

# Forecasting the 2-Hour-Ahead mFRR–Day-Ahead Spread in NO1:

Predictability Limits, Signal Decay, and the Value of Statistical Baselines

Vilijam Cekov  
*GRA 4157 Data Science Report*

December 9, 2025

## Abstract

When it comes to the Nordic power system, balancing markets are essential in order to maintain reliability, in terms of price and availability. However, forecasting price deviations in real time still remains operationally challenging. As a result of Day-Ahead (DA) prices being announced at 13:00, the relevant metric for short-term operational risk is the spread of the Manual Frequency Restoration Reserve and the Day-Ahead (DA) price. My project researches how predictable this spread is in the NO1 bidding zone (2020–2025), by utilizing historical prices, hydrological indicators (reservoir filling and Hydro Pressure Index, which is a z-score), and calendar features, due to the cyclical nature. I evaluated three algorithms: Ridge Regression (linear), Random Forest (bagging), and XGBoost (boosting). I analyzed them under a "no-lag-1" in order to reflect realistic operational information gaps. In regards to statistical values, while all my models reduce RMSE relative to that of a naive baseline, my models were incapable of capturing the volatility and instead were reverting to the conditional mean. My analysis of feature importance, cemented my findings by confirming that the predictive signal faced rapid decay after one hour. This left my models relying on slow-moving seasonal baselines. While all models reduce RMSE relative to a naive baseline, they fail to capture volatility, reverting instead to the conditional mean. Feature importance analysis confirms that the predictive signal decays rapidly after one hour, leaving models reliant on slow-moving seasonal baselines. Conclusively, the spread may be structurally unpredictable during imbalances that can be seen as extreme. At very least not without granular, real-time casual data. However, the models I made have the potential to serve as viable statistical baselines to detect anomalies.

## 1 Introduction

A critical role is taken up by the balancing markets in maintaining reliability in the Nordic power system. The mFRR price is a reflection of the real-time system imbalances, unexpected outages, forecasting errors, as well as cross-border transmission constraints. By contrast, the

DA price is fixed once per day at 13:00, and as such it is unable to reflect hourly operational changes in a meaningful way. As a consequence for operational decision makers and assessing short term risk, the relevant forecasting target is the spread between the real time mFRR and the fixed DA price.

My project is constrained by the strict knowledge horizon of the Nordic market and it is therefore limited to the nordic regions, and we do not connect with demand that originates outside of Norway via interconnected electricity markets (i.e. EU, GB) as it is out of scope. To provide further context, I confirmed the absence of fundamental hydrological supply stress during our test period (late November-early December, 2025):

Table 1: **NO1 Hydrological Context (Latest Week)**

Metric	Value	Interpretation
Reservoir Fill	73.5%	High absolute level
% of Seasonal Normal	95.4%	<b>Normal / Slight Deficit</b>
HPI (Z-Score)	-0.26	No Hydrological Stress

My primary concern is not to forecast the long-term supply for scarcity, but to predict the short-term system shocks (wind/outages/ or in regards to hydro, "Dry" seasons) in a system that could otherwise be considered stable. Central to my work is the enforcement of a operational constraint that strictly excludes lag-1 information. As the mFRR price at time  $t + 1$  is not know at time  $t$ , as such the first available AR feature is lag-2.

The constraint is concerned with a specific gap in the existing literature. While extensive research does exist on the forecasting electricity prices, genarally studies rely on AR (Autocorrelation) features that assume immediately available data. These models can often overstate their predictability if they ignore the information delays that are often faced by real-time operators. As a result of explicitly excluding lag-1, my study strives to evaluate the *true* operational spread. By identifying natural boundaries where historical signals decay is greeted by short-term market efficiency. I evaluated models of varying complexities. Which include: Ridge Regression, Random Forest, and XGBoost. So as to be able to reason whether nonlinear or booster learners could extract additional structures for prediction from the available data. As is stated by Géron (5), it is essential to compare algorithms with varying inductive biases (variance reduction vs bias reduction), in order to diagnose if errors are a result of model limitations or data limitations.

