Calibrated probabilistic weather intelligence for New Jersey energy affordability

Ilika Geospatial LLC | NJ AI Innovation Challenge (NJEDA, administered by Plug and Play NJ)
Technical whitepaper | Ilika Geospatial | November 15, 2026

Summary

New Jersey's electricity bills carry the cost of uncertainty about the weather: the margin a load-serving entity holds against peak demand, the hedge it buys against heating and cooling load, and the operations it runs on peak days are all sized against a temperature forecast whose error is rarely stated and rarely verified. This paper describes a platform that turns the 51-member ECMWF ensemble into calibrated probability distributions of daily temperature for New Jersey's utility zones, on the station basis that indices settle on, verified out of sample against the raw ensemble and against climatology, and served every day. It reports the accuracy of those distributions, explains the three-layer method, and sets out how a calibrated distribution acts on each point of the cost chain. Every accuracy figure is out of sample.

1. The problem

A weather forecast enters a utility's cost chain as a number with a judgement margin. The ensemble forecasts that the meteorological centres publish do contain an uncertainty estimate, but as published the spread is too narrow and the mean carries a systematic bias that depends on place and season: across 55 cities the raw ECMWF ensemble's 80 percent band covers about 51 percent of outcomes at a station, and its continuous ranked probability score (CRPS) against the settlement sensor is 2.50 degrees Celsius over the first fifteen days. A planner who takes that band at face value under-provisions; one who distrusts it adds a margin of their own and over-provisions. Either way the cost reaches the ratepayer.

2. The data foundation

- ECMWF ensemble forecasts: the 51-member IFS ENS at 0.5 degrees over the continental United States, daily initialisations since 2022, leads 1 to 15 days; the ECMWF extended-range ensemble to 46 days with its 20-year reforecast set; the SEAS5 seasonal ensemble with hindcasts to 1981; the 51-member stream ingested daily for 24 cities including the seven New Jersey zones.
- Truth and normals: ERA5 reanalysis daily temperature at every grid point since 1990, with 1991 to 2020 climatological normals; 59 station series (ACIS and GHCND) including Newark, Trenton and Atlantic City.
- Context: elevation, distance to coast, teleconnection loadings; the ENSO, NAO, AO, MJO, PDO and AMO indices back to 1970; New Jersey service territories and the published time-of-use schedules of PSE&G, JCP&L and Atlantic City Electric.
- Store: a ClickHouse Cloud feature store of about two billion rows, validated for gaps, with every forecast and score traceable to its source file and model version.

3. The method: three layers

Layer 1, ensemble calibration. EMOS (ensemble model output statistics): the ensemble mean and spread are corrected by four parameters per grid cell and calendar month, fitted by minimising CRPS against ERA5, separately for the short-range, subseasonal and seasonal regimes. Gradient-boosted regime models with 45 to 54 features (ensemble statistics, climatology, geography, teleconnection and MJO signals) were trained and scored

against this baseline under blocked time-series cross-validation; they match it within one to two percent, so the simpler model is kept in production and the richer models are held as references.

Layer 2, station anchoring. One affine correction per station and calendar month maps the grid distribution to the station observation, (Tmax + Tmin) / 2 at the airport sensor on which indices settle, fitted by the same CRPS objective on 2022 to 2026 forecast and observation pairs. New stations are added without retraining Layer 1.

Layer 3, delivery. Fifty-one calibrated members per city and day, written by the daily pipeline to a live table and served to the dashboards and the data route; the member frame is the one the index calculator consumes, so a desk can replace its raw ensemble with the calibrated one without changing downstream code.

Scoring. CRPS is the training objective and the headline score; the mean absolute error of the median and the coverage of the 80 percent band are reported beside it. Every score is out of sample under leave-one-year-out or blocked cross-validation; no forecast is scored on a year its model saw.

4. Results

4.1 Against the raw ensemble, grid truth, 55 cities

Lead	Days scored	Raw ensemble CRPS (C)	Calibrated CRPS (C)	Change
0 to 7 days	504,630	2.39	1.14	+52%
8 to 15 days	573,480	2.57	1.71	+34%
16 to 30 days	1,634,580	2.29	1.97	+14%
31 to 46 days	1,714,920	2.35	2.01	+14%

Phase 0 baseline of the 60-day plan: 2022 to 2026 daily initialisations (leads 1 to 15), 2004 to 2025 extended-range initialisations (leads 16 to 46), ERA5 daily mean truth, leave-one-year-out. The raw ensemble's short-range score includes its offset against the daily-mean convention, which Layer 1 removes.

