**Japan Day-Ahead → Imbalance: Interim Report**

**Overview**

This document summarizes work to date on forecasting the Japan day-ahead to imbalance spread. It covers the current production baseline, preliminary results, and early hypotheses for why the signal works in Japan.

**Scope & period.** Data spans Japan’s regional pricing zones from **31-03-2022** to **31-12-2023**. For EDA and model testing, **31-09-2023** is used as the hard cutoff between train and test.

**Data available for modeling**

* **Prices:** Imbalance and spot prices, with flags describing calculation method (market vs. additional pricing mechanism).
* **Forecasts:** EC12 and EC00 for residual demand/load, solar, wind, temperature, and consumption.
* **Actuals:** Pumped-hydro storage and the **reserve-rate** variable used in the dual-price imbalance calculation.

*Note:* For **2024–2025**, JEPX/TSO/vendor datasets contain additional fields that should be incorporated in follow-on analysis.

**Production Baseline (for reference)**

The current production signal is a simple trend model: an **EWMA** with **10-day half-life**, computed **per 30-minute bucket**. Positions are taken from the sign/magnitude of this EWMA by bucket.

**Performance (test period)**

* [Figure: Cumulative PnL — Tokyo]
* [Figure: Cumulative PnL — Kansai]

**Why does the trend work (in Japan)?**

Performance is noticeably stronger in Japan than in Europe or the US. Two market features appear to explain much of the hourly shape of the spread.

**Mean spread by hour**

[Figure: Mean spread by hour]

**1) Reserve-rate threshold effects (17:00–20:00)**

The most pronounced pattern is a **negative spread** during **17:00–20:00**. This likely reflects **risk preferences around the reserve-rate mechanism**: participants bid up **spot** for later hours when the threshold is most likely to be hit, pulling spot above imbalance in those windows.

* [Figure: Frequency of reserve-rate threshold hit by hour]

**2) Pumped-hydro behavior & ramp constraints**

Pumped hydro accounts for roughly **~15%** of “production” in Tokyo (similar in other regions). Operators tend to **buy in the spot** when power is cheap early in the day and **sell later in the evening** when spot is stronger—reinforcing the hourly shape of the spread. In practice, **ramp constraints** may limit how quickly outage can be adjusted, strengthening the intra-day pattern observed.

* [Figure: Yearly mean of hydro outage by hour]

*Additional details and supporting plots are provided in the Appendix.*

want me to go on and reword **“Simple Improvements to the model”** (half-life sweep, floor handling, clipping/ACF) in the same style next?

**Simple Improvements to the Baseline**

**1) Half-life tuning**

We first tested whether the 10-day half-life in the baseline EWMA is close to optimal.  
Result: across a sweep of half-life values, performance peaks near the current setting; gains from further tuning are marginal.

* [Figure: EWMA half-life vs. R²/performance]

**Takeaway:** keep **HL = 10 days** as the default until other features materially change the optimum.

**2) Handling the spot floor**

Japan’s spot can clear at a hard floor of **0.01**. Short positions perform poorly on floor days because spot can’t move lower while imbalance can. This shows up clearly in conditional bin plots.

* [Figure: Bin plots — floor hit vs. not]

Two simple remedies:

1. **Masking:** avoid opening shorts during floor conditions. Simple, but it ignores sizing.
2. **Conditional averaging:** maintain **separate rolling means** of the spread for (a) floor-hit periods and (b) non-floor periods, and size/tilt positions accordingly.

**Takeaway:** conditional means preserve signal while reducing systematic floor bias.

**3) Spike memory / clipping**

Because EWMA averages **all** past values, large spikes (≈ 3σ events) bleed into subsequent days and bias the bucket-level signal. We observe elevated positive autocorrelation right after spike days that then decays back to baseline.

* [Figure: ACF around spike events — before vs. after]

Working remedy: maintain a **rolling mean of prior spike outcomes** and clip or dampen their influence in the EWMA update so the model **reacts quickly** on the day, then **forgets faster** afterward.

