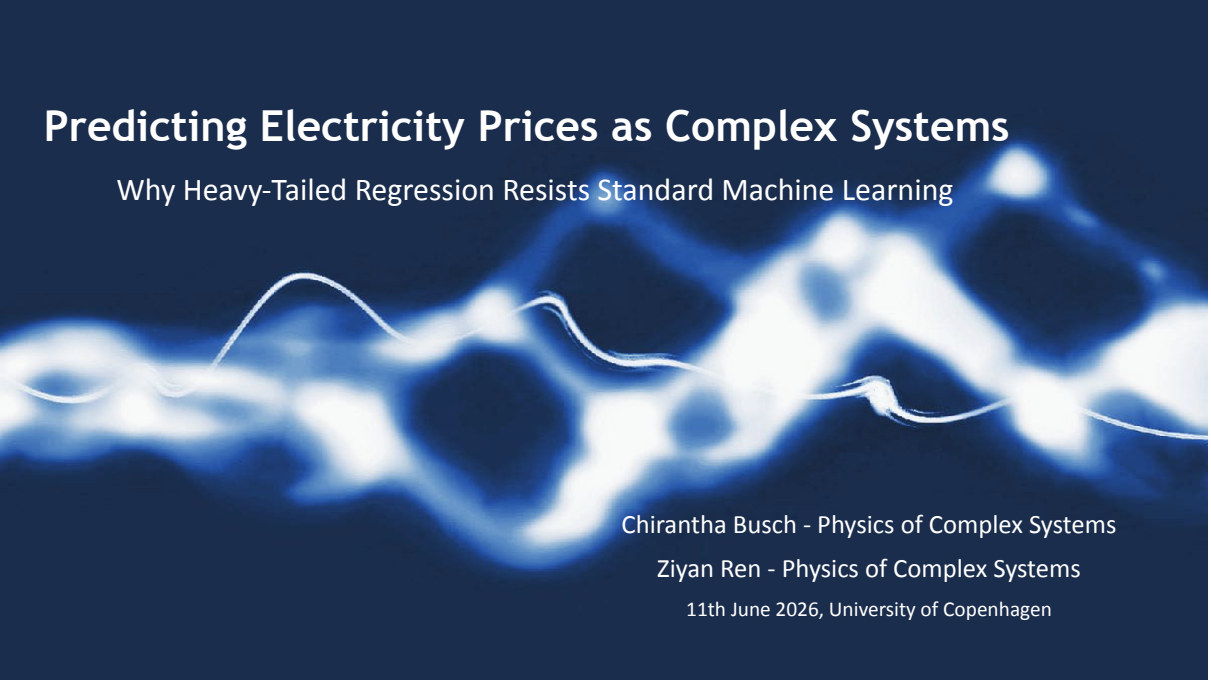
Predicting Electricity Prices as Complex Systems

Why Heavy-Tailed Regression Resists Standard Machine Learning



Chirantha Busch - Physics of Complex Systems

Ziyan Ren - Physics of Complex Systems

11th June 2026, University of Copenhagen

The Complexity of Future Predicting

"Over a dozen generations of men have pored over these equations...to the last decimal place, and put them together again. They've watched nearly four hundred years pass and against the predictions and equations, they've checked reality, and they have learned."

— Isaac Asimov, Second Foundation

We built a price prediction model driven by the joint feedback between natural and social variables to explore the structural limitation

It is beyond the model itself

Baseline → Spike failure → Cascade fails → SHAP diagnosis → Loss sweep → External coupling

Multiple-Layer Factor

Layer A — Global Events and Energy Shock:

Geopolitical events

Public reaction

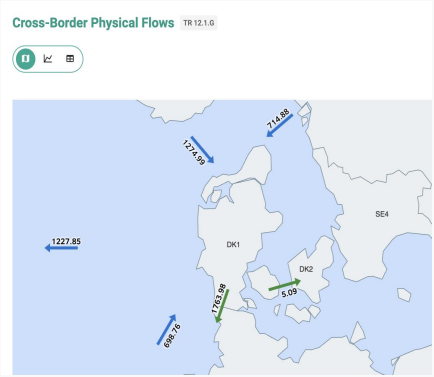
Market crisis

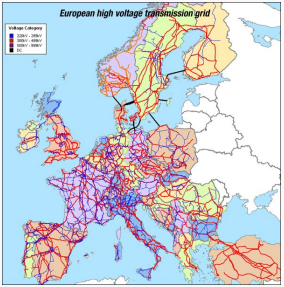
Layer B — European Power-Market Coupling:

Cross-border flows

European network

Energy distribution





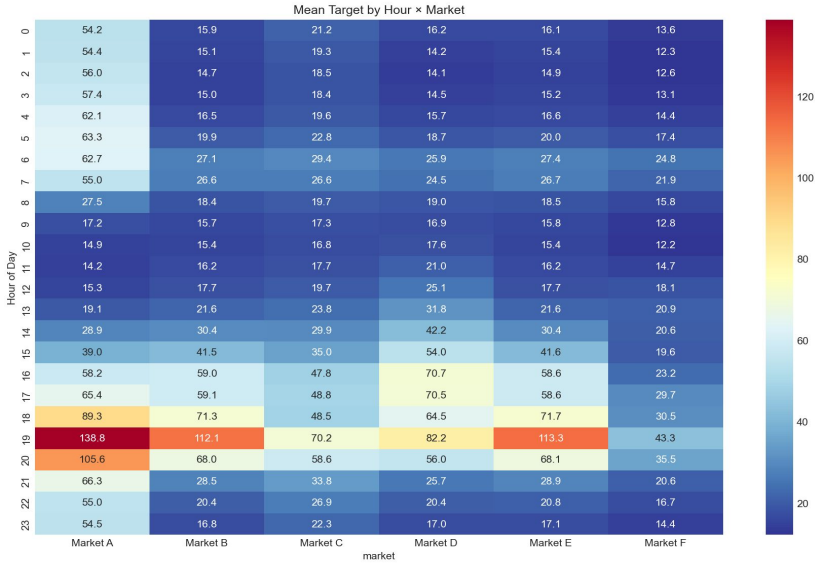
Layer C — Local Market and Weather:

