.03 USD/MWh, 7.1% relative MAE reduction versus weekly seasonal naive; 90% nominal interval coverage is 90.7%. Complete per-model, per-horizon metrics, runtime and block-bootstrap MAE intervals are stored in experiments/results/summary.json. The interval on mean absolute error resamples contiguous groups of 168 ordered predictions (500 draws); it is approximate and does not establish economic value.

Demand / validation-selected models

Horizon	Model	MAE	RMSE	Bias	Coverage
1h	ridge_1	871.45	1129.90	30.63	90.0%
6h	lightgbm	1055.87	1411.93	-125.05	93.0%
24h	lightgbm	1191.45	1638.88	-64.48	93.2%
48h	lightgbm	1525.27	2114.00	-102.91	92.4%
168h	xgboost	1731.29	2402.31	-104.62	93.6%

Price / validation-selected models

Horizon	Model	MAE	RMSE	Bias	Coverage
1h	ridge_100	5.84	7.53	3.12	91.8%
6h	ridge_1	6.25	8.00	3.16	91.4%
24h	ridge_100	6.03	7.79	3.35	90.7%
48h	ridge_100	6.60	8.41	4.08	88.8%
168h	ridge_1	6.36	8.32	2.05	85.4%

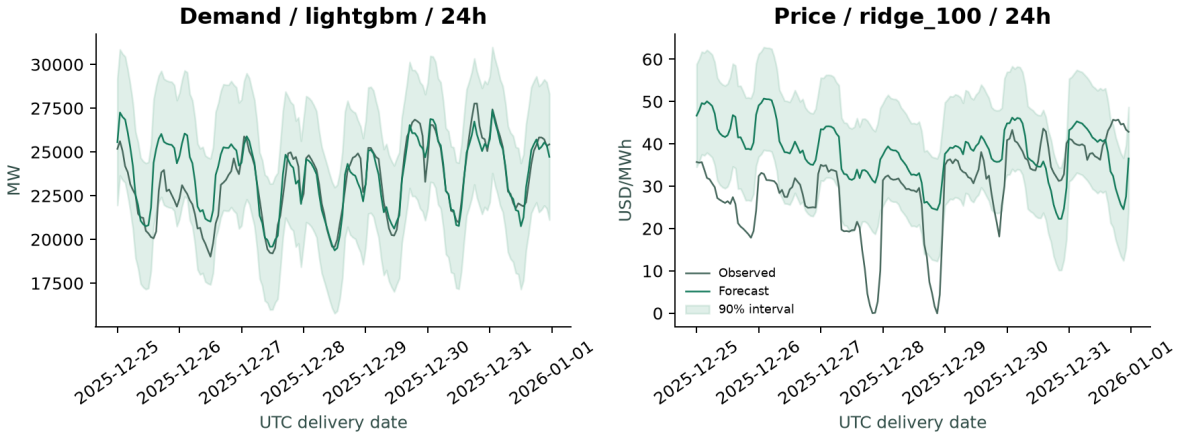


Source: CAISO OASIS; GridCast calculations

All reported test models use the same eligible origins per target and horizon. Horizon sample counts differ because labels crossing quarter boundaries are excluded.

21 / Model comparison

Tables report every candidate, including weak models. Selection uses validation performance, while test tables provide an honest final comparison. A different model winning a test column does not authorize retrospective reselection. A seven-day horizon is one point prediction at t+168h, not a seven-day optimized load schedule.



Source: CAISO OASIS; GridCast calculations

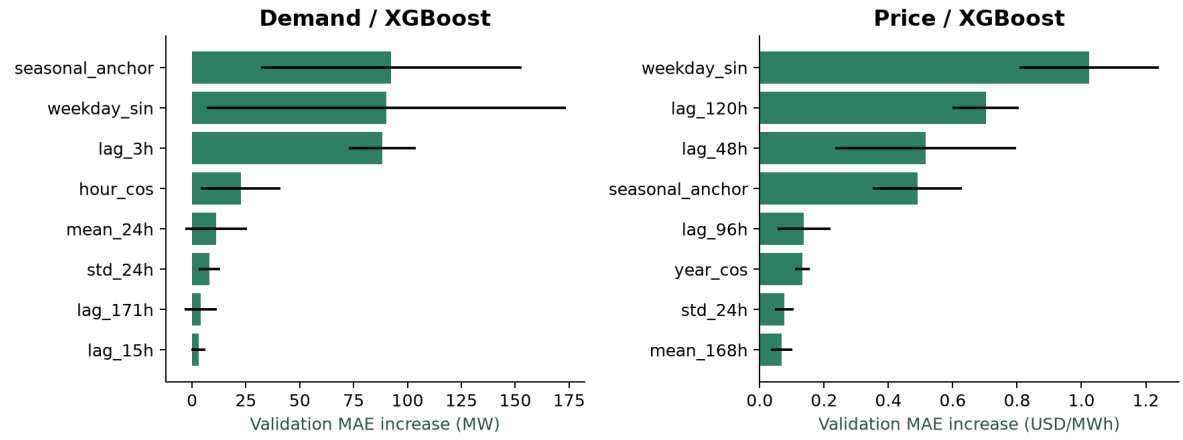
The last available week is shown for the model selected on validation. Split-conformal bands are calibrated before the quarter, not arbitrarily drawn.

22 / Forecast error analysis

Segments report season, local hour, weekday/weekend, high-value observations and large hourly changes. High-value thresholds are the 95th percentile of pre-fold history; volatility events use the 95th percentile of past absolute hourly changes. These are post-outcome diagnostic slices, not predictors. The H2 test only covers summer, fall and part of winter; it cannot establish spring test reliability. Extreme-weather-stratified performance remains unimplemented.

23 / Model explainability

Validation-only permutation importance reports MAE increases after feature shuffling, with three repeats and a 12-hour sample stride. The dashboard permits model/horizon selection. SHAP and partial-dependence studies are deferred. Shuffling correlated time-series lags can create unrealistic combinations, so importances are diagnostic predictive associations rather than independent effects or causal drivers.

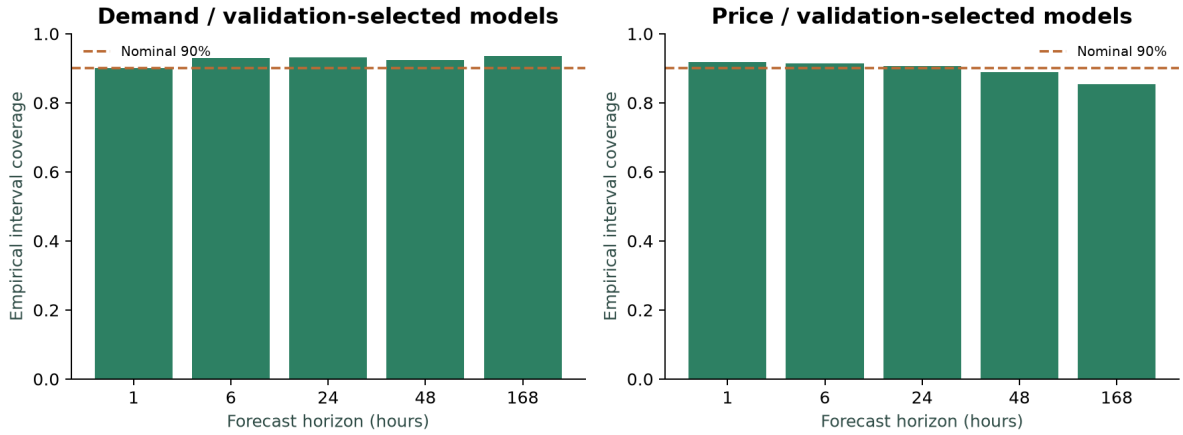


Source: CAISO OASIS; GridCast calculations

Validation-only permutation importance uses three repeats on a 12-hour subsample. Correlated lags can divide or mask importance; negative values indicate noisy attribution.

24 / Forecast uncertainty

Split-conformal intervals add/subtract a pre-fold absolute-residual quantile with finite-sample rank correction [10]. Calibration is separate from fitting; its labels are available before each forecast fold. Nominal coverage is 90%. Report empirical coverage, widths and segment coverage. Time dependence, market shift and event rarity mean the usual exchangeability guarantee should not be asserted. Scenario predictions do not inherit these bands.



Source: CAISO OASIS; GridCast calculations

Serial dependence and distribution shift invalidate a blanket exchangeability guarantee. Empirical coverage, event coverage and interval widths must guide use.

25 / Scenario analysis

The API scales historical lag levels, rolling means, standard deviations and the weekly anchor by 1 plus a user-selected percentage, then recomputes the prediction using the same eligible model artifact. Zero perturbation reproduces the stored forecast. This is a hypothetical feature sensitivity experiment; it is not a validated temperature response, renewable dispatch model or causal policy simulator.

26 / Operational applications

Supported uses are historical method comparison, peak-error assessment, demand planning research and price-risk monitoring research. Planning relevance must be assessed from the exact horizon and availability assumptions. Forecast errors are not converted to dollars because no procurement policy, settlement model or validated trading strategy exists.

27 / Limitations

The application is a tested local research release, not a production-certified forecasting service. Limitations include two years, a single price hub, revision-vintage uncertainty, assumed publication delays, short-horizon DAM targets possibly already public, no archived weather forecasts, limited tuning, coarse quarterly retraining, time-dependent conformal residuals, and no renewable or holiday covariates. No deployed status is claimed.

28 / Future improvements

Prioritize archived issue-time publication vintages and gate-closure-aware DAM evaluation, then population/load-weighted weather forecast archives, regional prices, renewable generation and net-load features. Add untouched prospective evaluation, drift alerts with calibrated thresholds, scheduled jobs with lock/atomic promotion, live inference and controlled public deployment. Compare SARIMAX/ETS and larger neural models only if their incremental value justifies laptop cost.

29 / Conclusion

GridCast now executes an authentic-data acquisition-to-evaluation pipeline and serves artifact-backed historical inference with explicit provenance. Its strongest contribution is a reproducible, inspectable comparison under stated information assumptions. Operational readiness requires prospective validation and the unresolved data-vintage work.

30 / References

References below link to official source documentation and methodological material. Data remain governed by source-specific terms; the repository's code license does not relicense CAISO or weather data.

1. CAISO OASIS portal
https://www.caiso.com/systems-applications/portals-applications/open-access-same-time-information-system-oasis

2. CAISO API specification
https://www.caiso.com/documents/oasisapispecification.pdf

3. CAISO OASIS FAQ
https://www.caiso.com/documents/oasis-frequently-asked-questions.pdf

4. CAISO terms
https://www.caiso.com/privacy-terms-of-use

5. EIA-930 methodology
https://www.eia.gov/electricity/930-content/about.html

6. Open-Meteo historical weather
https://open-meteo.com/en/docs/historical-weather-api

7. NOAA global hourly archive
https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database

8. CEC electricity data
https://www.energy.ca.gov/data-reports/energy-almanac/california-electricity-data

9. Forecasting: Principles and Practice - accuracy
https://otexts.com/fpp3/accuracy.html

10. Conformal prediction introduction
https://arxiv.org/abs/2107.07511