**Takeaway:** controlled clipping improves responsiveness without locking in spike bias.

**4) Combined tweaks (Floor + Clip)**

Applying the floor-aware conditioning **and** spike-aware clipping yields small but consistent gains versus the raw EWMA.

**Takeaway:** modest improvements across error metrics and PnL; serves as a stronger reference baseline before adding new features/hypotheses.

Yes—adding a short “bridge” block right after **Data available for modeling** will make the rest flow. Drop in these subsections (keep them tight—1–3 lines each) and then move into the baseline model.

Here’s a clean set you can paste in (use Heading 2/3 in Word):

1. **Data quality & preprocessing**
   * Coverage, missingness, outliers; resampling/alignment to 30-min.
2. **Leakage controls**
   * What was known at decision time; vendor lag rules; forecast cutoffs.
3. **Target definition**
   * Exact formula for day-ahead → imbalance spread; units; sign convention.
4. **Feature snapshot (current)**
   * Brief list by category (Prices, Forecasts EC00/EC12, Actuals: reserve-rate, pumped hydro).
   * Note: extended features for 2024–25 deferred to backlog.
5. **Train/validation/test split**
   * Calendar + hard cutoff (**31-09-2023** test start); region handling.
6. **Backtest protocol & metrics**
   * Walk-forward / bucketed evaluation; MAE/MSE/R², IC, hit-rate, PnL, t-stat, ESS.
7. **Costs & trading constraints**
   * Transaction costs, borrow/shortability, position limits, execution assumptions.
8. **Reproducibility notes**
   * Seeds, environment, data versions.

Modelling procedure

We know look at constructing new models for the day ahead to imbalance spread. In general, the following modelling procedure was followed.

We started with general exploratory data analysis this consisted of analysing time series, bin plots different interactions over time weeks, months, and yearly. Then we would look at a relationship more closely and construct a hypothesis for why it existed before starting some testing.

Final after repeating this process for many features we would construct a model, examine its residuals and see if there are any more features which could be predictive / useful.

Once a subset of features was chosen, we ran, mutual information / correlation / tree gains feature selection methods to see if there were any useful features which were left out.

Lastly after testing multiple models on the training set we examined the performance of one final model on the testing set.

Signals

We categized the signals into 4 broad based categories

Time based

The first signal to be found useful, is the hour of day signal this was shown earlier and has a clear trend throughout the year. This is informative as a feature but also has consequences later as any other feature such as solar, rdl, ect which have this shape will also have a high predictive power on its own but may not add to the analysis.

Next, we have time-based features such as if it is a holiday or is it a Sunday, both for the current day and the day before. On the days after these days reserves will be filled to a higher level due to cheap energy in the spot prices. This leads to a negatively biased imbalance price for the entire next day.

Price Based

The second feature which is useful is the EMWA price feature this was the original one shown earlier

Next, we have spot price data, the spot price data only must be lagged a day, it gives a good sense of the current bidding strategies, we have detrended it by taking away the mean of the spot to give a relative spot price within the day. We can see values which are largely likely around 5 – 7 have a high negative imbalance and values low around 12 have high positive imbalance. (so, explaining some of the hourly trend, but changes week on week)

Actuals

Hydro again has important both intraday and the level it is storing from the previous few days,

Forecast Based (Data Vendor)

Next is the rdl difference this tells us for a specific 30-minute increment if rdl is increasing or decreasing we can see it is negative trend which means as rdl gets higher the imbalance becomes negative this is likely an overcorrection in the spot market due to forecasted high rdl

Temperature is a key feature in Japan due to its climate there are two types of features one is due to the trend is weather is it increasing or decreasing and how this effects the spread. (increasing temperature difference shortens reserves leading to a higher imbalance than the spot closes). The second is a pure feature which is if it is extremely hot or cold which can happen in Japan this effects the imbalance price in Japan.

Wind, wind is the most variable forecast, we quantify the level of uncertainty in the wind by getting the sum of the difference between the ec00 and ec12 forecasts. We can see if the forecast is overestimating or underestimating the wind and how that has influenced past Imbalance spreads which we use in our model.