Load, wind, solar, weather……

Our Data

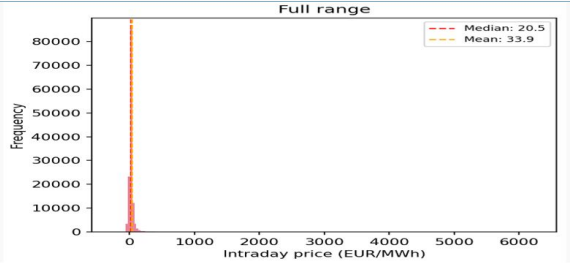
6 markets, hourly, 2023–2025.
133k train / 13k test rows
~25 features each.
Three complications:

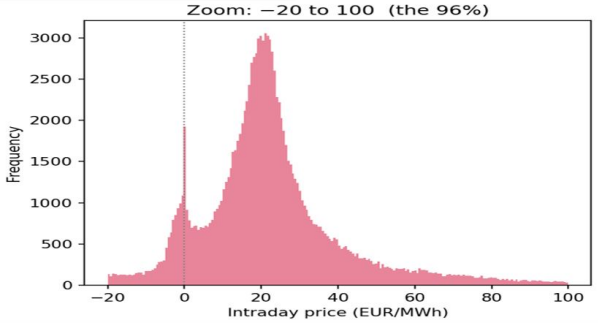
1. **Market A**: 46% of all spikes.
2. **Distribution shift**:
   train: 2023 Jan - Aug 2025
   test: 2025 Sep - Nov
3. **Block missingness**: test has 65× more missing weather data

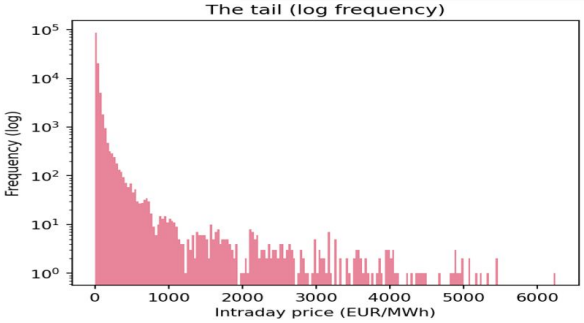


Mean price by hour × market — note Market A and the 17–20h ramp.

Extreme By Nature





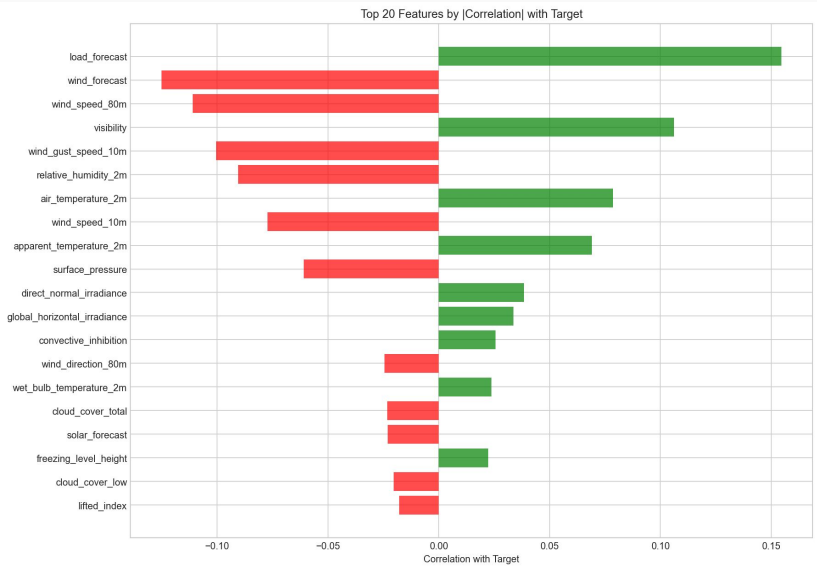


Price distribution: full range, zoom, and log-scale tail

Baseline: Four-Model Gradient-boosting Ensemble

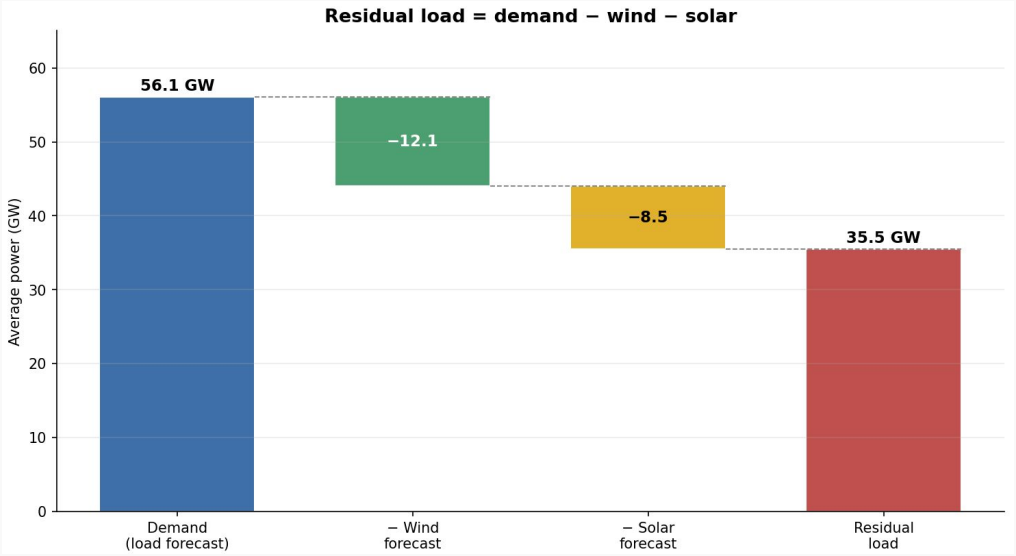
- **LightGBM ×2**
  XGBoost
  CatBoost
- ~60 ML features:
  - `energy balance`
  - `market stress`
  - `nonlinear physics`
  - `ramps/shocks`
  - `temporal regime`
  - ……
- Huber loss, sample weights, OOF-tuned blend