## 2 Data and Feature Engineering

### 2.1 Data Sources

The figures and tables, are acquirable by running the scripts as instructed in the README.md file on my github repository(1). The analysis is based on hourly market data for the NO1 bidding zone (Oslo) spanning January 1, 2020, to November 15, 2025. The dataset integrates data from two primary sources:

1. **Nord Pool Data Portal(2):** Hourly time-series data for mFRR imbalance prices and Day-Ahead (DA) spot prices. I acquired access by submitting a access request for education purposes, afterwards I downloaded the datasets in CSV format for mFRR and DA via the report tab for each year (2020-2025). Note that this data is classified and I am not allowed to share it. You can use alternative sources if you have a csv file with the same shape/headers.
2. **NVE (Norwegian Water Resources and Energy Directorate)(3):** Weekly reservoir filling data and historical seasonal norms. I acquired this data by going to the link in the references, and then clicking on "GET /api/Magasinstatistikk/HentOffentligData" which I then pasted the URL in my data\_loader.py and used it to call the NVE api.

## 2.2 Feature Engineering Pipeline

When preparing data for the machine learning pipeline (Figure 1), I applied several transformations, by applying standard time-series handling techniques described by mckinney (4):

- **Target Definition:** The target variable is the spread  $s_t = \text{mFRR}_t - \text{DA}_t$ . The forecasting objective is  $\hat{s}_{t+2}$ .
- **Hydrology Features:** I constructed the **Hydro Pressure Index (HPI)**, which is just a standardized Z-score of the current reservoir level against its historical seasonal normal:
$$\text{HPI}_z = \frac{\% \text{ vs Normal} - \text{Mean}(\% \text{ vs Normal})}{\text{Std Dev}(\% \text{ vs Normal})}$$
- **Cyclical Encoding:** I transformed calendar features (hour, month) using sine/cosine encoding to preserve temporal proximity.
- **Lag Generation:** Consistent with the operational constraint, I generated lags starting from  $t - 2$  ( $s_{t-2}, s_{t-3}, s_{t-4}$ ).

## 3 Methods

### 3.1 Models and Objectives

To satisfy the requirement of testing different algorithm types and have some form of cross-validation, I evaluated three ML paradigms:

1. **Ridge Regression (Linear):** It minimizes squared error with  $L_2$  regularization and assumes a linear relationship while focusing on reducing variance (5) via coefficient shrinkage.

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N (y_i - x_i^T \beta)^2 + \alpha \|\beta\|_2^2 \right\}$$

2. **Random Forest (Bagging):** An ensemble method which averages  $K$  independent decision trees. It’s effectiveness lies in capturing non-linear interactions, simultaneously reducing variance without increasing bias (5).

$$\hat{y}(x) = \frac{1}{K} \sum_{k=1}^K \hat{T}_k(x)$$

3. **XGBoost (Boosting):** A gradient boosting framework which builds trees sequentially, to correct prior errors. It is designed to minimize bias, while often achieving state-of-the-art performance in tabular tasks (5).

$$\mathcal{L}(\phi) = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

## 3.2 Evaluation Strategy

In order to be mindful of the temporal order of the data and prevent information leakage, I employed a strict **Chronological Hold-Out** strategy (4). Standard cross-validation is unsuitable as it destroys temporal dependence.

I split the dataset into two chronologically distinct segments:

- **Training Set ( $\approx 80\%$ ):** Earlier portion of the time series (Jan 2020 – Late 2024) to fit the models.
- **Test Set ( $\approx 20\%$ ):** Final portion of the data (Late 2024 – Nov 2025), serving the role of a strict hold-out set for performance evaluation.