Forecast Based Local

Reserve rate – As detailed earlier and in the appendix, Japan moves to an alternate pricing system when the Imbalance price is hit. We have a model detailed in additional documentation about how this model works – below shows the difference in spread

Region Coupling – The wide area that a region is considered in has a large impact on the imbalance price, this is because the different regions have different compositions of energy stack so therefore a place like Tohoku which neighbours Tokyo spread increase and decrease drastically as it is linked or not linked. We have a model for predicting the coupling between regions.

Hydro – people are overbuying later in the evenings they do not need this amount of demand – therefore utilities demand less ->

It means that In order to satisfy this demand hydro over sells in the spot market, it buys up to a point then tries to buy as much more as possible within the day until the difference between the imbalance price to buy at 12 and the spot price to sell at 6 is greater then the spot price at 12 to the spot price at 6

**Modelling procedure**

We build new models for the day-ahead → imbalance spread by iterating from exploratory patterns to simple tests and then to a combined model.

* **Exploration.** Start with EDA: time-series views, bin plots, and interactions across hours, weekdays/holidays, months, and the full year.
* **Hypothesis → quick test.** When a relationship looks real, write a short hypothesis for *why* it exists and run a small, leakage-safe test.
* **Residual check.** After a first pass model, examine residuals by hour/weekday and around notable days (spikes, floor hits) to uncover missing effects.
* **Feature screening.** From the candidate set, use mutual information, correlation (conditional on hour), and tree gain/importance to see what truly adds value and what is redundant.
* **Finalize for holdout.** After trying a few model forms on the training period, evaluate one **final** specification on the test period.

*(All tests use only information available at decision time; forecasts are used with their proper cutoffs; previous-day spot/imbalance are lagged.)*

**Signals**

We group signals into the same broad buckets you outlined. Each item notes what it captures and how we build it at a high level.

**Time-based**

* **Hour of day.** Clear, stable intra-day shape visible through the year. Useful as a standalone predictor and as a *conditioning* factor; many other features share this shape, so we de-mean/encode by hour to avoid double-counting.
* **Calendar (holiday/Sunday and day-after).** Days following Sundays/holidays tend to see reserves refilled on cheaper spot, biasing the next day’s imbalance lower. We include simple flags for these cases (both the day itself and D+1).

**Price-based**

* **EWMA of spread (baseline).** The original signal: EWMA with ~10-day half-life by 30-minute bucket.
* **Previous-day spot profile (relative within-day level).** Use yesterday’s spot, de-meaned within its day (or ranked) to get a relative profile of bidding. Higher relative spot levels tend to precede more negative imbalance; lower levels the opposite. This helps explain part of the hourly pattern but does vary week-to-week.

**Actuals**

* **Pumped hydro (level and intra-day behavior).** Both the intra-day pumping/selling pattern and the storage level over recent days matter. We include current state/trajectory indicators and simple ramp proxies.

**Forecast-based (data vendor)**

* **Residual demand/load (RDL) change.** The RDL *difference* by 30-minute bucket signals tightening/loosening; we observe a negative relationship (higher forecast RDL → more negative imbalance), likely from spot over-correction.
* **Temperature (trend and extremes).** Trend in temperature (rising/falling) affects reserves intraday; extreme hot/cold regimes in Japan also shift imbalance pricing. We capture both the trend and “extreme” buckets.
* **Wind (uncertainty & bias).** Wind is the most variable forecast. We quantify uncertainty via the EC00 vs. EC12 difference (|EC00−EC12|), and we track whether forecasts tend to over/under-estimate and how that lined up with past spreads.

**Forecast-based (local market structure)**

* **Reserve rate.** When the reserve-rate threshold is hit, Japan switches pricing mechanics, which materially shifts the spread. We use the reserve-rate information (as described earlier / in the appendix) to condition the signal and show the spread differences under those states.
* **Region coupling.** Wide-area linkage (e.g., Tokyo with Tohoku) changes the effective stack. Because neighboring regions have different mixes, coupling on/off can swing spreads. We include a simple coupling indicator based on historical linkage patterns.