Test RMSE ≈ 28.4



Top features by |correlation|. Residual load (load−wind−solar) is the strongest driver.

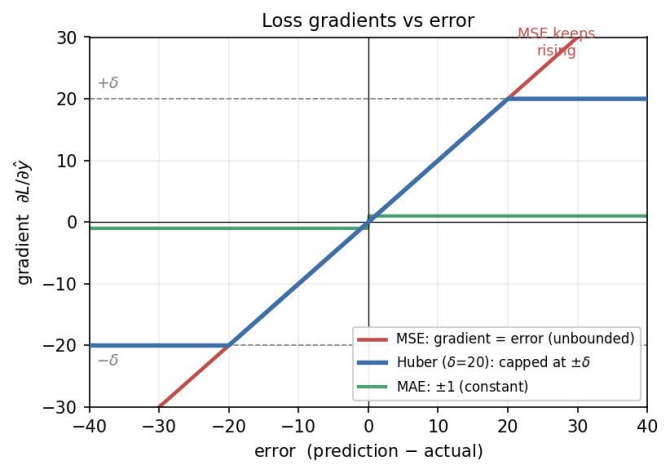
Residual Load



Why Huber Loss as a Ceiling

$$\frac{\partial L}{\partial \hat{y}} = \begin{cases} (\hat{y} - y) & |\hat{y} - y| \leq \delta \\ \delta \cdot \text{sign}(\hat{y} - y) & |\hat{y} - y| > \delta \end{cases}$$

Under MSE a 6,000 spike sends a gradient of 6,000. So it destabilises training. Huber caps it, stabilising the 96% normal rows.

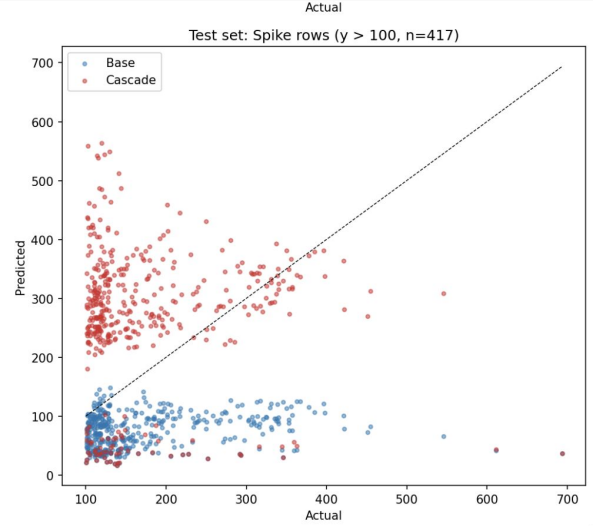


Stability vs. spike reach: this one tension drives the whole project.

Diagnosis: the Baseline Under-Predicts Spikes

- 20 worst predictions = 24% of total error - all under-predicted spikes.
- Market A RMSE **54.9** vs. 16–21 elsewhere.
- Predicted [−20, 225] vs. actual [−110, **694**].

A systematic, one-directional failure on the tail.



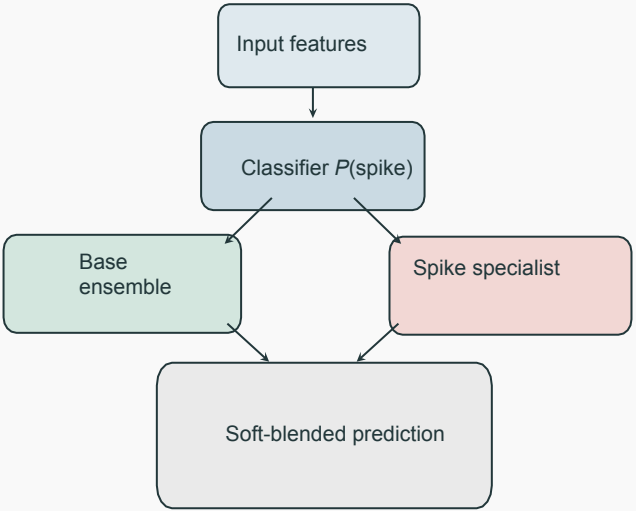
Baseline predictions stay far below the diagonal on spike rows.

ATTEMPT 1 - Splitting the Normal and the Extreme

Split the problem into two

- **Stage 1 – Classifier:** P(price>100),
  AUC 0.92
- **Stage 2 - Specialist:** trained only on spikes, free to reach high.
- **Blend** by spike probability.

Tested **60 configurations** (thresholds, specialists, blends, calibration).