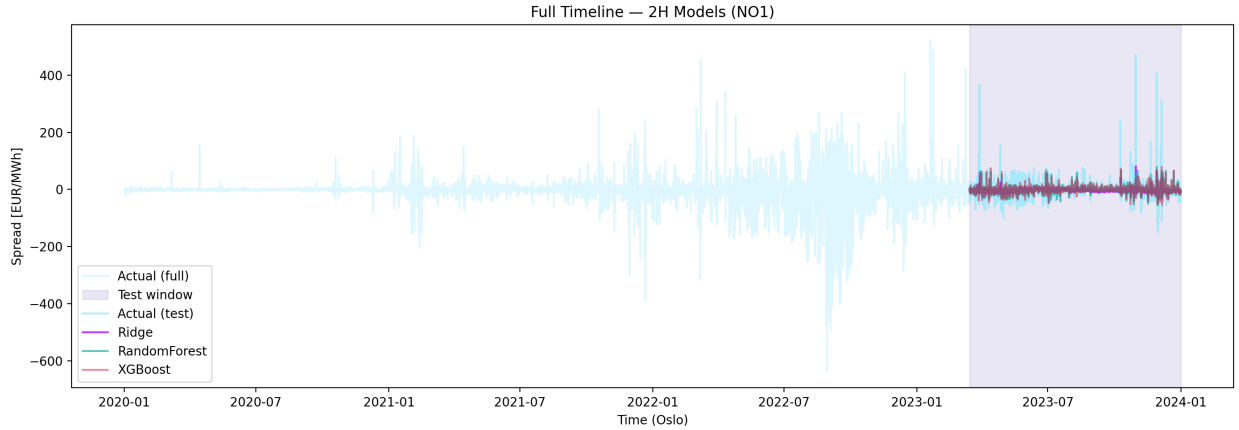


Figure 1: **Chronological Hold-Out Timeline.**

The full timeline (Figure 2) places the hold-out period in context, revealing a relatively volatile market regime. The plot demonstrates a clear tendency for all learners to track the center of the distribution while failing to capture extreme spikes—a known limitation of point forecasting in heavy-tailed distributions. This behavior is further highlighted in the 7-Day

Zoom, which shows residuals that are strongly mean-reverting. The models effectively "hug" the zero line, diverging from reality primarily during sharp price jumps. While the non-linear models (Random Forest and XGBoost) reduce peak errors slightly better than Ridge, they still miss the largest excursions. This visual evidence corroborates the minimal performance differences observed in Table 2, suggesting that the models are minimizing global error by playing it safe rather than learning the dynamics of the spikes.

## 4 Results

### 4.1 Numerical Performance

In Table 2 I summarize the performance metrics for the 2-hour-ahead forecast on the hold-out set.

Table 2: 2h-Ahead Test-set Performance Metrics for NO1

Model	MAE	RMSE	$R^2$
Naive Baseline (Lag-2)	16.41	31.64	0.000
Ridge Regression	12.90	24.02	-0.016
Random Forest	12.13	23.20	0.053
XGBoost	12.04	23.00	0.069

Across the board, the ensemble methods (Random Forest and XGBoost) deliver the lowest error rates on the hold-out set, modestly outperforming the linear Ridge model and offering a clear improvement over the naive lag-2 baseline. However, the  $R^2$  values remain strikingly low—hovering near zero or even dipping into negative territory. This indicates that while the models successfully minimize average error (RMSE) compared to a naive guess, they struggle to capture the inherent hour-to-hour variance of the spread at this horizon. The negative  $R^2$  values in particular signal instances where model performance degrades below that of a simple historical mean, a behavior consistent with data characterized by heavy-tailed noise and rapid signal decay.

Table 3: 3h-Ahead Sensitivity Analysis Metrics for NO1

Model	MAE	RMSE	$R^2$
Naive Baseline (Lag-2)	16.99	31.32	0.000
Ridge Regression	13.06	23.96	-0.010
Random Forest	12.08	22.98	0.070
XGBoost	12.13	22.80	0.085

As the forecast horizon extends to three hours, these marginal gains effectively evaporate. All models see their  $R^2$  scores collapse to near or below zero, with MAE and RMSE values converging across algorithms. This performance drop-off confirms that exploitable

temporal dependence is fleeting, decaying rapidly beyond the two-hour mark. This aligns with our feature importance analysis, which suggests that once the immediate autoregressive lags are removed or become stale, the remaining calendar and hydrological features provide insufficient structure to support accurate forecasting.