Note: in the subsequent section the performance including outputs from these models is left in the index, as implementing would require a combination of models

OK now next section, so using these features two models were chosen, the first is a Weighted Bayesian regression and the second is a Light GBM model.

Weighted Bayesian regression

We found the model performed best when the following features were chosen EMWA price feature, spot price detrended, hour of day, Holiday and Sunday Indicator, difference in temperature,

The covariance and decay parameters were chosen based off intuition / heuristics of model in similar zones and then validated against a walk forward cross validation on the training set which had similar agreement in parameter choices.

Bin plots of performance

PNL plots of performance

Rolling Betas for both Tokyo and Kansai

We can see a clear improvement of the model over the base EMWA model

Light GBM

The next model that was analysed was a Light GBM model, the Light GBM used all of the above features outlines earlier. Two different training methods were used the first was a walkforward cross validation on the depth of tree, no . estimators, min samples per leave etc. the second was a walk forward cross validation on a fixed parameters with a fairly large tree and the reg\_lambda parameter was cross validated.

Bin plots of performance

PNL plots of performance

Rolling Betas for both Tokyo and Kansai

Summary of Models

The following table summarizes the best 3 models from looking across all metrics and cost efficiency of implementing it seems that the light gbm model performs the best.

**Model candidates**

We tested two models built from the features above: a **Weighted Bayesian regression** and a **LightGBM** model. Both were evaluated in a leakage-safe, walk-forward setup and compared to the baseline EWMA.

**Weighted Bayesian regression**

**Feature set (final):**  
EWMA spread (by bucket), previous-day spot (de-trended within day), hour-of-day, Holiday/Sunday flags (incl. day-after), and **ΔTemperature**.

**Estimation approach:**  
Covariance and time-decay parameters were initialized from heuristics (similar regions) and then **validated by walk-forward cross-validation** on the training period. The chosen parameters showed consistent picks across folds.

**Diagnostics:**

* [Figure: Bin plots — Weighted Bayesian vs. baseline]
* [Figure: Cumulative PnL — Tokyo & Kansai]
* [Figure: Rolling betas — Tokyo & Kansai]

**Result in brief:**  
Clear improvement over the pure EWMA baseline, with more stable betas by hour/season and better conditioning around floor/holiday regimes. Interpretability remained high (signs match hypotheses).

**Notes:**  
Works well for additive effects and smooth trends; less flexible for non-linear interactions (e.g., hour × reserve-rate proximity).

**LightGBM**

**Feature set:**  
All features outlined earlier (time-based, price-based, actuals, and vendor forecasts), including those used in the Bayesian model.

**Training regimes tested:**

1. **Walk-forward hyperparameter search** over tree depth/num\_leaves, n\_estimators, min\_data\_in\_leaf, etc.
2. **Large fixed tree** with only **reg\_lambda** cross-validated via walk-forward (regularization-only tuning).

**Diagnostics:**

* [Figure: Bin plots — LightGBM vs. baseline]
* [Figure: Cumulative PnL — Tokyo & Kansai]
* [Figure: Rolling betas / SHAP-style stability]

**Result in brief:**  
LightGBM captured non-linearities and interactions (e.g., hour with spot profile/temperature) and **outperformed** the Bayesian model on most error metrics and after-cost PnL, with month-to-month stability comparable to or better than the Bayesian model.

**Notes:**  
Slightly higher operational complexity; feature importance helps interpretability, but not as transparent as regression betas.

**Summary of models**

[Table: **Top 3 specifications**]  
*Columns:* Model | Feature set (short) | Validation method | Test MAE | Test R² | After-cost PnL (Tokyo/Kansai) | Stability (% positive months) | Notes

**Takeaway:**

* **Winner:** *LightGBM* — best overall across metrics and cost efficiency to deploy.
* **Benchmark:** *Weighted Bayesian* — strong, interpretable baseline that improves on EWMA and provides clear sanity checks (rolling betas).
* **Control:** *EWMA* — retained for monitoring and drift detection.

If you want, I can also write a one-paragraph “Model selection rationale” to drop under the table (why LGBM is chosen now, and when we’d fall back to Bayesian).

Ask ChatGPT