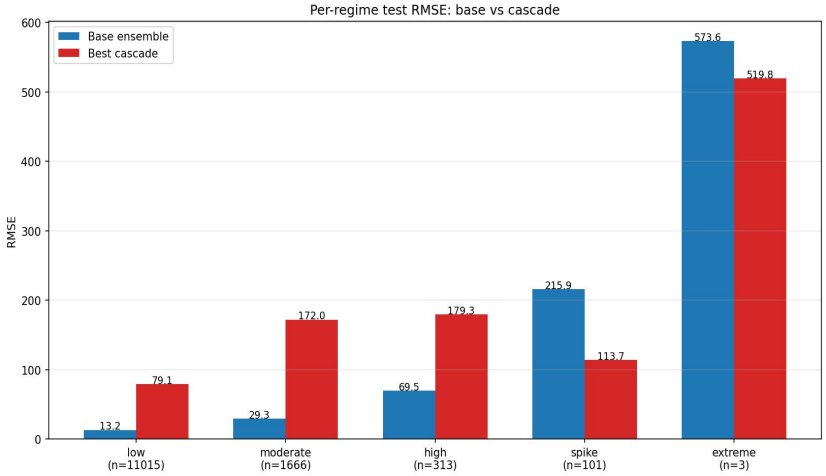


The Failure of the Cascade

- 0 of 60 configs beat the baseline on test.
- Best by CV (OOF **46.5**) scored **99.8** on test -- 3.5 × worse.
- Degradation is statistically significant (bootstrap, Diebold–Mariano).

Key finding

OOF-vs-test rank correlation = **0.06**: under shift, CV cannot pick the model that generalises.



Per-regime RMSE: the cascade helps spikes but wrecks the normal regime (84% of rows). Net: much worse.

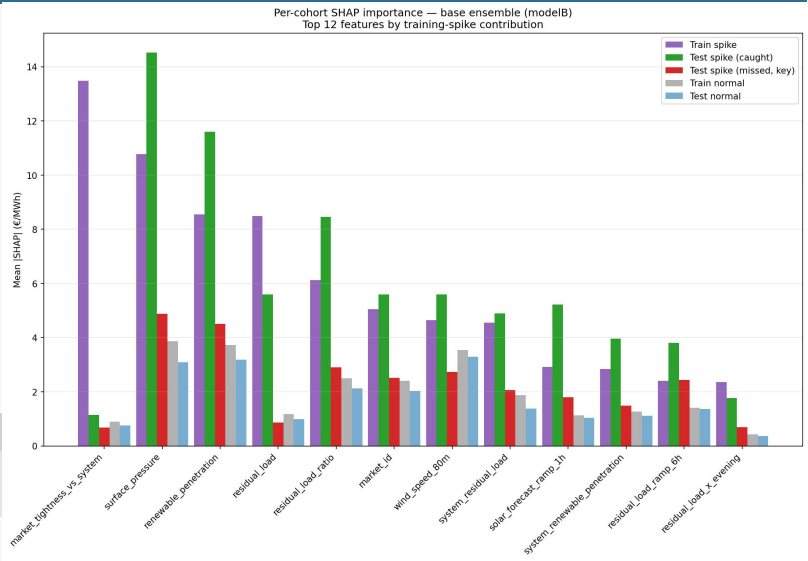
ATTEMPT 2 - Splitting the Mislead and the Absent

SHAP: *training spikes* vs. *test false positives*.

- Test precision only **18%** (302 false alarms).
- Their SHAP profiles are 90 -98% identical to real spikes.
- Every feature says "spike."

Failure mode:

The outcome shifted, not the features. The world changed.

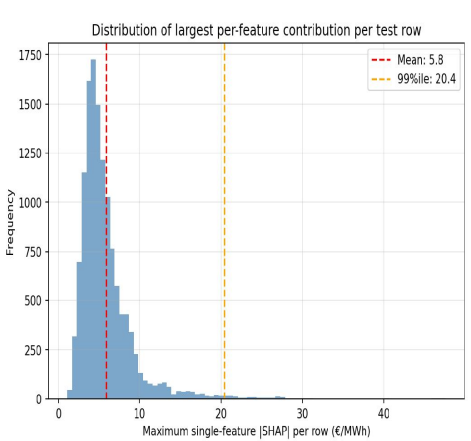


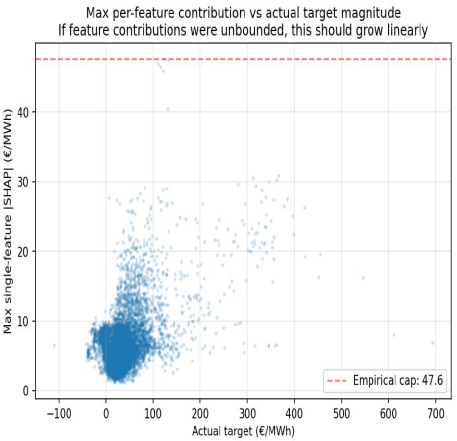
Per-feature SHAP: spikes vs. false positives. The bars line up , so identical spike evidence.

Why Does the Regressor Miss Spikes Entirely?

SHAP on the spikes the regressor *missed*:

- 417 test spikes: 18% caught, 26% missed entirely.
- Missed: actual **159**, predicted **36** (base rate).
- SHAP matches training *normals*, not spikes.





Failure mode: feature-incomplete

The features don't carry the signal. Not wrong but blind.

Max single-feature |SHAP| vs. price: flat (~5–18) even into the hundreds which is never enough to reach a spike.

Same Features, Opposite Reasons

	Classifier	Regressor
Failure	Feature-*misleading*	Feature-*incomplete*
Symptom	Fires on non-spikes	Misses real spikes
SHAP like	Training *spikes*	Training *normals*
Cause	Outcome shifted	Signal absent

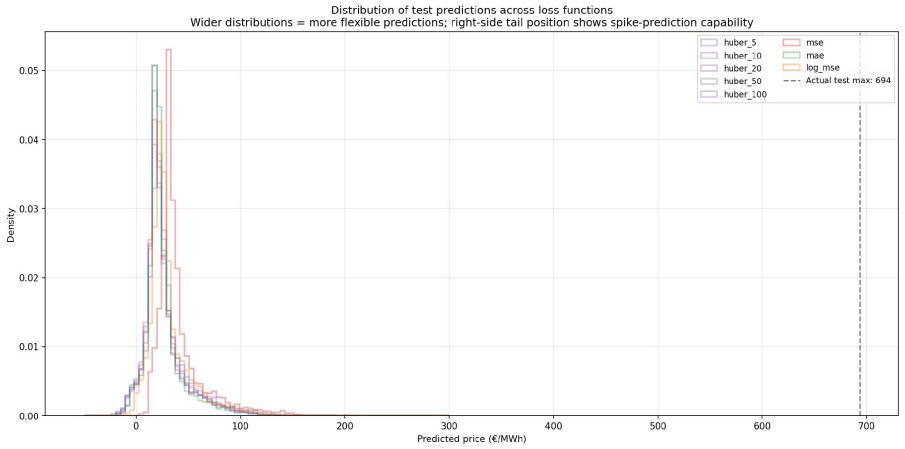
Same data, same features but opposite reasons. RMSE alone could never show this.

ATTEMPT 3 - Converging the Higher and the Lower

Hypothesis: the gradient cap bounds predictions. So we swept the loss.

Huber $\delta \in \{5, 10, 20, 50, 100\}$ plus MSE, MAE, log-MSE; only the loss changes.

Refuted:



- Even MSE (no cap) tops out at **263**.
- Actual test max is **694**.
- No loss comes within 60% of real spikes.

The Limitation Is Structural

Three independent attempts converge

The spike limit is not any single choice, not the architecture, the model component, or the loss/δ. It is a property of the **data + features + model class under distribution shift**.

- **Cascade:** architecture doesn't rescue it.
- **SHAP:** the signal is absent from the features.
- **Loss sweep:** every loss is bounded far too low.

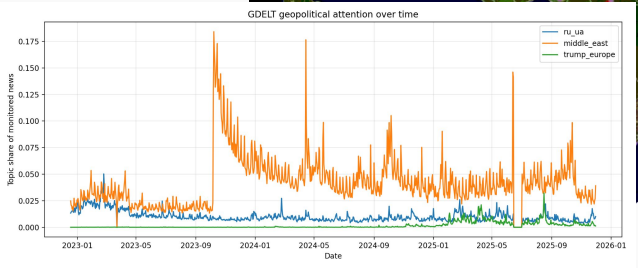
Generalising any heavy-tailed regression under these conditions.

ATTEMPT 4 - We Live in a Society…

Filter Keywords
Russian-Ukrainian War
Middle-Eastern Conflict
Trump's NATO/EU policy
(54 new features)

Intensity Comparison
Daily impact
lagging/rolling

Period Nodes
1d 2d 3d 1w 2w



A Global Database of Society

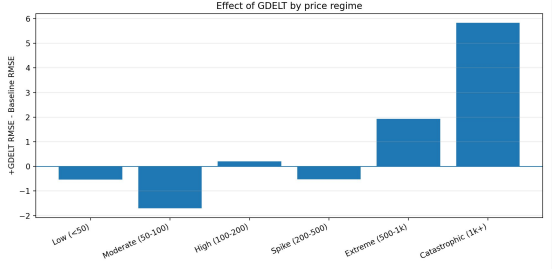
Supported by Google Jigsaw, the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

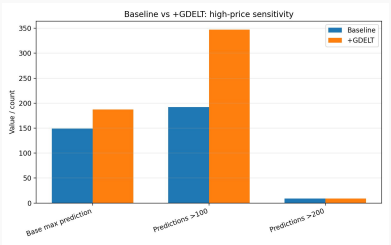
The Irrational Economic Man

Stablise ordinary regimes

⇩

Create more chaos in chaos

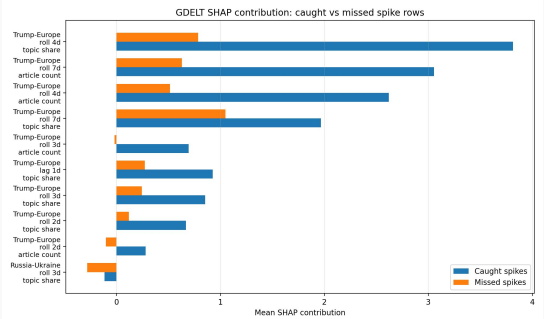




Predictions >100: 192 → **347**
Base max prediction: 148.8 → **187.1**

- Russia–Ukraine high-attention days
  → **higher** mean/max price and **higher** spike rate
- Trump-Europe → **Strongest** at catching attentions
- Middle East → no positive association in this sample

Power Hierarchy?



Evaluation

What worked

- Good RMSE in ordinary regimes.
- Multi-angle, mechanistic diagnosis
  cascade, SHAP, loss sweep, external
- Statistically grounded spike failures.