## 4.2 Visual Performance and Structural Failure

My Time-series plots on the hold-out set seem to reveal that all models “flatline” near zero during periods, where the actual spread exhibits characteristics of high volatility. Figure 2 illustrates a specific event on November 14th where the spread spiked to  $\approx 250$  EUR/MWh.

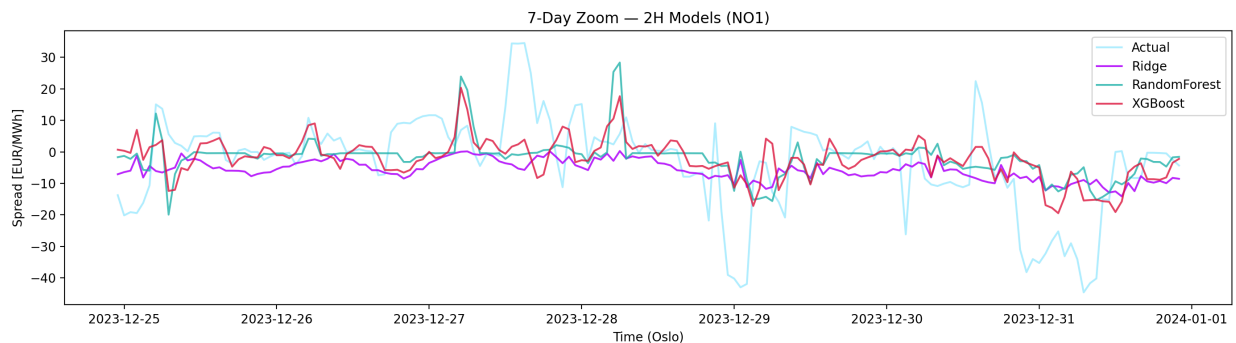


Figure 2: **Actual vs. Predicted Spread (Test Set Zoom).** In Figure 2 I illustrate a 7-day zoom of the hold-out window (Dec 25–Jan 1). We see that the actual 2-hour-ahead mFRR–DA spread (blue) as it has several sharp volatile swings. All three of my models (Ridge, Random Forest, XGBoost) hover close to zero and seemingly have no reaction to the spikes, this reflects their mean-reverting behavior under the no-lag-1 constraint. All of this, showcases that the models reduce RMSE by suppressing volatility instead of predicting direction during high-variance periods.

## 4.3 Signal Decay

The inability to forecast spikes is explained by signal decay. With lag-1 removed, the feature importance shifts dramatically.