4.2 Against the settlement sensor, 59 stations

Score	Raw ensemble	Layer 1 (grid)	Layer 1 + 2 (station)
CRPS (C)	2.50	1.95	1.73
MAE of the median (C)	3.28	2.71	2.44
Coverage of the 80 percent band	0.51	0.67	0.77

Leads 1 to 15 days, 1.19 million forecast and observation pairs, 2022 to 2026, blocked leave-one-year-out by initialisation year.

4.3 New Jersey zones

Zone	Lead	Raw CRPS (C)	Calibrated CRPS (C)	Change	Raw daily HDD error (F)	Calibrated HDD error
Atlantic City	0-7 d	1.48	1.26	+15%	2.58	2.30
Atlantic City	8-15 d	1.83	1.78	+3%	3.49	3.36
Newark / New York City	0-7 d	1.18	1.06	+10%	2.07	1.99

Zone	Lead	Raw CRPS (C)	Calibrated CRPS (C)	Change	Raw daily HDD error (F)	Calibrated HDD error
Newark / New York City	8-15 d	1.63	1.53	+6%	3.09	3.00
Trenton	0-7 d	1.43	1.24	+13%	2.60	2.32
Trenton	8-15 d	1.99	1.90	+5%	3.82	3.64

Station	Layer 1 CRPS	Layer 2 CRPS	Layer 1 cov80	Layer 2 cov80
Atlantic City	1.95	1.91	0.65	0.77
Newark	2.04	1.86	0.57	0.77
Trenton	1.73	1.77	0.73	0.78

At Newark the 80 percent band of the station-anchored distribution covers 77 percent of outcomes against 58 percent for the grid distribution and 32 to 58 percent for the raw ensemble; the stated probability is the realised frequency, which is the property a planner needs.

5. The platform

- Residential Consumer view: hour-by-hour demand risk for the seven New Jersey zones against the household's own utility time-of-use schedule (PSE&G, JCP&L, Atlantic City Electric), with an AI energy advisor that names the hours to avoid and the hours to shift load into.
- Utility Operations view: the New Jersey service territories, calibrated P10/P50/P90 bands, a 14-day risk calendar, ranked peak stress windows flagged where they fall inside the tariff peak, and demand response watch windows.
- Trading and risk view: settlement distributions for temperature contracts benchmarked daily against market pricing.
- Daily pipeline on AWS: ensemble ingest, observations, calibration, pricing and a run report at 06:00 Eastern every day; a completed run on each of the last 30 days.

6. How the work reduces electricity bills in New Jersey

Procurement: a planner who procures against a calibrated ninetieth percentile instead of a judgement margin on a single value releases the part of the reserve that covered forecast error the forecast no longer makes; the Milestone 5 case study works through the arithmetic for a Newark-scale system under stated assumptions. Hedging: the hedge volume set against a reliable P90 is neither padded against a band that was too narrow nor wasted on one that was too wide. Operations: demand response is called on the days and hours when calibrated risk and the tariff peak coincide. Households: load shifted out of the hours where high demand risk and peak price coincide lowers the bill directly and lowers the system peak that sets capacity cost for everyone.

7. Limits and next steps

- Skill beyond two weeks is modest and comes from the dynamical ensemble; calibration makes it reliable and keeps it honest against climatology. The platform reports where the forecast has no skill rather than hiding it.
- The gradient-boosted regime models did not beat the simpler calibration by more than year-to-year variation; the next gain at subseasonal range is a better or multi-model input, which is in progress with a

European index partner.

- Two of the seven New Jersey zones (Morristown, Toms River) use the grid calibration until a station series is added for their airports.
- Realised savings depend on each utility's book and policy and are the utility's to measure on its own archive with the distributions now served daily; this paper reports forecast accuracy and the mechanism, not realised savings.

About

Ilika Geospatial LLC, Edison, New Jersey, builds calibrated probabilistic weather intelligence for energy and insurance. This work was supported by the NJ AI Innovation Challenge (New Jersey Economic Development Authority, administered by Plug and Play NJ). https://ilikallc.com, info@ilikallc.com.