Not solved

- Spike magnitudes.
- Model selection under shift.
- Remaining hidden drivers

Future work

- Richer temporal models
  Sequential NN (LSTM/TCN)
  Hybrid Physical-ML
- Layer B: European energy coupling
  direct grid topology
  cross-border flows
- Layer A: Global Factor
  correlation and causation
  market response
  tariffs and taxes

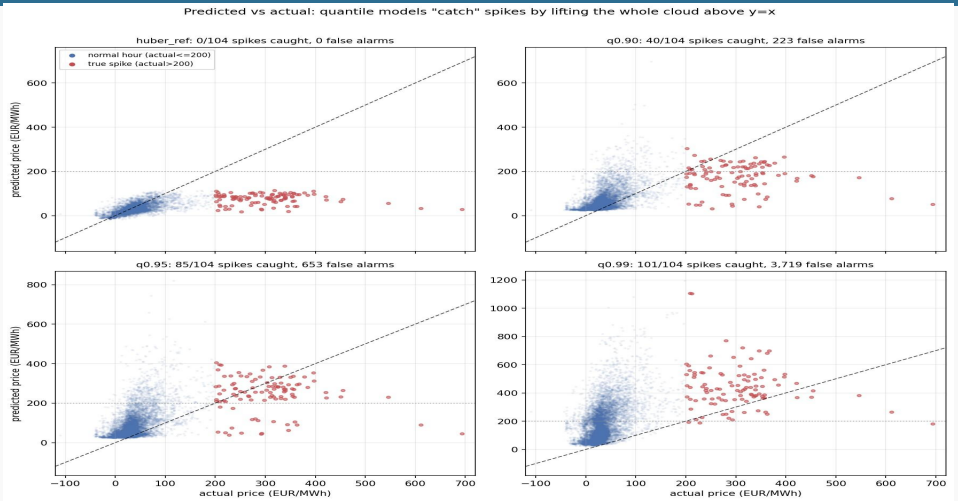
Conclusions

1. We made a predictor then we asked the harder question.
2. A principled cascade failed to generalise which was shown rigorously.
3. SHAP revealed two failure modes: a misled classifier, a blind regressor.
4. A loss sweep proved the spike bound is structural.
5. External factors radicalise the performance
6. **Lesson:** under distribution shift, interpretability and honest negatives beat another decimal of RMSE.

Thank you ☺️
Questions welcome.

Appendix

A0: quantile regression , spikes “caught” only by flooding



Identical features, CV split and sample weights to Model A , **only the loss changes**. The 0.99-quantile reaches **1,195** (true ma
694), so the ceiling is *not* representational; yet it flags **3,820** rows >200 against **104** real spikes (~37:1), and overall RMSE
worsens **29** → **170**. Spike recall rises only by lifting *every* prediction, the limit is the information in the features, not the loss.

A1 - Feature engineering (~60 features)

Energy balance

- $\text{residual load} = \text{load} - \text{wind} - \text{solar}$ (strongest driver, corr 0.22).
- Renewable penetration, wind/solar share, market tightness vs. system.

Non-linear physics

- Wind power $\propto$ wind speed^3; capped turbine curve.
- Effective solar = GHI $\times(1 - \text{cloud})$.
- Heating/cooling degrees (U-shaped demand).

Temporal

- Cyclical hour/month/day-of-week (sin/cos).
- Regime flags (evening peak, solar hours, night).

Lags & rolling windows

- Residual load lags 1/2/3/6/12/24h.
- Rolling mean/std/max over 6/12/24h, per market.

Cross-market

- Mean/std/min/max of weather across markets per hour.
- Each market's deviation from the cross-market mean.

Ramp & forecast-error proxies

- 1/3/6h rate-of-change of load, wind, solar.
- Weather-implied vs. provided forecast mismatch.

14 zero-importance features were pruned.

A2 - Cross-validation strategy

- Four strictly chronological folds, each with a ~3-month validation window, mimicking the test horizon:

Fold	Train through	Validate on
1	Aug 2024	Sep–Nov 2024 *(same months as test)*
2	Nov 2024	Dec 2024–Feb 2025
3	Feb 2025	Mar–May 2025
4	May 2025	Jun–Aug 2025

- No future data ever leaks into training. Lags/rolling features computed *within* each market's timeline.
- Early-stopping patience = 200 rounds, max 15,000 rounds.
- Despite this care, OOF rankings did not transfer to test (see main slides) — the shift is genuine, not a CV bug.

A3 - Base ensemble details

Models and typical blend weights

Model	Framework	Huber δ	Weight
A	LightGBM	10	~0.35
B	LightGBM	20	~0.50
C	XGBoost	10	~0.00
D	CatBoost	10	~0.15

Sample weights (counteract Huber gradient clamping): 1× normal, 2× moderate (50–100), 3× large (100–300), 5× extreme (> 300), 2× negative (< -30). **Market-bias correction:** per-market additive offset at 50% strength, applied after blending. **Note on the ensemble optimiser:** grid search zeros out spike-capable models (e.g. XGBoost), because overall RMSE is dominated by the 84% normal rows. Any model trading 0.5 normal RMSE for 5.0 spike RMSE gets weighted to zero — a structural reason the ensemble cannot self-correct toward spikes.

A4 - Cascade configuration grid (60 configs)

- **Spike threshold** $\tau \in \{100, 150, 200\}$ - defines what counts as a spike for the classifier and specialist.
- **Specialist objective:** MSE vs. log-MSE (log compresses the tail).
- **Blend type:**
  - *Hard:* route each row to one model by a probability threshold.
  - *Soft:* $\hat{y} = (1 - p^k)\,\hat{y}_{\text{base}} + p^k\,\hat{y}_{\text{spec}}$, sharpness $k \in \{1, 1.5, 2\}$.
- **Calibration:** isotonic vs. raw classifier probabilities.

Calibration caveat: isotonic calibration fit on OOF showed ECE = 0.000 — a memorisation artifact. Calibrated soft-blends scored a mean test RMSE of **95.6** vs. **42.8** for raw — calibration actively hurt under shift.

A5 - Statistical testing

Bootstrap confidence intervals (1,000 resamples of the test set):

- Cascade vs. base, OOF: -0.28 RMSE, 95% CI $[-0.47, -0.10]$ - a *significant improvement*.
- Cascade vs. base, test: +71.44 RMSE, 95% CI [+68.69, +74.12] - a *significant degradation*.
- The signs are opposite on OOF vs. test which is the core of the generalisation failure.

Diebold–Mariano test (equal predictive accuracy):

- OOF: $p = 0.005$ (cascade better).
- Test: $p < 0.0001$ (cascade worse).

Rank correlation: Spearman $\rho = 0.063$ ($p = 0.63$) between OOF and test ranking across all 60 configs — OOF is essentially uninformative for test selection here.

A6 - Cascade per-regime breakdown

Why the cascade looks good on OOF but fails on test: it trades a large improvement on rare spikes for a moderate degradation on the common rows and the common rows dominate the aggregate.

Regime	Base RMSE	Cascade RMSE
Low / normal (majority of rows)	~13	~70 *(much worse)*
Spike + extreme	~215	~113 *(47% better)*

The spike improvement is real but on test, the normal-regime damage (multiplied across thousands of rows) overwhelms it. On OOF, the spike rows happened to be more predictable, so the aggregate flipped sign.

A7 - SHAP methodology

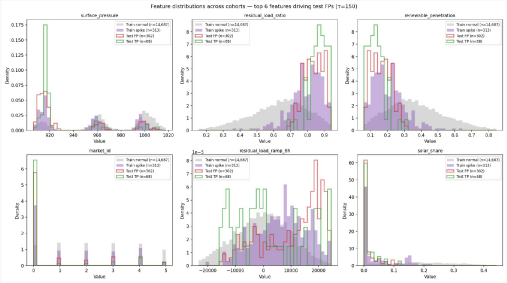
- **TreeExplainer** on the gradient-boosted models which are exact, fast SHAP values for tree ensembles.
- **Additivity check:** reconstructed predictions from SHAP values match model output to MAE = 0.0000 on the base regressor. The explanations are exact, not approximate.
- **Cohort analysis:** we compare mean |SHAP| per feature across defined cohorts (training spikes, training normals, test caught/missed/false-positive) to localise *where* the signal lives.
- **Distribution-shift quantification:** Kolmogorov–Smirnov tests on each top feature between training-spike and test-FP cohorts. So all top-15 features show $p < 0.01$ shifts, with ramp features shifting +106% to +177%.

A8 - Classifier SHAP: the feature-misleading evidence

- Top-feature cohort similarity (train spike vs. test FP): residual-load ratio **92%**, renewable penetration **91%**, system renewable penetration **98%**.

- Example FP - row 11952, Market A, 2025-11-23 00:00: actual price 33, predicted spike $p = 0.53$.

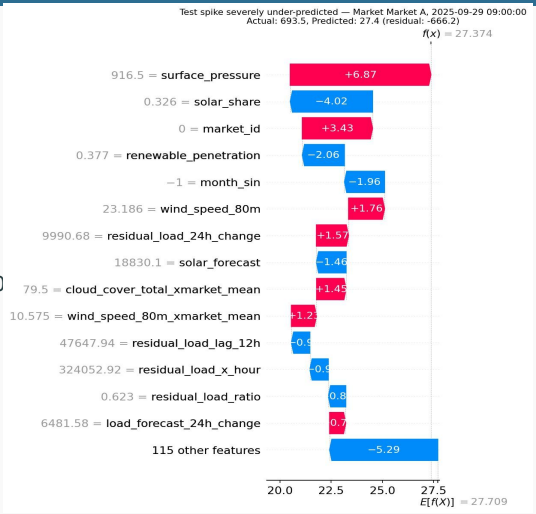
Surface pressure alone contributes +0.69 to the logit; every signal looks textbook-spike.



Feature value distributions: train-spike vs. test-FP cohorts. The cohorts overlap heavily - the classifier genuinely cannot separate them on the available features.

A9 - Regressor SHAP: caught vs. missed spikes

- **Caught** spikes: SHAP magnitudes *match or exceed* training-spike levels (ratios 1.07–1.38). The model had the signal and used it.
- **Missed** spikes: SHAP magnitudes match training-*normal* levels (ratios 0.74–1.26), *not* spike levels (0.05–0.66) - the signal was absent.
- **Worst single case:** row 4086, Market A, 2025-09-29 09:00 Actual **693.5**, predicted **27.4**. Top contributions roughly cancel — the model sees an ordinary morning.

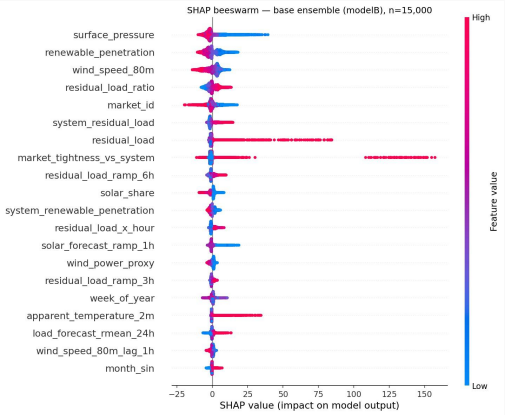


Waterfall for a missed spike: contributions are small and offsetting, landing the prediction near the base rate despite a

A10 - Bounded contribution across the whole test set

- Max single-feature |SHAP| across all 13,098 test rows: **47.6**.
- 99th percentile per-row max: **20.4**; mean per-row max: **5.8**.
- The max stays flat across regimes (low 5.0, moderate 8.6, high 11.3, spike 18.4). No feature is ever allowed to “shout.”

This is the empirical fingerprint of gradient capping but the loss sweep (next) shows the bound persists even without Huber.



Global beeswarm: every feature’s SHAP values are tightly bounded; none produces the large positive excursion a real spike would require.

A11 - Loss comparison: fullmetrics

Variant	Iters	CV RMSE	Test RMSE	Pred max	Max |SHAP|
huber 5	1734–5000	47.47	29.31	123	31.6
huber 10	938–4999	47.35	29.01	140	44.2
huber 20	418–3558	47.27	28.93	142	47.5
huber 50	255–916	**47.10**	28.61	136	39.6
huber 100	257–426	47.18	**27.93**	159	51.9
mse	79–273	54.24	30.35	**263**	52.5
mae	96–970	48.03	29.25	195	**72.3**
log mse	174–209	48.01	28.25	210	0.15*

*log-MSE SHAP is in log space, not comparable.**Two notable points:** (i) CV picks huber 50, test picks huber 100 - the OOF/test mismatch recurs. (ii) Our base used $\delta = 20$; $\delta = 100$ would have been ~3.4% better on test but CV would never have told us so.

A12 — Loss comparison: per-regime RMSE & the trade-off

Variant	low	moderate	high	spike	extreme
mae	**12.14**	32.14	74.58	224.13	591.28
huber 5	12.69	30.23	74.55	226.96	581.65
huber 20	12.95	30.30	71.38	221.91	582.97
huber 50	13.25	29.72	69.51	217.75	580.74
huber 100	14.13	29.39	64.79	205.81	573.71
log mse	14.63	29.87	66.17	205.28	572.09
mse	20.65	33.08	59.19	**186.63**	557.78

Read top-to-bottom: as you move from MAE to MSE, low-regime RMSE rises ~70% while spike-regime RMSE falls ~17%. This is the textbook normal-vs-spike trade-off, cleanly traced by the loss choice. The asymmetry is steep because normal rows are 84% of the data which is exactly why no loss “wins” overall by chasing spikes.

A13 - Why MAE produces the largest contributions

An unexpected result: MAE's max single-feature |SHAP| (72.3) is the highest of any loss, higher even than MSE. This contradicts naive gradient theory (MAE's gradient magnitude is a constant ±1). **The explanation is implementation-specific:**

- MAE's true Hessian is zero almost everywhere / undefined at the median.
- LightGBM substitutes a small constant Hessian.
- Leaf values are computed as $-\Sigma g / \Sigma(h + \lambda)$ - a small Hessian denominator inflates leaf magnitudes.
- Result: MAE builds fewer but more aggressive trees, concentrating large contributions on individual features.

Lesson: per-feature contribution bounds are not derivable from loss theory alone — the optimiser's implementation and early-stopping behaviour both mediate the outcome.

A14 - GDELT external-risk layer

Raw signal

- Article count: absolute number of matching news articles.
- Topic share: article count / total monitored articles.
- Topic share controls for variation in global news volume.

Feature construction

- lag1d
- rolling 2d / 3d / 7d / 14d
- shock vs 7d / 14d baseline
- event48h / event72h high-stress flags

Controlled experiment

- Same v17.1 ensemble.
- Same chronological CV folds.
- Same sample weights, Huber losses, and market-bias correction.
- Only change: add full-period GDELT features.

Feature count

- Baseline: 129 model features.
- +GDELT: 183 model features.
- Added 54 external-risk features.

	regime	baseline_base_rmse	gdelt_base_rmse	n	delta_rmse	improvement
0	Low (<50)	16.86	16.33	46478	-0.53	0.53
1	Moderate (50-100)	30.90	29.20	4220	-1.70	1.70
2	High (100-200)	68.07	68.27	1193	0.20	-0.20
3	Spike (200-500)	230.44	229.92	597	-0.52	0.52
4	Extreme (500-1k)	498.95	500.88	42	1.93	-1.93
5	Catastrophic (1k+)	1445.84	1451.67	24	5.83	-5.83

OOF effect

- Base OOF RMSE: 46.7995 → 46.6220.
- Conservative OOF RMSE: 46.7544 → 46.5574.
- The improvement is small but consistent.

High-price sensitivity

- Base max prediction: 148.8 → 187.1.
- Predictions >100: 192 → 347.
- Predictions >500 remain 0.

	method	oof_rmse	test_rmse	oof_spike_rmse	test_spike_rmse	test_minus_oof
0	Base ensemble	46.6220	26.8685	229.92	181.75	-19.7535
1	Hard cascade	46.6205	45.5095	229.78	135.69	-1.1110
2	Soft calibrated	46.2711	33.0096	218.98	156.95	-13.2615
3	Soft raw	48.0046	44.7065	207.63	122.06	-3.2981

A14 - Missing-data structure

- Test missingness is **4.54%** vs. train **0.07%** which was a 65× increase.
- Missingness is block-structured: of 111 affected delivery hours, **87** have *all six markets* missing weather simultaneously and entire forecast blocks drop out, not random cells.
- **Imputation pipeline:** (1) time-aware linear interpolation within each market; (2) fallback to market×month×hour median; (3) final forward/backward fill.
- This is itself a research question: the test distribution is degraded relative to training, compounding the temporal shift.

A15 - Reproducibility & code

- All randomness seeded (seed=42); LightGBM in deterministic mode.
- Pipeline: phase1 eda → phase2 features → phase3 (baseline / cascade) → phase4 shap → phase5 loss.
- Featured CSVs are regenerated from raw via phase2_features.py (not stored — 244 MB).
- Four detailed written reports (one per analysis phase, ~1,140 lines total) accompany the code.
- Repository: github.com/chirantha7/energy-price-prediction (private).

Each analysis phase is independently runnable and independently documented.

A16 - Supplementary plots index

Additional figures available in plots/ for Q&A:

EDA

- 02_target_by_market
- 03_target_by_hour
- 06_target_over_time
- 07_energy_forecasts_vs_target

Cascade

- cascade_predictions
- cascade_reliability

SHAP (base & classifier)

- shap_base_global_bar
- shap_base_cohort_comparison
- shap_base_waterfall_spike_caught
- shap_classifier_waterfall_*_conf_*

Loss

- loss_comparison_rmse
- loss_comparison_shap_pred