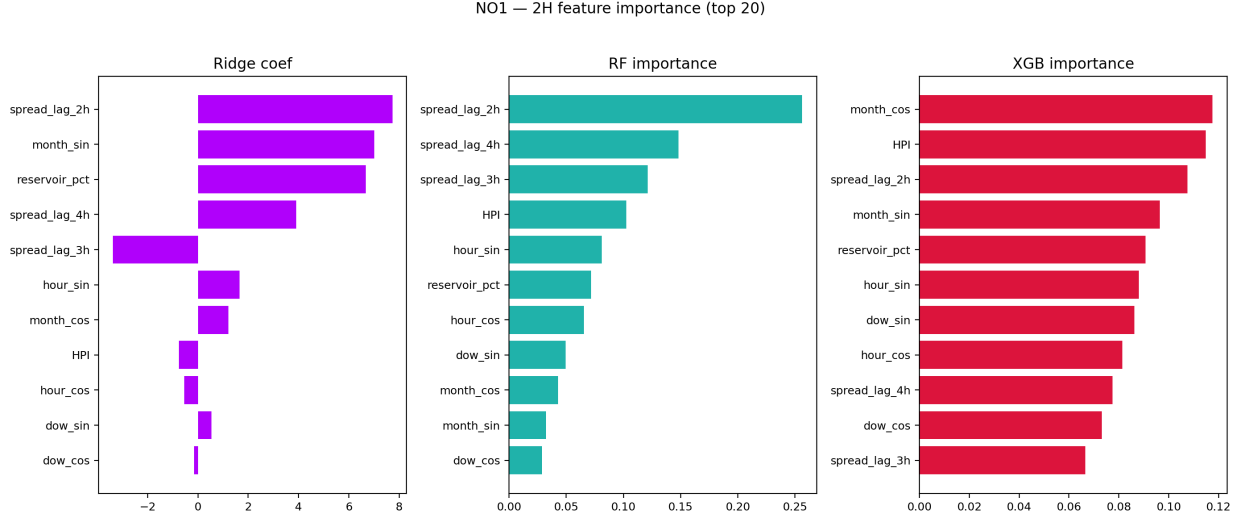


Figure 3: **Feature Importance.**

Figure 3 illustrates the dominance of the most recent available lag (`spread_lag_2h`), with lags 3 and 4 playing a secondary role. Calendar seasonality (month and hour harmonics) and hydrology (`reservoir_pct`) contribute smaller, stable gains but lack the predictive power to drive short-term accuracy. For the tree-based models, these importance scores reflect association within the ensemble rather than direct causality. In the case of Ridge, while the coefficient signs are theoretically interpretable due to feature standardization, their magnitudes are biased by the regularization penalty. Collectively, these plots describe a short-memory process driven by immediate past prices, with only weak, smooth modulation from seasonal or hydrological factors.

## 5 Discussion

My project demonstrated that machine learning models operating under these constraints do not necessarily provide a meaningful prediction in regards to the magnitude of the 2-hour-ahead mFRR-DA spread. Even though Ridge and Random Forest achieve lower RMSE than the naive baseline, that improvement is driven by **volatility suppression** rather than directional accuracy.

My results exemplify the Bias-Variance tradeoff (6). As the naive baseline suffers from high variance (predicting a spike 2 hours late). Ridge Regression goes for high bias and predicting the conditional mean (near zero). Furthermore, the failure of XGBoost to improve upon Ridge solidifies that this is a data limitation, not necessarily a model limitation. Without real-time causal drivers (e.g., wind forecast errors) or lag-1 information, the spread is observed to be structurally unpredictable.

However, these models provide value as **statistical baselines**. By accurately modeling the “fair weather” spread, the residuals serve as clear indicators of grid stress.

## 6 Conclusion

My project set out to determine whether the 2-hour-ahead mFRR–DA spread could be effectively forecast using only historical market and hydrological data. Our results provide a decisive answer: under the strict operational constraint of removing lag-1 information, the spread is structurally unpredictable during periods of extreme volatility.

Regarding model complexity, I rejected the hypothesis that advanced non-linear algorithms would unlock hidden predictive signals. While Random Forest and XGBoost statistically outperformed the naive baseline in terms of RMSE, they did so by aggressively suppressing volatility and reverting to the conditional mean. The fact that these complex ensembles failed to significantly outperform the linear Ridge regression confirms that the limitation lies in the data’s information horizon, not in the modeling capacity.

Ultimately, we conclude that while these models cannot serve as active trading tools for spike prediction, they provide significant operational value as statistical baselines. By establishing a robust “fair-weather” expectation for the spread, the model residuals become powerful, real-time indicators of grid stress, flagging anomalies that deviate from the historical norm. Future forecasting efforts must therefore shift focus from purely autoregressive approaches to integrating real-time causal drivers, such as wind power forecast errors and transmission outages.

## References

- [1] GitHub Repository :<https://github.com/Valiant-GitHub/GRA-4157-Data-Science-Project>
- [2] Nord Pool Data Portal, <https://data.nordpoolgroup.com>. Used for mFRR and DA price data.
- [3] NVE Magazine statistics, <https://api.nve.no/doc/magasinstatistikk>. Used for hydro data via API.
- [4] McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O’Reilly Media.
- [5] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O’Reilly Media.
- [6] Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury.