AARHUS UNIVERSITY

DEPARTMENT OF ECONOMICS AND BUSINESS ECONOMICS

ELECTRICITY PRICE FORECASTING IN THE ERA OF FOUNDATION MODELS

A COMPARATIVE ANALYSIS OF CLASSICAL TIME SERIES FORECASTING, MACHINE LEARNING AND FOUNDATION ZERO-SHOT FORECASTING IN EUROPEAN DAY-AHEAD MARKETS

Andreas Mørk Jensen (202108364), BSc in Economics

Paw Eskildsen Castellanos (202306176), BSc in Economics

Supervisor: Erik Christian Montes Schütte

1. May 2025

Bachelor Thesis

111407 Characters (Incl.Spaces) = 46,42 Pages

TABLE OF CONTENTS

Abstract

This thesis provides a comparative analysis of classical time series models, machine learning and zero-shot foundation models for forecasting European Day-Ahead electricity prices. Driven by the increasing integration of renewable energy, modern electricity markets exhibit extreme volatility and non-linear dynamics that have challenged traditional forecasting methods. Using hourly data from 2020 to 2024 across six European bidding zones, this study benchmarks twelve models. The empirical results demonstrate that classical linear models are insufficient for capturing volatile merit-order dynamics. While state-of-the-art multivariate foundation models achieve highly competitive point and probabilistic forecasts, though they present a memorization bias risk. Localized machine learning models consistently demonstrate a robust and competitive performance, particularly outperforming foundation models in the highly volatile bidding zones.

1. Introduction

Forecasting Day-Ahead electricity prices accurately is a complex and technically demanding problem in modern economics. Diverse market participants from utility companies and powerplants to financial traders depend on the ability to predict the price accurately, for optimizing bidding strategies, hedging risk and allocating capacity efficiently. A systematic failure of forecasting aggregated across millions of contracts, means misallocated capital, non-optimal dispatch decision and higher risk in the European Day-Ahead market (Lago *et al*, 2021).

The environment in which forecasting models must operate within, has changed in the last couple of years. The rising integration of renewable electricity sources has induced a structural break in the European Day-Ahead electricity markets, although some bidding zones more than others. Prices have transitioned from being a relatively predictable to showing highly complex nonlinear relationships, volatility clustering and often negative prices (Seitaridis, Thomaidis and Biskas, 2021). The consequence of this is that the prices have become more nonstationary, heteroskedastic, with large kurtosis, properties that collectively challenge the classical econometric assumptions under each model (Weron, 2014a). While these classical models have captured the seasonal cycles in electricity prices, they have ultimately become insufficient in the face of the transitioned and highly complex

Day-Ahead electricity market. The limitation of this has pushed the literature and research to implement other methods, such as machine learning models, which relax the assumptions and can understand nonlinear relationships (Lago *et al*, 2018). Most recently the emergence of time series foundation models has introduced the possibility of zero shot forecasting, in which a pretrained model can generate forecasts without having to train the model on your specific data (Ansari *et al.*, 2024a; Garza, Challu and Mergenthaler-Canseco, 2024a). Because of this rapid evolution, a research gap emerges. This thesis will fill this gap by empirically benchmarking, classic time series models, machine learning and foundation models.
Using hourly data from January 2020 to September 2024, this study evaluates twelve different forecasting models, two seasonal naïve benchmarks, SARIMAX, MSTL-ARIMAX, ETS, CatBoost, NHITS, Time-GPT, Chronos, Chronos V2, TimesFM and Lag-llama. Evaluating these models across six European bidding zones, West Denmark, Southern Norway, Germany, France, Poland and Spain.

This thesis is structured as follows. Chapter 2 provides a critical review of relevant literature within the field. Chapter 3 introduces the methodology of this thesis. Chapter 4 presents the data analysis. Chapter 5 introduces the theoretical background within each model and establishes the performance evaluation of each model across bidding zones. Chapter 6 provides a cross-model comparison for each bidding zone. Finally, this thesis concludes on our findings.

2. Literature

The ongoing integration of renewable electricity sources has induced a fundamental structural break in the data generating process of European Day-Ahead electricity markets. As zero-marginal cost increasingly dictates market clearing, prices have transitioned from exhibiting predictable load-driven seasonality to demonstrating profound non-linearities, extreme volatility clustering and almost not seen before negative prices (Seitaridis, Thomaidis and Biskas, 2021; 'ACER-CEER 2024 Market Monitoring Report', 2024). Therefore, economic literature surrounding electricity price forecasting has been forced to evolve rapidly to capture these complex dynamics (Seitaridis, Thomaidis and Biskas, 2021).

This chapter evaluates the methodological evolution, tracing the progression from classical times series models to machine learning models and the current state-of-art large-scale foundation models. However, before assessing the empirical efficacy of these models, it is important to ground the forecasting challenges in their underlying microeconomic reality beginning with the primary driver of price volatility: the merit order effect.

2.1 Economic Dynamics of European Electricity Markets

2.1.1 The merit order effect

The fundamental driver of the price in European Day-Ahead markets is the merit-order effect. In a uniform price auction, the market clearing price is determined at the intersection of aggregate demand and supply curve, where the units generating electricity is introduced in an ascending order based on short run marginal costs. As renewable electricity sources are near-zero in marginal cost they are placed at the bottom of the merit order curve (Sensfuß, Ragwitz and Genoese, 2008)

Consequently, high forecasted amounts of renewable electricity generation, shifts the aggregate supply curve to the right. This displaces expensive, thermal generation and puts significant downward pressure on the market-clearing price, which is documented extensively across integrated European bidding zones (Cludius *et al.*, 2014). The econometric challenges do not arise from the falling average prices, but instead from the increasing price variance. Because renewable electricity generation is highly weather dependent and exogenous to market price signals, the supply curve is subject to severe, high-frequency shifts.

As (Ketterer, 2014) demonstrates, the increasing use of renewable electricity generation significantly amplifies price volatility. During periods of high renewable electricity generation and low inelastic demand, the price can implode. The high technical costs associated with ramping down and restarting conventional baseload generation, leads operators to frequently submit negative bids to remain synchronized with the market. This dynamic introduces extreme non-linearities, price spikes and negative prices (Weron, 2014a). For forecasting this implies that classical linear assumptions are often violated, necessitating forecasting algorithms that are capable of capturing complex non-linear relationships between forecasted weather variables and market-clearing prices.

2.1.2 Market efficiency and non-storability constraint

In standard financial econometrics, the weak form Efficient Market Hypothesis (EMH) says that asset prices reflect all historical information, causing prices to follow an approximate random walk (Fama, 1970). This efficiency is enforced through intertemporal arbitrage: if a predicable price difference exists between two time periods, rational agents purchase and store the asset, this effectively smooths out the price curve. However, electricity violates this assumption due to the strict physical property, it is currently economically non-storable on a large scale (Bessembinder and Lemmon, 2002). The lacking ability to store electricity necessitates that supply and demand must be balanced at all times, this decouples electricity prices from standard financial assets dynamics, so existing financial models fail when considering electricity prices due to the mean reversion and persistence (Knittel and Roberts, 2005).

Econometrically this lack of storability gives electricity two statistical properties that dictate the choice of forecasting models. Rather than following a random walk, Day-Ahead electricity prices exhibit mean reversion. Sudden shutdowns and extreme weather phenomena can cause severe price spikes, the lack of storage suggests that these shocks are temporary, the prices will revert to the short-run marginal cost of the generation mix after the shock (Knittel and Roberts, 2005). Because market-clearing is linked to the instantaneous consumption of the electricity, prices exhibit predictable, deterministic seasonal frequencies of daily, weekly and annual cycles (Lucia and Schwartz, 2002). Therefore, any forecasting model must be explicitly configured to capture these seasonal components and mean reversion.

2.1.3 Market coupling and spatial integration

The merit-order effect and non-storability constraint drive price dynamics within a single bidding zone, the European Day-Ahead market is a fundamentally spatial integrated network (Zachmann, 2008). Through the Price Coupling of Regions framework, the EUPHEMIA algorithm optimizes cross-border capacity allocation to maximize total economic surplus (NEMO, 2020). As long as the interconnection infrastructure is not constrained, this algorithm forces prices across connected bidding zones to converge (Böckers and Heimeshoff, 2014). Therefore, a shock in one bidding zone will affect the prices in neighboring bidding zones.

However in reality the interconnections between zones are frequently subject to thermal and operational constraints (ACER, 2023). When the optimal electricity flow between bidding zones exceeds the available Net Transfer Capacity (NTC), the interconnection is not available, when this happens the markets decouple and the price formation in each bidding zone becomes isolated, driven only by localized supply and demand inelasticities (Keppler, Phan and Pen, 2016). For the purpose of forecasting across different bidding zones, this decoupling presents econometric challenges. Demonstrated by (Ziel and Weron, 2018) that univariate forecasting methodologies, that exclude these interconnectors are subject to omitted variable bias. To capture the complexities of these interconnected bidding zones, one must transition to using multivariate models.

2.2 Methodological evolution of electricity price forecasting

2.2.1 Classical Econometric and Time Series Models

The foundational literature in electricity price forecasting was previously dominated by classical linear time series models. Early empirical applications demonstrated that Auto-Regressive Integrated Moving Average (ARIMA) models and their variants (SARIMA), where effective at capturing the highly seasonal periods present in electricity demand (Nogales *et al.*, 2002; Contreras *et al.*, 2003). However, these purely univariate models ignore the physical and economic drivers of the market for electricity (Karakatsani and Bunn, 2008), therefore later literature transitioned to ARIMAX and SARIMAX models. By integrating covariates these models provided a much better representation of the complex electricity market (Zareipour *et al.*, 2006).

However as the integration of more renewable electricity expanded, the ARMA models, that assume linear relationships between variables and rely on the normality of the residuals, struggle to capture the volatility clustering and negative prices found in the modern markets (Weron, 2014a). Because of the lacking ability of the classical time series models, there has been a shift to more algorithmic learning models (Lago *et al,* 2018).

2.2.2 Gradient Boosting and Categorical Dynamics

To circumvent the parametric constraints of the classical time series models, the electricity price forecasting literature has increasingly adopted algorithmic learning models, specifically the ensemble decision trees. While comprehensive benchmark studies, such as (Lago *et al*, 2021), emphasize deep neural networks and LEAR methods as standard baselines, recent literature has demonstrated the distinct advantages of Gradient Boosted Decision Trees in handling the non-linear volatility of Day-Ahead electricity markets (Yıldız, 2023).

2.2.3 NHITS and Foundation Models

While tree-based ensembles like CatBoost effectively model non-linear relationships, they are fundamentally designed to process tabular cross-sectional data, requiring manual construction of autoregressive lags (Lago *et al.*, 2021). To capture deeper time dependent relationships without manual engineering, the electricity price forecasting literature expanded to Deep Neural Networks, specifically sequential architectures like Long Short-Term Memory networks (Zhang *et al*, 2022). However, early deep learning models were heavily criticized for their computational inefficiency over long horizons and their lack of econometric interpretability (Mashlakov *et al.*, 2021). To resolve the conflict between deep learning performance and computational inefficiency, literature has transitioned to hierarchical architecture, resulting in the Neural Hierarchical Interpolation for Time Series (NHITS) model (Challu *et al.*, 2022a).

The success of neural networks like NHITS paved the way for the most recent methodological advancement in the literature: the Time Series Foundation Models. Unlike the machine learning algorithms, which must be trained exclusively on localized, task-specific data, the foundation models use a transformer-based architecture pre-trained on massive datasets (Bommasani *et al.*, 2022). This massive scale introduces the capability for zero-shot forecasting, where the model relies on pre-trained global temporal dynamics to generate robust predictions on unseen target variables without requiring localized fine-tuning or bidding-zone-specific regressors (Ansari *et al.*, 2024a; Garza, Challu and Mergenthaler-Canseco, 2024a).

2.3 Identification of the Research Gap

The econometric literature surrounding Day-Ahead electricity markets has successfully established the microeconomic drivers of price volatility, moving progressively from linear classical models toward algorithmic models and deep learning models to capture non-linear merit-order dynamics.

Despite this methodological evolution, a empirical gap remains. Existing literature predominantly evaluates these advanced architectures in isolation or restricts cross-sectional analysis to a limited subset of highly correlated bidding zones under normalized market conditions.

This thesis directly addresses this gap. There is currently a distinct lack of comprehensive literature that simultaneously benchmarks classical time series methodologies against both localized algorithmic learning and state-of-the-art Foundation Models across a highly heterogeneous cross-section of European bidding zones. By utilizing classical time series models as a baseline and incorporating exogenous covariates across the highly volatile 2020-2024 period, this research provides a vital empirical comparison of predictive performance at the current frontier of time series econometrics.

3. Methodology

3.1 Data and market mechanics

The foundation of any robust econometric forecasting model relies on the integrity and consistency of its underlying dataset. To evaluate the predictive performance of classical time series models against modern machine learning and foundational models, the dataset must encompass periods of market stability and high volatility. This thesis relies on a multi-dimensional high frequency dataset of hourly observations to model the complex spatial and temporal dynamics of the European Day-Ahead electricity market.

All data preprocessing, model training, and forecasting evaluations were conducted in Python. To ensure full transparency and reproducibility, the complete codebase, is available in the attached repository (Jensen, A. and Eskildsen Castellanos, P. 2026)

3.1.1 Time Horizon

The empirical analysis is conducted using a sample period from January 1, 2020, to September 30, 2024. This sample period was selected for two crucial econometric reasons. First, it encompasses a period of significant macroeconomic stress, it specifically includes the 2021-2023 European electricity crisis (ACER/CEER, 2022). Training, testing and evaluating models across this structural break is mandatory to test whether non-linear models can successfully map shifts in the merit-order supply curve.

Second, the dataset is cut off at the end of September to avoid a fundamental structural change in the market structure. In October 2024, NordPool which services a large part of the western European countries, transitioned to Flow-Based Market Coupling (NordPool, 2024b). Extending the dataset beyond October 2024, would violate the assumption of spatial homogeneity. The resulting price convergence dynamics would be altered. By cutting the data in September 2024, the comparability of the models is preserved.

3.1.2 Bidding Zones

To capture the multidimensional nature of the Day-Ahead European electricity market, the dataset encompasses six distinct bidding zones, these zones were partly chosen for their interconnectivity, but also for their differences in generation mix:

- West Denmark (DK_1): The Danish bidding zone was chosen for its high reliance on wind (EMBER, 2024). This serves as a study for price volatility driven by wind-generated supply shocks. DK_1 also works as a link between the Nordic system (NO_2) and the western European systems (DE_LU).
- Southern Norway (NO_2): Was chosen for its high stability due to its reliance on hydro electricity (*Norway Insights*, 2025) , providing the models a more stable region to test against. Notably, NO_2 possesses effectively zero installed solar generation capacity.
- Germany/Luxembourg (DE_LU): As the most liquid market in Europe, it serves as the main price setter of the central western European region (EPEXSPOT, 2024). Germany relies on a vast amount of different generation methods making it a crucial inclusion for analyzing the effects of TTF and CO2 quotas. DE_LU has wind electricity generation as a major component of its generation mix (*SMARD, The electricity market in 2023*, 2023)
- France (FR): The French bidding zone serves as the representative model of a nuclear dominated baseload system (EMBER, 2024), providing a contrast to the renewable electricity heavy zones such as DK_1, DE_LU and ES. A defining characteristic of the French system is

its residential reliance on electric heating (International Electricity Agency, 2021). This implies that the forecasted load variable will have significant explanatory power in the models and will also have an impact on the seasonal component.

- Poland (PL): Poland is chosen to represent the thermal intensive energy generation, given its reliance on coal generated electricity (EMBER, 2024). This characteristic makes PL the zone that is most sensitive to fluctuations in the CO2 quotas.
- Spain (ES): The Spanish bidding zone was chosen due to its unique geographical characteristics. Being almost isolated when it comes to electricity interconnectivity (European Commisson, 2018). Furthermore, Spain's heavy investment in solar electricity (Seitaridis, Thomaidis and Biskas, 2021) allows the test of the models ability to map the highly daily cycles.

The deliberate selection of these six bidding zones ensures that the comparative analysis captures the full spectrum of European merit order. By incorporating regions dominated by wind, hydro, natural gas, nuclear and coal the methodological choice represents a sampling of different marginal cost functions. This structural heterogeneity is critical for evaluating the predictive power and robustness of classical time series models, machine learning models and foundational models.

3.1.3 Variable selection

The forecasting framework is centered around the endogenous variable, Y_t, which is the Day-Ahead electricity market clearing price. To capture the complexities of the merit order, a matrix of exogenous covariates, X_t, is introduced. This matrix consists of forecasted load, forecasted wind and solar production, Net transfer capacity (NTC) between neighboring zones for each zone, Title Transfer Facility (TTF) gas prices and EU ETS CO2 quota prices.

The high volatility of the 2021-2023 European electricity crisis showed that the marginal clearing prices are highly sensitive to macroeconomic covariates (Cevik and Zhao, 2025). Electricity price forecasting literature mandates the inclusion of exogenous regressors to accurately model the shift in the supply curve. However, incorporating these variables must strictly adhere to the ex-ante forward bias constraint.

Because the European Day-Ahead electricity market gate closure happens at 12:00 on day D, the exact informational state of the market at time of gate closure must be respected. Forecasted load, forecasted wind and solar production are published ex-ante, making these valid to use. However, the daily settlement prices for TTF and CO2 quotas are unknown at time of gate closure, these variables must be lagged by $D-1$ to eliminate any forward bias.

3.1.4 Data sources

To ensure reproducibility and avoid manual extraction errors, the entire dataset was constructed using APIs. The data is divided between institutional electricity market data and macroeconomic financial covariates. The institutional data includes the Day-Ahead market clearing price (Y_t), Forecasted Load, Forecasted wind and solar production and corresponding actual data, as well as NTC were all retrieved from European Network of Transmission System Operators for Electricity (ENTSO-E) using their REST API.

The TTF and CO2 quotas were retrieved from Yahoo finance using the python package, *yfinance.*

3.2 Data preprocessing and feature engineering

The raw data retrieved from ENTSO-E is subject to missing data, different time zones and grid anomalies. To ensure mathematical stability for the models, a preprocessing pipeline was implemented using pythons *pandas* library.

3.2.1 Time zone conversion

European bidding zones operate on different local time zones. ENTSO-E data is transmitted using the coordinated universal time (UTC), the forecasting models must use the local time to preserve the localized demand cycles and solar generation cycles and prevent forward bias.

To achieve this the UTC time index was converted to the specific local time zone of each bidding zone. This introduces the problem of daylight savings time, where there would be a 23-hour day in spring and a duplicate 25-hour day in autumn. However, this violates the underlying assumptions of seasonal models. Therefore, the time series was homogenized to a 60-minute frequency using mean-aggregation resampling, systematically identifying and resolving duplicate DST timestamps to enforce a continuous and uniform time series.

3.2.2 Missing values and injections

There are two categories of missing values within the dataset, API transmission errors and generation absences.

For transmission errors such as missing values in NTC or missing values of macroeconomic variables over weekends, a forward fill ($ffill$) insertion is used. This approach is critical for the covariates like TTF and CO2 quotes, carrying Friday's settlement prices through Saturday and Sunday.

For generation absences zero injections were applied. For instance, bidding zones lacking offshore wind capacity like Poland, generated fully missing arrays which were set to zero. Also, Norway which had no solar production, so both the forecasted and actual values required the injection of a zero-value array. This enforces uniform dimensionality across all six bidding zones.

3.2.3 Gate Closure constraint

The European Day-Ahead market operates under the EUPHEMIA clearing mechanism. This imposes strict gate closure at 12:00 on Day D for electricity delivery across the 24 hours of day $D + 1$ (NordPool, 2024a). Forecasting frameworks cannot utilize actual market data from the afternoon and evening of day D. This creates a 12-hour informational blind spot (12:00 to 00:00). To prevent forward bias and data leakage, all models in this study are anchored to this cutoff. Autoregressive lags are restricted exclusively to before noon actuals, while the target horizon relies strictly on Day-Ahead grid forecasts published by the operator prior to gate closure.

3.3 Distributional Transformations and Feature Scaling

3.3.1 Distributional properties

Unlike traditional financial assets, European Day-Ahead electricity prices do not follow a log-normal distribution (Weron, 2014a). Because electricity is economically non-storable at a large scale, the times series exhibits volatility clustering, right skewness and events on the left tail. In bidding zones with a large generation of renewable electricity, the inelasticity of supply during periods of high wind and solar generation frequently drives the market clearing price below zero.

These non-linear properties violate the assumption of constant variance (homoscedastic) and normality of residuals required by the classic time series models. Therefore, a mathematical transformation of the endogenous variable (Y_t) is required to stabilize the variance.

3.3.2 Yeo-Johnson

The Yeo-Johnson transformation is a variance stabilizing transformation. The Yeo-Johnson transformation is unlike the Cox-Box transformation able to handle positive, zero and even negative data which is a critical property when transforming Day-Ahead electricity prices, which as mentioned can have massive spikes, close to zero or negative prices. It achieves this by a piecewise function that shifts the data. This is mathematically defined as:

$$y_t = \begin{cases} \dfrac{(y_t + 1)^{\lambda} - 1}{\lambda} & if\ y_t \geq 0,\ \lambda \neq 0 \\ ln(y_t + 1) & if\ y_t \geq 0,\ \lambda = 0 \\ -\dfrac{(-y_t + 1)^{2-\lambda} - 1}{2 - \lambda} & if\ y_t < 0,\ \lambda \neq 2 \\ -\ ln(-y_t + 1) & if\ y_t < 0,\ \lambda = 2 \end{cases}$$

The specifics of how we are transforming the data is determined by the λ parameter, which is estimated using the Maximum Likelihood Estimation (MLE). To prevent forward bias, the transformation is only applied to the training data, doing this ensures that we do not choose a parameter λ that is calculated on the variance of the testing data. In the context of the Day-Ahead Electricity prices, applying this transformation helps with highly volatile, asymmetric and heteroskedastic data.

Although the Yeo-Johnson transformation effectively stabilizes variance, it introduces a necessary trade-off, re-transformation bias. Because models are trained on transformed data, their forecasts must be inverted back to the original scale for evaluation. However, because the Yeo-Johnson transformation is non-linear, this inversion is subject to Jensen's inequality (Duan, 1983).

$$E\left[f^{-1}\left(\hat{y}^{(\lambda)}{}_t\right)\right] \neq f^{-1}\left(E\left[\hat{y}^{(\lambda)}{}_t\right]\right)$$

Where f is our transformation function and $\hat{y}^{(\lambda)}{}_t$ the transformed forecast. Even if a forecast is unbiased in its transformed state, applying the strictly convex inverse function yields a systematically biased output that tends to under-predict. Nevertheless, given the highly volatile and heteroskedastic nature of Day-Ahead electricity prices, accepting this introduced bias is preferable to the alternative of forecasting on unscaled data.

3.4 Statistical Diagnostics

3.4.1 Stationarity and Seasonal Integration

As outlined by (Verbeek, 2017), in time series econometrics, we normally assume weak stationarity. This condition is satisfied if the data has the same distributional properties over time. This means that both the mean and variance of the series is non-time varying, and that the autocovariance doesn't depend on time but depends on the distance of the lag between the two observations. To determine the stationarity of the datasets, we apply two widely used tests, the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

3.4.1.1 ADF test

The ADF test investigates the presence of a unit root in the dataset. The test estimates the following regression model.

$$\Delta Y_t = \delta + \pi Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \epsilon_i$$

Where $\pi = \theta_1 + \theta_2 + \cdots + \theta_p - 1$ which is the sum of the coefficients of the AR process minus one. Under the null (H_0) we have that $\pi = 0$, and $H_1 : \pi < 0$. This way we can test if a unit root is present. $\Delta Y_t = Y_t - Y_{t-1}$ represents the first difference of Y_t, δ is simply a constant in the regression, $\sum_{i=1}^{p} \delta_i \Delta Y_{t-i}$ is the sum of the lagged first differences.

The ADF test provides a good framework for testing for unit roots, though it has one well known limitation (Kwiatkowski *et al.*, 1992). If we do not reject the null hypothesis, it does not necessarily mean that it is true, it just means that we do not have strong enough proof of it, risking that this may be because of noisy data. To address this problem and complement the ADF test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is utilized.

3.4.1.2 KPSS test

The KPSS test proposed by (Kwiatkowski *et al.*, 1992) fixes this problem, since its null hypothesis is that we do not have a unit root. The time series is decomposed into three different components, a deterministic trend, a random walk and a stationary error term.

$$y_t = \gamma t + r_t + \epsilon_t$$

Where y_t is our data which is composed of γt the time trend component, $r_t = r_{t-1} + v_t$ which is the random walk component and the error term ϵ_t which is assumed to be stationary. The logic for testing for stationarity revolves around checking the variance of the random walk errors. If $\sigma_v^2 = 0$ then r_t will be non-time varying and hence will become constant, which then means that the data is explained by a deterministic trend with stationary errors around this trend. Hence the null hypothesis is described as follows: $H_0: \sigma_v^2 = 0$. This means the data is trend stationary and $H_1: \sigma_v^2 > 0$ which means that the random walk component exists, indicating that the data is non-stationary. To test this, KPSS uses the Lagrange Multiplier (LM) approach. Firstly, we run an auxiliary regression of y_t, we then save the residuals and calculate the partial sums given by $S_t = \sum_{i=1}^{t} \epsilon_i$.The test statistic then looks as follows.

$$KPSS = T^{-2} \sum_{t=1}^{T} \frac{S_t^2}{\hat{\sigma}^2}$$

Where T^{-2} is simply a scaling factor so that the test statistic does not explode as the sample goes to infinity. By running both an ADF and a KPSS test, we now have a more robust assessment of stationarity.

3.4.2 Multicollinearity

While the exogenous variables used within our models introduce predictive power, they also introduce a classic problem, multicollinearity. This adds great risks and stability issues to the forecasting models, if the included covariates are correlated. If the covariates are highly correlated this makes the likelihood function very flat, making the estimation process difficult to find a coefficient that is stable (Greene, 2018). In the context of forecasting this can result in unprecise forecasts, because the model risks overfitting on the noise that is correlated within the covariates, giving it poor out-of-sample performance. To check if multicollinearity is an issue in our dataset, we use, Variance Inflation Factor (VIF) (Wooldridge, J. M., 2015). The idea behind this method is to check how much of the variance in the estimated coefficient is inflated due to collinearity with the other covariates. The VIF for a specific covariate is mathematically calculated as.

$$VIF_j = \frac{1}{1 - R_j^2}$$

The way this is estimated is through an auxiliary regression, where the variable in question acts as the dependent variable, while using all other covariates as predictors in the regression. If the variable we are testing has a R^2 of zero, this suggests that it holds all its own explanatory power. It is seen as standard practice to remove the covariate if it has a VIF of more than 10 (James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013).

3.4.3 Granger Causality

To validate the inclusion of the covariates, the Granger Causality Test is conducted. In time series econometrics, establishing Granger Causality proves that the lagged values of an exogenous variable (X_t) contain statistically significant information to forecast the endogenous variable (Y_t) (Shojaie and Fox, 2022).

Mathematically, the test compares a restricted univariate autoregression of the Day-Ahead electricity price against an unrestricted multivariate model. The unrestricted model using exogenous covariates looks as follows:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{p} \beta_j X_{t-j} + \epsilon_t$$

Where p represent the number of lagged observations included in the model. The null hypothesis (H_0) states that the exogenous variable (X_t) does not granger-cause the Day-Ahead electricity price (Y_t), expressed as:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

An F-test is conducted. By applying this test to the covariates, it justifies their predictive utility. Rejecting the null hypothesis confirms that these variables are not merely correlated with the electricity price but act as statistically significant indicators of the merit-order supply curve.

3.5 Forecasting Framework and Model Selection

To systematically evaluate the evolution of Day-Ahead electricity price forecasting. This thesis constructs competitive models ranging from classical time series models to state-of-the-art foundation models. The objective of this framework is not to merely pinpoint the model with the lowest error metric, but to empirically quantify the marginal predictive gains achieved by increasing architectural complexity, particularly in the context of extreme market volatility.

3.5.1 Naïve benchmarks

The efficacy of any advanced econometric, machine learning or foundational model must be benchmarked against its ability to outperform simple, non-parametrized models. Given the cyclical nature of the Day-Ahead electricity market, two baselines are chosen:

- Daily Naïve: This model assumes that the price at any given hour is the same as the day before. It captures the daily load cycle driven by the consumers and industry.

- Weekly Naïve: This model assumes the price is identical to the same hour the previous week, hence capturing the weekly patterns.

3.5.2 Classical Time Series Models

These models allow for the parametrization of the unobserved components of time series, namely trend, seasonality, autoregressive errors.

- ETS: employed as the primary univariate exponential smoothing baseline to capture temporal dynamics.
- SARIMAX: Standard ARIMA models often fails when it must address the overlapping daily and weekly seasonalities inherent to electricity markets. Therefore, this thesis employs an SARIMAX model allowing for a seasonal pattern, and the inclusion of exogenous covariates.
- MSTL-ARIMAX: The SARIMAX model is limited to using one seasonal pattern, Day-Ahead electricity prices often exhibit multiple seasonalities. Therefore, this thesis utilizes Multiple Seasonal-Trend decomposition using Loess (MSTL) to isolate these seasonal components prior to fitting an ARIMAX model, allowing for the integration of the exogenous covariates.

3.5.3 Machine Learning Models

- CatBoost: Among gradient-boosting decision trees, CatBoost is selected specifically for its handling of categorical variables via symmetric tree structure and ordered boosting. Because Day-Ahead electricity prices are dictated by categorical factors (hour of the day, day of the week, month of the year), allowing CatBoost to extract non-linear patterns, serving as the benchmark for multivariate Machine learning.

- NHITS: Is selected for its hierarchical insertion and down-sampling mechanisms. This specific architecture allows the model to efficiently ingest 168-hour historical lags alongside exogenous covariates without overfitting. By sampling data at different frequencies, NHITS can simultaneously capture broad market trends and high-frequency hourly volatility, serving as the optimal deep learning benchmark for this thesis.

3.5.4 Foundational Models

- Chronos (t5-small & V2): Chronos tokenizes continuous time series data into discrete bins, transforming the forecasting problem into a probabilistic language modeling task. The t5-small variant possesses a verifiable release date whose training data predates the 2023-2024 out-of-sample test period, establishing a zero-leakage baseline. Chronos V2 is included to represent the updated architecture. By comparing V2 against its t5-small predecessor, this thesis attempts to isolate whether predictive gains in newer foundation models are driven by algorithmic improvement, or merely forward bias from updated training data.

- TimesFM, TimeGPT, Lag-llama: Completes the selection of foundational models, these models are the state-of-the-art of foundational models (along with Chronos V2). These models are, however, released in a spectrum of dates after the in-sample-period from 2021-2023.

All the models feature a rolling window for out-of-sample testing.
The analysis is accompanied by a further theoretical explanation of each model in section 5.

3.6 Evaluation Metrics and Statistical Significance

To assess the predictive performance of the evaluated models, this thesis employs multiple predictive metrics. The models are graded not only on their point forecast accuracy but also on their probabilistic measure and statistical robustness. Given the volatility of Day-Ahead electricity prices, relying on a single error metric would yield an incomplete assessment of the models.

3.6.1 Point forecasts Metrics

The accuracy of the deterministic forecasts is evaluated using four distinct performance measures, each capturing a different economic aspect of prediction error. Let N denote the total number of out-of-sample observations, y_i the realized actual price and $\hat{y}_t$ the forecasted price.

Mean absolute error (MAE): MAE provides a linear measure of the average size of the forecasting errors. It assigns equal weight to all errors regardless of their size, hence it serves as a robust, interpretable baseline indicator of a models accuracy:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

Root Mean Squared Error (RMSE): Unlike MAE, RMSE applies a quadratic penalty to deviations before averaging, making it sensitive to outliers. In electricity price forecasting, RMSE is a critical risk management metric. A model that performs well but fails to predict a price spike during an electricity crisis will be penalized by RMSE:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Weighted Absolute Percentage Error (WAPE): Standard percentage errors are mathematically undefined for electricity prices at or near zero. WAPE resolves this limitation by dividing the sum of the absolute errors by the sum of the actual values. This effectively volume-weights the error, providing a stable percentage deviation even with negative prices:

$$WAPE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i|}$$

Mean Absolute Scaled Error (MASE): To compare predictive performance across bidding zones with varying baselines price levels, a scale independent metric is required. MASE scales the out-of-sample MAE of the evaluated model by the in-sample MAE of the seasonal naïve baseline. Let T represent the total number of training observations and m the seasonal period. A MASE < 1 indicates that the proposed model outperforms the naïve benchmark:

$$MASE = \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$$

3.6.2 Probabilistic Forecast Metric

Point forecast are inherently insufficient for optimal bidding strategies under uncertainty. To evaluate the estimated distributions, the models prediction intervals are judged using the Winkler Score.

Winkler Score: For a given $(1-\alpha)$ prediction interval bounded by (L_i, U_i), the Winkler Score penalizes the model for excessively wide intervals and applies a scaled penalty if the realized price y_i falls outside the predicted bounds. The average Winkler Score is defined as:

$$W_i = \begin{cases} (U_i - L_i) + \frac{2}{\alpha}(L_i - y_i) & if\ y_i < L_i \\ (U_i - L_i) & if\ L_i \leq y_i \leq U_i \\ (U_i - L_i) + \frac{2}{\alpha}(y_i - U_i) & if\ y_i > U_i \end{cases}$$

$$Winkler\ Score = \frac{1}{N}\sum_{i=1}^{N} W_i$$

3.6.3 Diebold-Mariano test (DM)

It is not conclusive to say that Model A is better than Model B, because it yields a marginally better RMSE. In finite samples such a difference may just be the result of random variance rather than an algorithmic improvement.

To ensure econometric correct results, a Diebold-Mariano (DM) test is employed. The DM test evaluates the null hypothesis that the predictive results of two competing models are equal. Let $d = g(e_{1t}) - g(e_{2t})$ represent the loss differential between the errors of the two models at time t. The DM test statistic is constructed as:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}}$$

Where $\bar{d}$ is the sample mean of the loss differential and $\hat{V}(\bar{d})$ is the consistent estimate of its asymptotic variance. If the DM test exceeds the critical threshold at the 5% significance level, the null hypothesis is rejected. This empirically confirms that one model provides improved predictive performance.

4. Preliminary Data Diagnostics

This section bridges the gap between the theoretical methodology and empirical forecasting analysis, by examining the statistical properties of the data. Establishing the underlying distributional characteristics, stationarity and casual dynamics of the selected variables is a prerequisite for the econometric assumptions of the forecasting models.

4.1 Summary Statistics and Distributional Properties

The dataset encompasses hourly observations from January 1, 2020, to September 30, 2024, across six European bidding zones. Table 4.1 presents the centralized moments and distributional properties of the data.

Table 4.1: Summary Statistics for Bidding Zones

Bidding Zone	Obs (N)	Mean	Std. Dev.	Min	Max	Variance	Skewness	Kurtosis*
DK_1	41616	98.43	102.82	-440.10	871.00	10571.36	2.39	7.60
DE_LU	41616	107.52	105.95	-500.00	871.00	11224.98	2.19	6.35
ES	41616	92.56	71.43	-2.00	700.00	5101.69	1.24	2.56
FR	41616	115.87	119.68	-134.94	2987.78	14323.15	2.55	17.99
NO_2	41616	86.69	94.76	-61.84	844.00	8979.79	2.50	8.45
PL	41616	101.13	63.82	-61.64	771.00	4073.37	2.26	10.20

An economic analysis of the summary statistics reveal several structural traits of the European power grid. There is extreme variance and significant positive skewness across all six bidding zones. The volatility of the market is driven by the 2022 electricity crisis. The maximum values reflect the scarcity pricing during the crisis. It should also be noted that the French bidding zone (FR) has both extreme variance and high maximum value, this is caused from a one-day electricity price spike in 2022 which comes from a combination of exceptionally low nuclear availability due to maintenance issues, historically low hydro levels from drought, and high gas prices (International Energy Agency, 2020).

The minimum values confirm the frequent occurrence of negative clearing prices, especially in zones with high integration of renewable electricity (DK_1, DE_LU). Furthermore, the high excess kurtosis confirms that price spikes are not normal deviations but rather structural fat tailed events. This excess kurtosis empirically validates the methodological necessity of employing the Yeo-Johnson transformation to stabilize the variance.

4.2 Stationarity analysis

To avoid spurious regression in the classic time series models, the stationarity of the endogenous variable must be confirmed. Both the ADF test and KPSS test is employed.

The ADF test evaluates the null hypothesis of a unit root (non-stationary), the KPSS test evaluates the null hypothesis of trend stationarity, requiring a constant variance around a constant mean or trend. The empirical results are shown in (A1)

The ADF test decisively reject the null hypothesis of a unit root, confirming that the transformed series is mean-reverting. In electricity markets, this aligns with economic intuition, prices do not drift indefinitely but are forced back to equilibrium by demand elasticity and the merit-order effect.

However, the KPSS test rejects the null hypothesis of trend stationarity. This indicates that the series is variance non-stationarity. While the Yeo-Johnson transformation successfully stabilized the baseline, the magnitude of the 2022 electricity crisis introduced some structural volatility that mathematically makes it difficult to have strict variance stationarity over the full 2020-2024 sample.

In conclusion, the conflicting results between the ADF and KPSS test reveals the fundamental econometric challenge of modeling Day-Ahead electricity prices. For our forecast procedure, this has a profound implication, the classical time series models will likely struggle to maintain stable error variances and reliable confidence intervals. This further validates the inclusion of machine learning and foundation models which do not have these strict requirements for the input data.

4.3 Multicollinearity analysis

The inclusion of exogenous covariates introduces the econometric risk of multicollinearity, which can inflate standard errors. The variance Inflation Factor (VIF) was calculated for the exogenous covariates.

A standard econometric rule is that a $VIF > 5$ indicates moderate multicollinearity, a $VIF > 10$ indicates strong multicollinearity, requiring intervention. To keep this section concise, the full VIF diagnostic table is reported in appendix (A2).

The VIF test reveals that the grid covariates (forecasted load, wind and solar) and the macroeconomic variables (TTF, CO2 quotas) exhibit independence, consistently scoring below the critical threshold ($VIF > 3$)

However, strong multicollinearity was detected within the NTC variables. Most notably export and import from Great Britain (GB). The interconnector between FR and GB had a VIF score of more than 990, and the interconnector between GB and DK_1 had a score of more than 63. This suggest that the transfer capacities are highly symmetrical and co-move. These highly colinear NTC variables have been dropped to prevent standard error inflation.

4.4 Exogenous Predictive Utility

To empirically justify the inclusion of the covariates, the predictive utility of the covariates was validated using the Granger Causality test. Given the daily cycle of Day-Ahead electricity market, the test was evaluated at the primary daily lag ($m = 24$). The complete table of F-tests and corresponding p-values of all the covariates are provided in appendix (A3).

The results reject the null hypothesis of non-causality for the fundamental grid covariates, specifically Forecasted load, wind and solar as well as the lagged TTF prices across most of the evaluated bidding zones. This empirically validates the theoretical dynamics of the European merit-order mechanism, discussed in section 2.1. The forecasts of the renewable electricity generation, dictate the horizontal shift of the supply curve, while the TTF price dictates the marginal cost of the dispatchable thermal generation required to clear the market.

Overall, the diagnostics firmly justify a multivariate forecasting framework, proving that exogenous grid fundamentals contain critical explanatory power.

4.5 Spectral analysis

Day-Ahead electricity markets are highly seasonal in both supply and demand patterns. To empirically validate and quantify these seasonal patterns, a spectral analysis is conducted utilizing Fourier decomposition. This way we can see how much of the variance is explained by different cyclical frequencies, this is further explained in appendix (A4).

The result of this analysis is visualized in the periodogram in appendix (A4). The periodograms plot the spectral density, calculated as the squared coefficients $\alpha_k^2 + \beta_k^2$ on a logarithmic scale against the period length in hours. It is observed that the most prominent seasonal patterns are at 24 and 168 hours, which is one day and one week respectively. For this reason, these are the used frequencies. This further validates the choice of including seasonal components in the deployed models.

5. Analysis

Having established the theoretical complexities of the European Day-Ahead electricity market, defined the methodological approach needed to handle them and examined the statistical properties of the data, this chapter presents the empirical analysis. The primary objective is to forecast and evaluate their performance across three main categories of models, the classical time series models, machine learning models and foundation models. The analysis chapter is structured in the following way, each individual model is examined following a similar structure. First, the specific theory and mathematical mechanics of the model will be outlined, second the out-of-sample forecasts are generated over the testing period, finally a comprehensive evaluation is conducted using the established point and probabilistic performance metrics. All residual diagnostics are found in the appendix (A6).

5.1 Classical times series models

5.1.1 Seasonal Naïve models

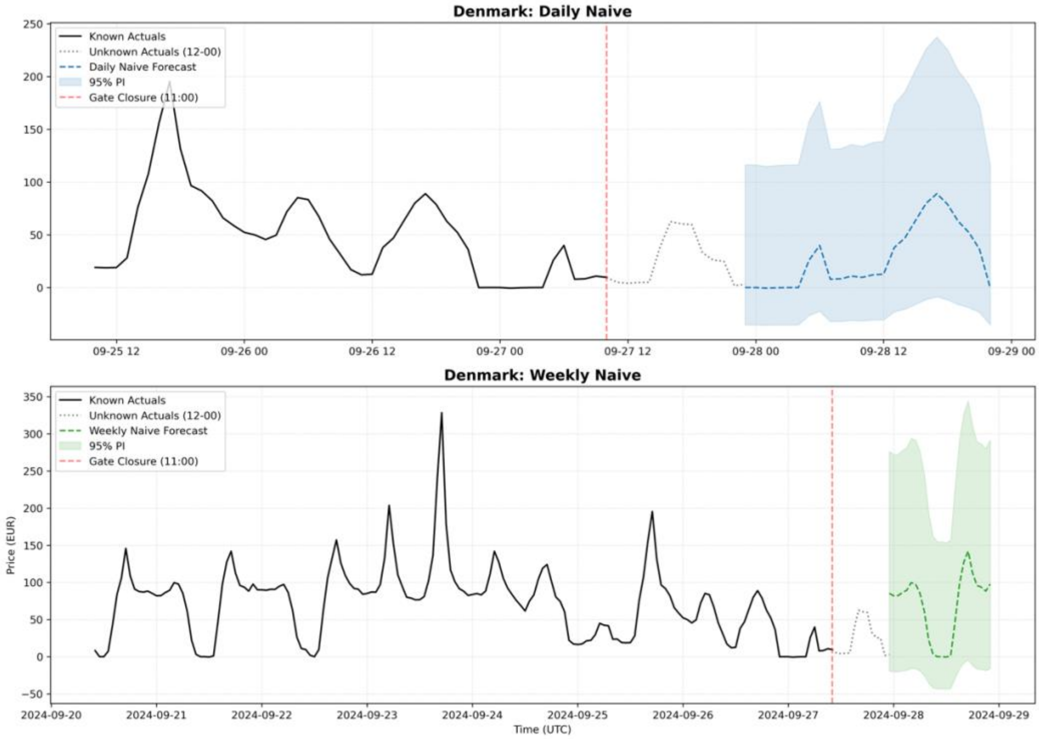
We begin the empirical analysis with the seasonal naïve models. These models will serve as a forecasting baseline. The reason for using the seasonal naïve approach is because, Day-Ahead electricity prices are heavily dictated by the seasonal patterns. The main patterns observed in the spectral analysis of the data in chapter 4.5 are the daily ($m = 24$) and weekly ($m = 168$) cycles. These two seasonal patterns are not introduced to one model, but rather two separate models. This way we can also evaluate which seasonal pattern is the strongest. Formally the h step ahead forecast is mathematically defined as follows.

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

Where m is as mentioned the seasonal period, k represents $(h-1)/m$. This accounts for the number of cycles completed within the forecast horizon prior to $T + h$.

For the seasonal naïve models, we encounter a methodological challenge. The model must adhere to the market gate closure constraint to prevent forward bias.

Once the daily and weekly point forecasts have been generated in the transformed space, we construct the 95% prediction intervals by using the standard deviation from the historical residuals, the prediction intervals are calculated as $\pm 1{,}96 \cdot \sigma$. These prediction intervals and point forecast are then reverted to the original scale. The forecasting looks as follows for a given week and daily forecast.



To show the process of forecasting with seasonal naïve models, we will use the zone of DK_1 as an example. This plot clearly shows the gate closure constraint, preventing forward bias. The graph shows that the last information is available before the gate closure is at 11:00. Since, it is impossible to run the model and execute the bidding strategies with the observation at time 12:00. For the daily model it uses what the prices were today from 00:00 to 11:00 and from 12:00 to 23:00 what the price was yesterday. It is also seen that for the weekly model it uses what the price was 168 hours ago.

Table 5.1: Naïve Models Performance Metrics

Country	Model	RMSE	MAE	WAPE (%)	Winkler Score
DK_1	Daily Naïve (m=24)	44.48	31.43	40.25	279.52
	Weekly Naïve (m=168)	49.18	35.52	45.49	266.55
DE_LU	Daily Naïve (m=24)	43.13	29.47	34.37	240.75
	Weekly Naïve (m=168)	44.83	31.26	36.47	257.20
FR	Daily Naïve (m=24)	34.16	24.18	31.51	259.78
	Weekly Naïve (m=168)	37.55	27.42	35.74	214.76
PL	Daily Naïve (m=24)	36.28	24.59	23.83	407.01
	Weekly Naïve (m=168)	36.53	25.75	24.95	198.72
ES	Daily Naïve (m=24)	32.61	22.08	30.60	346.54
	Weekly Naïve (m=168)	37.54	26.94	37.33	208.47
NO_2	Daily Naïve (m=24)	25.45	17.27	25.85	247.41
	Weekly Naïve (m=168)	29.77	21.01	31.46	175.50

Across all bidding zones we observe that the seasonal daily naïve model consistently outperforms the seasonal weekly naïve model. This is observed across all the point forecasts performance measures. This shows that the autocorrelation is better across day to day basis rather than week to week basis. This is also expected since a week is a significant amount of time and hence weather trends and other external shocks have faded out and hence do not have the same effect.

Further analysis of the different models will not include the detailed explanation of market gate closure constraint, as it is the same as the seasonal naïve models. The graph of the forecast will not be included either as it adds no value to the analysis.

5.1.2 ETS

We now turn to classical time series models, the Exponential Trend Smoothing model (ETS). The main idea of the ETS framework, is that recent observations carry more information than older ones, the model applies exponentially decaying weights to the coefficients. Including ETS allows us to understand if smoothing models are better than autoregressive models, for the Day-Ahead electricity markets.

The model divides the data into three separate components, error, trend and seasonality. Each component can be one of the following state, additive, multiplicative or none, noted as (A), (M) and (N) respectively. Because of computational constraints, we will not use any grid search algorithm to find the most effective model, and will instead assume the following model, ETS (A, N, A). This specific combination, says that we assume the errors are additive, meaning that the errors do not get multiplicatively bigger over time. There is no deterministic trend over time, hence the none (N) term and that the seasonality is additive. For each model, there is two different types of equations, the measurement equations which describes the observed data, and state equations, which describe the unobserved data. Mathematically these are defined as follows:

$$y_t = l_{t-1} + s_{t-m} + \epsilon_t$$

$$l_t = l_{t-1} + \alpha\epsilon_t$$

$$s_t = s_{t-m} + \gamma\epsilon_t$$

Where in the measurement equation y_t is the price at time t, which is constructed from the level l_{t-1}, the seasonal component s_{t-m} and the errors ϵ_t assumed to be $\epsilon \sim NID(0, \sigma^2)$. The parameters in the state equations α and γ define the smoothing amount for the level and seasonality, respectively. The parameters and the initial states are estimated with the function *ETS* from the *statsforecast* python package, which uses maximum likelihood estimation. To ensure consistency and comparability, we apply the same rolling window approach, with the simulated gate closure constraint to the ETS framework. For this model, covariates are not integrated, forcing the model to make predictions based solely on the lagged price. This yields the following results across our bidding zones.

Table 5.2: ETS Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	55.88	42.12	53.94	1.340	249.41
DE_LU	56.65	43.08	50.26	1.462	264.01
FR	44.89	34.11	44.45	1.411	219.50
PL	61.57	43.46	42.11	1.767	358.48
ES	47.05	36.24	50.22	1.641	281.97
NO_2	36.92	26.43	39.58	1.531	188.35

We observe the worst performance in PL, with an RMSE (61,57) and MAE (43,46), also the biggest prediction intervals, a winkler score of 334,77. One would expect Poland to have a flatter supply curve and therefore a more stable price, the reason for the poor performance of the ETS framework in Poland is a combination of the following. Because of the lower variance environment of Poland, the model accustoms to this, when an unexpected shock then happens, the model has a hard time capturing this and adjusting. Another possible explanation could be that Poland is more prone to smaller regime shifts, for example when there is a power plant outage, it does not just take a couple of hours to revert back to its mean, this in combination to the forced ETS(A,N,A) model, that does not allow it to adjust through the trend component, hence the model expect the price to revert back to its mean, which it does not. In general, it has better performance in the more volatile bidding zones, although still worse than the seasonal naïve approach.

5.1.3 SARIMAX

Having established the baseline models and the univariate ETS model, the first of the autoregressive class, is the Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model.

The SARIMAX model represents the parameterization of the time series, capturing linear dynamics while simultaneously estimating the deterministic impact of the exogenous covariates. The SARIMAX $(p, d, q)(P, D, Q)_m$ is defined using the lag operator (L) as follows:

$$\phi_p(L)\Phi_p(L^m)(1-L)^d(1-L^m)^D y_t = c + \sum_{i=1}^{k} \beta_i X_{it} + \theta_q(L)\Theta_q(L^m)\epsilon_t$$

Where y_t is the endogenous variable, the Day-Ahead electricity price at time t, $\phi_p(L)$ and $\theta_q(L)$ represent the non-seasonal autoregressive (AR) and moving average (MA) polynomials of orders p and q. While $\Phi_p(L^m)$ and $\Theta_q(L^m)$ represent the seasonal AR and MA components of orders P and Q, at the seasonal frequency m. X_{it} is the matrix of exogenous covariates, where β_i represent their estimated linear coefficients. c is a constant drift term and ϵ_t represents the white noise error term, assumed to be normally distributed.

Implementing the SARIMAX equation into a operational Day-Ahead electricity price forecasting method, requires strict adherence to the markets gate closure constraint.

Standard recursive forecasting, where you forecast $t + 1$ and then using that prediction to forecast $t + 2$ is highly susceptible to forecast error accumulation. This thesis implements a direct multi-step forecasting strategy. Rather than forecasting a single continuous time series, the dataset is split into 24 independent hourly sub-series. The model trains 24 distinct models in parallel, each optimized for every hour of the day. This aligns with the market clearing mechanism EUPHEMIA, where all 24 hours clear simultaneously.

To adhere to the market gate closure constraint, the information set is dynamically restricted based on the target hour. For morning hours $(h < 12)$, the actual realizations of day D are observable by the 12:00 gate closure, therefore the model trains on data up to day D and executes a 1-step ahead forecast for day $D + 1$. For afternoon and evening hours $(h \geq 12)$ the model is restricted to training on data ending on day $D - 1$ and executes a 2-step ahead forecast for day $D + 1$.

A recognized limitation of the direct strategy is that the independently hourly models lose cross hourly context, the $h = 14$ model does not know what happened at $h = 13$. To remedy this without violating gate closure, custom features were engineered. The model is supplied with aggregated historical context from day $D - 1$. These features anchor each independent hourly models forecast to the previous day's broader trends (A6).

A continuous rolling window is employed. At the beginning of each 30 day out-of-sample rolling block, the Hyndman-Khandakar algorithm is used via the *StatsForecast* library to search the parameter space and identify the $(p, d, q)(P, D, Q)_m$ orders minimizing AIC. The 30-day refitting is due to computational constraints. The seasonal frequency is set to $m = 7$ to capture the weekly cycle within the hourly data.

The optimally identified orders are then passed to the *statsmodel* state-space engine. This engine fits the maximum likelihood coefficients for the AR, MA and β parameters.

Table 5.3: SARIMAX Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	37.34	28.01	35.86	0.891	215.35
DE_LU	36.63	26.83	31.26	0.910	207.44
FR	31.08	24.22	31.56	1.002	168.97
PL	30.52	21.14	20.49	0.860	185.60
ES	27.20	20.51	28.40	0.929	161.51
NO_2	25.85	19.31	29.01	1.118	161.50

Evaluating the direct SARIMAX model across the six bidding zones reveal that the bidding zones electricity portfolio is the primary factor that determines how well the SARIMAX model performs.

The model exhibits the highest point forecast errors in the DK_1 (RMSE 37,34) and DE_LU (36,63). This poor performance is a consequence of these zones high integration of non-dispatchable wind generation. Because wind is inherently stochastic, it causes a horizontal shift in the merit-order supply curve. When this interacts with the inelastic demand curve, this creates price volatility and price spikes that the autoregressive polynomials of the SARIMAX model struggle to capture. However, it must be noted that both zones scored a MASE of approximately 0.90, indicating that the model successfully integrated the exogenous covariates to narrowly outperform the naïve baseline benchmark.

However, in bidding zones dominated by dispatchable baseloads, the point forecasts improve significantly. Poland whose baseload is heavily dominated by coal-fired generation, has a much flatter and more predictable supply curve. This translates into the models best relative performance, achieving a WAPE of 20.49% and a MASE of 0.86. France (FR) which is dominated by nuclear baseload generation also exhibits lower errors (RMSE 31.08) compared to the solar and wind heavy bidding zones. Nuclear provides stable and continuous generation. However, the MASE of FR sits at just 1.002 indicating that the SARIMAX model did not capture anything new compared to the naïve benchmark.

Spain (ES) has strong point forecast accuracy (RMSE 27.2) and stable probabilistic measures. Spain is dominated by solar generation, which unlike wind follows a deterministic daily pattern. The SARIMAX model seasonal autoregressive component easily captures this daily pattern, yielding strong predictions and a WAPE of 28.40%

Southern Norway (NO_2) is fascinating, in absolute terms the model achieves its lowest errors here (RMSE 25.85, MAE 19.31). Because Norway relies almost entirely on hydro-reservoirs, producers can store potential electricity and dispatch it when needed, smoothing price volatility. However, this stability causes the SARIMAX model to fail relative to the naïve baseline benchmark, yielding a MASE of 1.118. Because hydro-dominated prices are autocorrelated, the water available today is virtually identical to the water available tomorrow, the simple naïve benchmark is superior.

5.1.4 MSTL-SARIMAX

In chapter 4.5 the importance of seasonality in the Day-Ahead electricity market was established. The general SARIMAX models discussed in the previous chapters are only able to handle one seasonal pattern natively. This problem is the motivation to interchange how SARIMAX handles seasonal patterns to utilizing Multiple Seasonal Trend decomposition using Loess (MSTL). This approach makes it possible to filter different types of seasonality at the same time. The MSTL approach is developed by (Bandara, Hyndman and Bergmeir, 2021). The core idea behind it is to decompose the time series y_t, into multiple seasonalities, a trend and a remainder component. This is mathematically defined as.

$$y_t = T_t + \sum_{i=1}^{M} S_{i,t} + R_t$$

Where T_t is the trend component, R_t is the remainder component and $S_{i,t}$ represents the i-th seasonal component. For the seasonality in the electricity market it is $S_{24,t}$ and $S_{168,t}$. Once the data has been decomposed by deploying the MSTL method, the trend and the remainder can then be modelled using the ARIMAX framework. ARIMAX is mathematically expressed as:

$$\phi(L)(1-L)^d y'_t = c + \sum_{j=1}^{k} \beta_j X_{j,t} + \theta(L)\epsilon_t$$

Where $y'_t = y_t - S_{24,t} - S_{168,t}$ is the seasonally adjusted price, (L) is the lag operator, $\phi(L)$ represents the autoregressive patterns, $\theta(L)$ represents the moving average parameters, d is the amount of differencing applied to the data set in order to achieve stationarity and $X_{j,t}$ are the covariates with its corresponding coefficient β_j and the error term ϵ_t.

To evaluate the MSTL-ARIMAX model, we implement a rolling window approach, with the gate closure constraint satisfied. This ensures no forward bias. Furthermore, the model also integrates the covariates $X_{j,t}$.

Just like the computational constraint encountered in SARIMAX, we estimate the optimal ARIMA parameters (p, d, q), once every 720 hours, which is equivalent to 1 month. These parameters are then forced upon the model for the rest of that month's forecast. This is done with the *Autoarima* function from the *statsforecast* python package. One the parameters have been chosen; the function uses MLE to estimate the coefficients.

Table 5.4: MSTL-ARIMAX Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	37.25	26.84	34.38	0.854	208.53
DE_LU	32.70	23.55	27.47	0.799	201.20
FR	29.31	21.59	28.14	0.893	186.98
PL	31.67	22.31	21.62	0.907	173.37
ES	28.85	20.39	28.26	0.924	143.45
NO_2	31.66	21.22	31.78	1.229	147.30

The performance across the different bidding zones will now be analyzed following the same structure as in the previous chapters.

We see that both the Danish and German bidding zones are the most difficult for MSTL-ARIMAX to forecast, with Denmark having the highest RMSE (37,25) and MAE (26,84), followed by Germany with RMSE (32,7) and MAE (23,55), while also both of them having the largest prediction intervals showed by the Winkler score. One of the reasons why the model is worse under these conditions is because of the high dependence both Denmark and Germany have on wind sources. Because wind is non-dispatchable and because of the merit order effect discussed earlier, this creates the difficult dynamics of volatility clustering, price spikes etc. Poland and France consist mainly of dispatchable sources. Here we observe that the two bidding zones are relatively close to each other in performance, Poland with RMSE (31,67) and MAE (22,31), one of the reasons for this performance shift in Poland arises from the fact that the electricity market being mainly dominated by coal powerplants resulting in a flatter supply curve, hence making the price less volatile and more predictable (Zachmann *et al.*, 2023). Whereas in the bidding zone for France we observe RMSE (29,31) and MAE (21,59), as mentioned, France electricity market is highly dominated by nuclear power, even though its less flexible than a normal power plant, the price is more stable compared to Denmark and Germany. Hence, we observe significantly better performance measures in both countries. Another bidding zone where the model has a mid-performance is in Norway. As established in the SARIMAX framework the water used for the hydro powerplants can be stored as potential energy, resulting in a more stable price. The best performing bidding zone for this model is Spain. We observe the best Winkler scores of 145. We also observe RMSE (28,85) and MAE (21,22). As discussed earlier, because of the high electricity production coming from solar power in Spain, following a daily pattern. The decomposition of m = 24 within MSTL, captures most of the solar variation.

5.2 Machine learning models

As seen in the previous chapters, the classical time series models struggle to capture nonlinear relationships. To address these limitations, machine learning models will be deployed, as they have had increasingly importance within electricity price forecasting (Lago *et al*, 2018). Among machine learning, two separate categories were chosen, Gradient Boosted Decision Trees and Deep Neural Networks. Specifically, CatBoost and NHITS will be covered in the following chapters.

5.2.1 CatBoost

The CatBoost model has a decision tree structure. A decision tree splits the predictor space into a number of regions $R_1, \ldots, R_M$. The regression decision tree can be formalized mathematically as:

$$F(X) = \sum_{m=1}^{M} c_m \cdot 1_{(x \in R_m)}$$

Where c_m is the mean of the training observations within the terminal node R_m. The regions are found using a recursive binary splitting algorithm. The goal of the recursive binary splitting algorithm is to find the regions $R_1, \ldots R_M$ that minimize the Residual Sum of Squares:

$$RSS = \sum_{j=1}^{M} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

The general problem with these types of trees is that they are very prone to high variance and overfitting. However, CatBoost uses a gradient boosting approach. Gradient boosting models the prediction as the sum of the weak learners. Defined as follows:

$$F_M(x) = \sum_{m=1}^{M} f_m(x)$$

Where $F_M(x)$ is our final prediction and $f_m(x)$ are the weak learners. Boosting, constructs trees sequentially so that each new tree learns from the mistakes of the previous trees. Before we can start training our model, a loss function *(MultiQuantile)* is defined. For the model to recognize which direction it should go, we take the gradient of the loss function. This is mathematically defined as.

$$r_i = -\frac{\partial L\left(y_i, F_{m-1}(x_i)\right)}{\partial F_{m-1}(x_i)}$$

At step m we compute the negative gradient for each observation i. The new weak learner $f_m(x)$ is trained to predict the residual r_i, this makes the model take small steps in the direction of the steepest decent, hence minimizing the loss. Now that the algorithm knows the direction, it must also know how big of a step it should take, which is defined by γ:

$$\gamma = \arg min_{\gamma} \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \gamma f_m(x_i)\big)$$

When γ has been found we can now update the model:

$$F_m(x) = F_{m-1}(x) + \gamma f_m(x)$$

The residuals are calculated in the following way:

$$r_i = y_i - M_{\sigma(i)-1}(x_i)$$

Where y_i is the actual data and $M_{\sigma(i)-1}(x_i)$ is a supporting model trained only on the preceding examples in that permutation. By subtracting the support model's prediction from the true value y_i we get unbiased residuals.

Time features are added such as, hour of the day, day of the week and current month. Autoregressive features are also added such as price lags for 24, 48 and 168 hours. The mean of the morning price 00:00-11:00 at day D is also created. Covariates are also integrated. To prevent forward bias, we implement a simulated gate closure.

Table 5.5: Catboost Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	25.70	18.15	23.27	0.578	137.84
DE_LU	18.53	12.12	14.13	0.411	94.28
FR	22.21	16.26	21.24	0.672	138.37
PL	19.87	13.35	12.94	0.543	111.32
ES	17.53	12.87	17.87	0.583	106.82
NO_2	20.71	14.28	21.58	0.827	116.43

After having established the theory and the methodological integration of our CatBoost model we can now evaluate the results in out-of-sample performance. Denmark remains the most challenging to forecast because of its volatile nature with an RMSE (25,70) and MAE (18,15). However, in the other most volatile bidding zone we have, Germany, CatBoost manages to significantly reduce its RMSE and MAE to 18,53 and 12,12 respectively.

In the two other zones dominated by more dispatchable sources, as France and Poland, we observe as expected better performance than in the Danish bidding zone. In France RMSE (22,21) and MAE (16,26). In Poland RMSE (19,87) and MAE (13,35). In the Spanish bidding zone, we observe the best performance from CatBoost, with RMSE (17,53) and MAE (12,87), this is likely because of the seasonal patterns of solar. This is encoded as a feature in the model, in the form of the hourly time feature, this shows that CatBoost is able to detect this seasonality without any seasonal decomposition. It must also be mentioned that because of the models robust capabilities against heteroskedasticity, the otherwise used variance stabilizing transformation of Yeo-Johnson is not used in this case, showcasing its abilities to remain stable even under volatility clustering.

Most notably when evaluating the confidence intervals of CatBoost, we observe a great performance. This shows the model can capture periods with high volatility and open the intervals up, and when in stable markets or time periods shrink these down, accurately mapping the true market risk.

5.2.2 NHITS

While tree-based models like CatBoost excel portioning feature spaces, they lack an understanding of temporal dependencies. To model complex non-linear dynamics, the empirical analysis advances to deep neural networks, employing the Neural Hierarchical Interpolation for Time Series (NHITS) model (Challu *et al.*, 2022b).

NHITS relies on a multi-layer perceptron built upon a twofold residual topology (Challu *et al.*, 2022b). The network is composed of multiple stacks each containing multiple blocks. In standard residual networks, a block maps and input to an output. In the NHITS model, each block b processes the sequential input to produce two different outputs: a forward forecast $\left(\hat{y}^{(b)}\right)$ and a backward backcast $\left(\hat{x}^{(b)}\right)$. The backcast is subtracted from the input, leaving only the unexplained residual error to be passed to the next block:

$$x^b = x^{(b-1)} - \hat{x}^{(b-1)}$$

The final prediction is the sum of the forecast from the blocks:

$$\hat{y} = \sum_{b=1}^{B} \hat{y}^{(b)}$$

The innovation of NHITS is its hierarchical interpolation and multi-rate data sampling. By using MaxPool mechanisms to down-sample the input sequence at varying frequencies, the early stacks capture trends and the later stacks capture high-frequency hourly volatility (Challu *et al.*, 2022b).

To get the probabilistic risk rather than just the point forecast, the network minimizes the Mulit-Quantile Loss (*MQLoss*) function during gradient descent. For a given quantile $\tau \in (0,1)$, the loss is defined as the asymmetric absolute error:

$$L_\tau(y, \hat{y}_\tau) = \max\left(\tau(y - \hat{y}_\tau), (\tau - 1)(y - \hat{y}_\tau)\right)$$

Using the NHITS model for Day-Ahead forecasting requires strict adherence to the market gate closure constraint to prevent forward bias. The model is configured with a look back window of 168 hours to simultaneously forecast a 24-hour block. The model integrates exogenous covariates.

The autoregressive price lags were engineered using a daily lag. For hours before 12:00 gate closure, the network uses $t - 24$ prices, for afternoon hours it uses $t - 48$, simulating trading constraints.

Table 5.6: NHTIS Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	26.66	18.32	24.22	0.583	134.60
DE_LU	25.42	16.14	19.35	0.548	115.69
FR	26.60	14.97	20.51	0.619	103.97
PL	23.37	15.61	15.42	0.635	115.48
ES	20.10	13.28	18.50	0.601	101.65
NO_2	15.83	10.66	16.81	0.617	99.41

Evaluating the out-of-sample performance demonstrates the capabilities of hierarchical neural networks. Across all six bidding zones, the NHITS model achieved a MASE between 0.583 and 0.635 confirming that the neural network successfully reduced the forecasting error compared to the naïve baseline.

In the wind dominated Danish zone, NHITS achieved an RMSE of 26.66 and an MAE of 18.32. However, consistent with the results of the CatBoost model, the neural network demonstrated its strongest relative performance in the highly interconnected German (DE_LU) zone. In Germany, NHITS yielded an RMSE of 25.42, an MAE of 16.14, and the lowest MASE across the bidding zones at 0.548. This divergence supports the hypothesis that NHITS can take advantage of Germanys large matrix of NTC limits to capture complex cross-border flow dynamics, successfully mitigating the stochasticity of wind generation better than in more geographically isolated zones like Denmark.
In the solar dominated ES zone, the model successfully extracted the deterministic daily solar cycle, achieving an excellent RMSE of 20.10 and an MAE of 13.28.

The Southern Norway (NO_2) zone recorded the lowest absolute errors, with an RMSE of 15.83 and an MAE of 10.66. This proves that the NHTIS can utilize exogenous predictive signals to outperform the massive autocorrelation of hydro reservoirs.

The NHITS model successfully provides probabilistic stability. The Winkler Scores remained highly compressed across all regimes, capping at 134.60.

5.3 Foundation models

As established in the preceding analysis chapters, classical time series models have trouble capturing the non-linear relationships and rely on strict assumptions that are simply not feasible in extreme volatile Day-Ahead electricity markets. However, machine learning models can acknowledge these nonlinear relationships and do not impose the strict assumptions, they have the disadvantage of having to remain localized. Furthermore, they require manual engineering of features. To resolve this, we dive into foundation models and their performance. Most foundation models are transformer-based architecture pretrained on a large amount of data. Hence, they can perform zero shot forecasting, meaning that there is no requirement for training. They only need a context window, that compared to machine learning and classical time series models is very small. The following chapter will explain the theory behind each model and analyze its performance across bidding zones.

5.3.1 Time-GPT

Developed by Nixtla, Time-GPT is the first major foundation model specifically designed for time series forecasting (Garza, Challu and Mergenthaler-Canseco, 2024b). Time-GPT uses a generative pretrained transformer architecture. It has a encoder decoder transformation architecture. The model is pretrained on 100 billion datapoints, with many diverse datasets. These datasets were not stabilized in any way, meaning that the model has been exposed to data with noise and outliers at different scales, increasing the robustness of the model. The architecture can be visualized in A6. Time-GPT takes the input and models it autoregressively. The objective of the model is to estimate the conditional distribution of future values given past values and the available covariates, this is mathematically defined as:

$$P\left(y_{[t+1:t+h]} \middle| y_{[0:t]}, x_{[0:t+h]}\right) = f\left(y_{[0:t]}, x_{[0:t+h]}\right)$$

Where $y_{[t+1:t+h]}$ are the future values over the forecast horizon h, given the data available up until time t. $y_{[0:t]}$ are the past values and future covariates being $x_{[0:t+h]}$.

We require prediction intervals that are accurately able to assess the actual risk of a given point interval. Classical time series models construct these intervals assuming that the errors are normally distributed. Time-GPT takes another approach that does not require this strict assumption about the distribution of the errors. Time-GPT uses what is called conformal prediction. It generates historical predictions using cross-validation obtaining the actual and forecasted values. After obtaining the actual and forecasted values, the model can calculate the conformity scores. These conformity scores are then added and subtracted for each point forecast to generate the prediction intervals.

Since Time-GPT is not open source, the model must be accessed through an API key. The market gate closure constraint is applied. Using the same rolling window approach as the classical time series models and machine learning models ensure comparability across models. Because the model has a large context window, all 26280 observations prior to the testing period were used.

Because Time-GPT comes as a pretrained model, we are running the risk of the model having memorized the out-of-sample data set, this makes it look like the model has a great performance out-of-sample, when the model has simply memorized the next actual value. Since the model was published in October 2023 and there is no record of the model being trained afterwards, this imposes the memorization risk in our out-of-sample period ranging from January 2023 to September 2024. This also suggests that the remaining 11 months, from November 2023 to September 2024, is the true zero shot period. Another factor that must be considered is that because of Time-GPT not being open-source, it acts as a black box for the forecaster, this means that in theory the model could be updated regularly, this would completely invalidate all out-of-sample performance.

Table 5.7: Time-GPT Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	28.04	19.70	25.22	0.627	279.30
DE_LU	21.88	14.50	16.91	0.492	212.02
FR	25.78	18.64	24.29	0.771	258.03
PL	20.47	14.22	13.78	0.578	200.46
ES	20.58	14.66	20.30	0.664	213.07
NO_2	23.34	16.66	24.96	0.965	228.98

Having established the architecture and introduced the methodological approach we can now evaluate the out-of-sample performance. In general, we see a much lower spread in performance across bidding zones, this hints that the model is much better at adapting to different electricity structures. We observe that Denmark, remains among the most difficult to forecast. We observe RMSE (28,04) and MAE (19,7) in this bidding zone. In the other extreme we have the polish bidding zone, with an

RMSE (20,47) and MAE (14,22), this is by far the best performing zone for this model, additionally we observe the best Winkler score, showing that here Time-GPT is better at assessing the risk of the market.

5.3.2 Chronos

The Chronos foundation model developed by Amazon (Ansari *et al.*, 2024b), differs fundamentally from other foundation models because of its T5 structure. The Chronos architecture operates as a Large Language model (LLM), essentially being a text-to-text transfer transformer. This is achieved by converting the data into a sequence of text, then the model outputs another sequence of text which is then converted back to numeric.

The core lies in the tokenization process. This consists of two primary processes, mean scaling and quantization. When a time series is put into the model it undergoes mean scaling. This process normalizes the data while keeping the overall structure. The main problem arises when the data contains extreme heteroskedasticity. In highly volatile markets the non-stationarity causes clipping, this means that Chronos cannot differentiate between large spikes and very large spikes, because it assigns the same token to the extreme outliers. Making the model blind to the magnitude of the outliers.

The scaled values are clipped. From here it quantizes the data into 4049 uniform discrete bins. Each bin is for the model a specific text token, this process transforms the data into the language that the model understands, although this makes the model lose precision, since two values that are different can end up being categorized in the same bin.

The chosen Chronos model is the Chronos-t5-small, which has 46 million parameters, a larger model was not chosen because of computational constraints. This model in specific has a context window of 512 data points, establishing a baseline for the time series. To maintain a valid comparison across models, the same gate closure is implemented. This ensures that there is no risk of forward bias. A disadvantage of the Chronos t5-small architecture is the inability to include covariates as an input. Hence, the forecast is generated based solely on the lagged prices. Although it must be mentioned that because the model was published in 2024, it overlaps our out-of-sample period, hence there is a risk of memorization bias.

Table 5.8: Chronos t5-small Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	40.76	28.26	36.15	0.899	210.83
DE_LU	38.77	26.06	30.35	0.884	207.01
FR	29.49	21.68	28.21	0.897	170.51
PL	30.39	21.00	20.33	0.854	160.61
ES	29.16	20.15	27.88	0.913	155.86
NO_2	22.00	15.44	23.14	0.894	108.98

Having established the architecture and methodological approach, we can now evaluate its out-of-sample performance. We see as established in the preceding chapters, that Denmark and Germany are the most difficult to forecast accurately, we observe a RMSE of (40.76) and (38.77) respectively. Along with the highest Winkler scores of (210.8) and (207.01). This clearly indicates that a combination of the model having no covariates and the highly volatile market conditions make these two zones very difficult to forecast. In contrast, we see in the Norwegian bidding zone, the best performance with RMSE (22.00) and MAE (15.44) as well as the lowest Winkler score (108.98). We observe that the Norwegian bidding zone is easier for the model to predict when only having past price as an input, this is because of the more stable nature of this bidding zone. The rest of the countries lie between these bidding zones, with medium performance.

5.3.3 Chronos V2

With the first version of Chronos covered, we can now dive into the new variation of the Chronos t5. Chronos V2 utilizes a new architecture which focuses on solving the limitations of its predecessor. The main limitation of the first version of Chronos was the inability to use covariates. Another limitation of the first version was the tokenization process, making it less precise. The Chronos V2 transitions to a more widely used architecture, the encoder transformer architecture as seen in Appendix A8. Chronos V2 standardizes the data and then transforms it, this transformation stabilizes variance and minimizes the influence of extreme. Another architectural change is the fact that Chronos V2 now possesses a group attention mechanism, because of this, Chronos V2 can see patterns not only on past data, but also in the specified covariates time series (Ansari *et al.*, 2025).

A significant concern with newer foundation models is the risk of the model memorizing datasets it was trained on, hence showing much better performance than it is actually capable of producing. Chronos V2 relies heavily on diverse real-world datasets, but this is only the case for univariate datasets. Its ability to understand the relationships across covariates is driven by synthetic data. The developers of the model have utilized what is known as multivariatizers, which generate complex relationships across datasets, to teach the model how to interpret covariates. Because of the synthetic generation of covariate mechanisms, the risk of the model having memorized the entire datasets is not likely, but there is still a risk regarding the actual price series, since the paper was submitted October 2025, which overlaps with our out-of-sample data (Ansari *et al.*, 2025).

Regarding the methodological approach, the same gate closure constraints were implemented. Because of its new capabilities, Chronos V2 was supplied with the covariates. The model's maximum context window size is 8192 observations, hence all these are used.

Table 5.9: Chronos V2 Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	27.20	18.23	23.32	0.580	109.91
DE_LU	21.76	13.63	15.88	0.463	88.35
FR	22.51	16.07	20.91	0.665	100.37
PL	19.33	13.00	12.59	0.529	73.86
ES	16.67	11.79	16.31	0.534	66.48
NO_2	19.91	13.95	20.91	0.808	88.09

Having established the methodological approach and architecture, we can now evaluate the performance of the model. In general, we see that the performance of the new version is significantly better than the old version of Chronos, this is likely mainly because of the introduction of covariates, which in the electricity market are of great importance. We observe that the most difficult bidding zone to forecast remains to be Denmark, with a RMSE (27,2) and MAE (18,23). In contrast the other highly volatile bidding zone, Germany is now within the middle range of performance, this may be caused by the overall structure of the German market. While Germany has a lot of wind power, it also has a massive solar generation capacity, which as established earlier, introduces a very predictable pattern, because this pattern is identifiable for Chronos V2, the model performs better in Germany than in Denmark. In Germany the model achieves an RMSE (21,7) and MAE (13,6). The best performance is observed in Spain, here we see an RMSE (16,67) and MAE (11,79), this is because, Spain's electricity production follows a highly predictable pattern. This is also reflected in the Winkler score, which in Spain is 66,48 and in comparison in Denmark it is 110, showing that the unpredictability of Denmark makes the model have bigger prediction bounds or gets penalized more because of unpredicted price spikes.

5.3.4 Times FM

Developed by Google research, the Time Series Foundation model (TimesFM) represents a different approach to zero-shot forecasting. While the Chronos models relies on LLM style tokenization and Time GPT functions as a black box, TimesFM is a open-weights, decoder-only transformer designed to process continuous time series directly (Das *et al.*, 2024; Google, 2024). To increase predictive accuracy and probabilistic forecasts, this thesis deploys the newest iteration of the model, TimesFM 2.5.

TimesFM uses a patching mechanism. Instead of feeding the model data point by point, the continuous timeline is chopped into fixed size, non-overlapping patches. These patches are then translated into the network's latent space. Finally, positional encodings are attached to each patch, acting as timestamps to ensure the transformer perfectly retains the original order (Das *et al.*, 2024). TimesFM uses the Multi-Head-Self-Attention mechanism.

The TimesFM 2.5 model used in this thesis features 200 million parameters and was pre-trained on a massive training set containing over 100 billion real-world and synthetic data points (Das *et al.*, 2024). The architecture of the TimesFM model is illustrated in the appendix (A8).

Implementing a pre-trained foundation model for Day-Ahead electricity markets requires strict adherence to the market gate closure constraint. To ensure a valid comparison across all foundation models, the exact same gate closure methodology applied to Time-GPT and Chronos was implemented here. The context window was configured to 1,024 hours, allowing the models attention mechanism to capture trends alongside weekly cycles. The model is fed with exogenous covariates.

As with Time-GPT and Chronos, deploying a pre-trained model introduces the risk of data memorization. TimesFM is continuously updated, the newest version released in April 2026. Although the exact cut-off date of TimesFM's training set is not explicitly published, European ENTSO-E electricity data up to late 2024 was likely included in the pre-training mix. This must be accounted for when evaluating its out-of-sample performance.

Table 5.10: TimeFM Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	31.24	22.22	28.45	0.707	221.31
DE_LU	35.41	15.79	18.42	0.536	172.11
FR	23.06	17.33	22.59	0.717	169.15
PL	20.23	14.06	13.62	0.572	133.87
ES	18.23	13.78	19.10	0.624	129.88
NO_2	23.53	15.93	23.85	0.922	175.52

Evaluating the out-of-sample performance of the TimesFM model demonstrates the potential and limitations of zero-shot foundation models in Day-Ahead electricity markets. The model achieved a MASE ranging from 0.536 to 0.922 across all six bidding zones. Although the model demonstrated high capabilities in some regions it failed to provide a significant improvement in others.

The empirical results reveal a fascinating result in the German (DE_LU) zone. TimesFM delivered an exceptional MAE of 15.79 and the lowest MASE across all six bidding zones at 0.536. However, DE_LU recorded the recorded the highest RMSE in the dataset at 35.41. This discrepancy implies that the TimesFM model accurately captured the dynamics of the market for the majority of hours, but the model struggles to accurately capture the price spikes, resulting in a penalization of the RMSE metric.

In zones dominated by stable baseloads or predictable daily cycles, TimesFM exhibited its strongest absolute performance. Poland (PL), supported by its predictable coal baseload, achieved an excellent MAE of 14.06 and the lowest WAPE of 13.62%. The solar heavy Spanish (ES) zone recorded the lowest RMSE across the six bidding zones (18.23) and the lowest MAE (13.78). This confirms that the TimesFM model excels at isolating and extrapolating highly deterministic patterns, such as solar generation and thermal dispatch.

The Winkler Scores remained bounded between 63.54 (ES) and 104.83 (DK1). This structural stability confirms that the model accurately maps the true underlying risk of the market.

The Southern Norway (NO_2) zone still struggles with the problem of high autocorrelation due to its reliance on hydro-power. This results in a weak MASE score of 0.922, although it achieved a relatively moderate RMSE (23.53) and MAE (15.93). This highlights that the TimesFM model struggles to outperform simple naïve benchmarks in zones driven by strong autocorrelation.

One interesting note on the residual diagnostics (A6.9) of the TimesFM model is that a divergence occurs in Germany. Where TimesFM generates an extreme kurtosis and a large positive skewness. This highlights a zero-shot inference failure where the model completely missed a massive price spike. Interestingly, this specific DE_LU outlier is the only instance where the ARCH-LM test confirms stable variance.

5.3.5 Lag-Llama

To complete the foundational model analysis, this study evaluates the Lag-Llama model. Lag-llama has a decoder-only transformer architecture. The most significant difference relating to the other univariate foundation model, is that it doesn’t group data into text-based batches. When the model has seen the data and evaluated it, the model tries to predict the distribution based on the students t-distribution, predicting the mean, variance and degrees of freedom. The students t-distribution is characterized by its fat tails, equipping Lag-Llama to handle the high volatility and volatility clustering that is observed in the Day-Ahead electricity market. An illustration of the architecture of the Lag-Llama model can be seen in appendix (A9). Lag-Llama has a severe limitation; the model is strictly univariate. The performance of the model will show how well a complex univariate foundation model performs.

The methodological approach for this model is the same as the others, we implement a strict gate closure, preventing forward bias. Because Lag-Llama outputs the parameters of the distribution rather than a point forecast, Monte Carlo sampling is utilized to generate the possible outcomes.

Specifically, the model did this 20 times for each forecast from the distribution, then the 5th, 50th and 95th percentiles are extracted to generate the point forecasts and its bounds. Critically, 20 simulations is a very low number, but because of computational constraints this was the achievable number of simulations. A much better and precise result could be achieved with a higher number of simulations.

Table 5.11: Lag-Llama Model Performance Metrics

Country	RMSE	MAE	WAPE (%)	MASE	Winkler Score
DK_1	53.95	40.92	52.37	1.302	318.63
DE_LU	58.12	43.64	50.85	1.481	347.81
FR	50.32	38.28	49.86	1.583	294.29
PL	59.96	43.23	41.89	1.758	366.94
ES	50.00	37.54	51.97	1.700	299.21
NO_2	34.14	23.77	35.64	1.376	223.17

In general, we observe that the model is significantly worse at predicting the price across all bidding zones. These results are a combination of the model only being able to handle univariate input and the small number of Monte Carlo simulations. Its best performance across the bidding zones is in Norway, here we observe a RMSE (34,14) and MAE (23,77), the significantly better performance in Norway is partly because of the large hydro generation that stabilizes the price. However, in the other more covariate dependent zones, the lack of covariates places a significant burden on Lag-Llama. In Poland (PL) which relies heavily on coal power, the inability to ingest the CO2 and TTF prices result in the highest RMSE (59,96). Similarly in other countries with high influence from wind and solar, such as Denmark, Germany and Spain, we observe an RMSE ranging from 58,12 to 50. The predictive failure really arises when looking at MASE, here we observe that all countries have a MASE much large over 1, showing that the model performs significantly worse than a simple seasonal naïve model. We also observe that the models is accompanied with very high Winkler Scores, because the model is blind to exogenous shocks, the model recognizes the high uncertainty and outputs very wide prediction intervals. Ultimately, we see that the empirical performance of a univariate model, even with its complex architecture, is simply not sufficient to understand the markets dynamics, resulting in the bad performance that has just been covered.

6. Comparison

Having established the theoretical architectures and evaluated the individual out-of-sample performance of the twelve forecasting models, this chapter combines the empirical results.

The objective is covering a comprehensive cross model comparison, benchmarking classical time series model against machine learning and zero shot foundation model. This analysis tries to identify the current state-of-the-art of Day-Ahead electricity price forecasting across the six selected bidding zones.

Table 6.1: Comparative Model Performance Metrics

Country	Model	RMSE	MAE	WAPE (%)	Winkler Score
DK_1	Daily Naïve	44.48	31.43	40.25	243.13
	Weekly Naïve	49.18	35.52	45.49	287.97
	ETS†	55.88	42.12	53.94	249.41
	SARIMAX	37.34	28.01	35.86	215.35
	MSTL-ARIMAX	37.25	26.84	34.38	208.53
	CatBoost	25.70	18.15	23.27	137.84
	NHITS	26.66	18.32	24.22	134.60
	TimeGPT	28.04	19.71	25.22	279.30
	Chronos	40.76	28.26	36.15	210.83
	ChronosV2	27.20	18.23	23.32	109.91
	TimesFM	31.24	22.22	28.45	221.31
	Lag-Llama	53.95	40.92	52.37	318.63
DE_LU	Daily Naïve	43.13	29.47	34.37	247.13
	Weekly Naïve	44.83	31.26	36.47	281.86
	ETS	56.65	43.08	50.26	264.01
	SARIMAX	36.63	26.83	31.26	207.44
	MSTL-ARIMAX	32.70	23.55	27.47	201.20
	CatBoost	18.53	12.12	14.13	94.28
	NHITS	25.42	16.14	19.35	115.69
	TimeGPT	21.88	14.50	16.91	212.02
	Chronos	38.77	26.06	30.35	207.01
	ChronosV2	21.76	13.63	15.88	88.35
	TimesFM	35.41	15.79	18.42	172.11
	Lag-Llama†	58.12	43.64	50.85	347.81
FR	Daily Naïve	34.16	24.18	31.51	201.05
	Weekly Naïve	37.55	27.42	35.74	221.78
	ETS	44.89	34.11	44.45	219.50
	SARIMAX	31.08	24.22	31.56	168.97
	MSTL-ARIMAX	29.31	21.59	28.14	186.98
	CatBoost	22.21	16.26	21.24	138.37
	NHITS	26.60	14.97	20.51	103.97
	TimeGPT	25.78	18.64	24.29	258.03
	Chronos	29.49	21.68	28.21	170.51
	ChronosV2	22.51	16.07	20.91	100.37
	TimesFM	23.06	17.33	22.59	169.15
	Lag-Llama†	50.32	38.28	49.86	294.29
PL	Daily Naïve	36.28	24.59	23.83	206.12
	Weekly Naïve	36.53	25.75	24.95	206.76
	ETS†	61.57	43.46	42.11	358.48
	SARIMAX	30.52	21.14	20.49	185.60
	MSTL-ARIMAX	31.67	22.31	21.62	173.37
	CatBoost	19.87	13.35	12.94	111.32
	NHITS	23.37	15.61	15.42	115.48
	TimeGPT	20.47	14.22	13.78	200.46
	Chronos	30.39	21.00	20.33	160.61
	ChronosV2	19.33	13.00	12.59	73.86
	TimesFM	20.23	14.06	13.62	133.87
	Lag-Llama	59.96	43.23	41.89	366.94
ES	Daily Naïve	32.61	22.08	30.60	182.93
	Weekly Naïve	37.54	26.94	37.33	199.18
	ETS	47.05	36.24	50.22	281.97
	SARIMAX	27.20	20.51	28.40	161.51
	MSTL-ARIMAX	28.85	20.39	28.26	143.45
	CatBoost	17.53	12.87	17.87	106.82
	NHITS	20.10	13.28	18.50	101.65
	TimeGPT	20.58	14.66	20.30	213.07
	Chronos	29.16	20.15	27.88	155.86
	ChronosV2	16.67	11.79	16.31	66.48
	TimesFM	18.23	13.78	19.10	129.88
	Lag-Llama†	50.00	37.54	51.97	299.21
NO_2	Daily Naïve	25.45	17.27	25.85	153.88
	Weekly Naïve	29.77	21.01	31.46	191.74
	ETS†	36.92	26.43	39.58	188.35
	SARIMAX	25.85	19.31	29.01	161.50
	MSTL-ARIMAX	31.66	21.22	31.78	147.30
	CatBoost	20.71	14.28	21.58	116.43
	NHITS	15.83	10.66	16.81	99.41
	TimeGPT	23.34	16.66	24.96	228.98
	Chronos	22.00	15.44	23.14	108.98
	ChronosV2	19.91	13.95	20.91	88.09
	TimesFM	23.53	15.93	23.85	175.52
	Lag-Llama	34.14	23.77	35.64	223.17

Best performance (underlined), Worst performance (†)

6.1 Linear vs. non-linear inference

A not surprising but profound finding derived from the metrics, is the failure of the classical time series models, when subjected to the volatility of the merit-order dynamics.

Across all six evaluated bidding zones, the classical linear baselines (SARIMAX, MSTL-ARIMAX, and ETS) consistently yielded the highest point-forecast errors. The pure univariate ETS model performed the worst, scoring high RMSEs ranging from 36.92 to 61.57. While the integration of covariates and MSTL decomposition significantly improved the linear baseline, these models remained constrained by their autoregressive nature, struggling to capture price spikes induced by sudden wind burst or inelastic supply shocks.

On the other hand, architectures capable of non-linear inference, the gradient boosted decision trees (CatBoost) and the deep neural networks (NHITS) established an improved predictive performance. CatBoost dramatically reduced the RMSE across the six bidding zones. In Germany, CatBoost yielded an RMSE of 18.53 and an MAE of 12.12, representing an approximate 43% reduction in error compared to the MSTL-ARIMAX baseline. The NHITS models yielded similarly great results, achieving the lowest RMSE in Norway (15.83). These metrics confirm that to successfully forecast the high volatility of the Day-Ahead electricity market, forecasting algorithms must possess the capacity to capture non-linear relationships.

6.2 Multivariate models vs. univariate zero-shot models

The introduction of foundation models introduces an interesting debate. Does the large parametrization and zero-shot pattern recognition outweigh the inclusion of exogenous covariates.

The empirical results provide a definitive answer. The univariate foundation models (Lag-Llama and Chronos (t5-small)) were outperformed by the multivariate models across every single bidding zone. Lag-Llama, despite utilizing a students t-distribution to map fat-tailed variance, yielded poor point-forecast errors, yielding an RMSE of 59.96 in Poland and 53.95 in Denmark. Similarly, although the LLM-style tokenization of Chronos (t5-small) performed adequately in highly autocorrelated zones, achieving an RMSE of 21.99 in Norway, it failed entirely in volatile zones, yielding an RMSE of 40.75 in Denmark.

A univariate model is blind to incoming weather fronts, geopolitical gas shocks, and bottlenecks in the interconnections. The necessity of the multivariate data is proven by examining the updated foundation models. When the Chronos model was updated to the V2 version, allowing it to integrate exogenous covariates, its performance improved drastically. In Germany, the introduction of covariates caused the Chronos RMSE to plummet from 38.77 (t5-small) to 21.75 (V2). Similarly, the multivariate TimesFM 2.5 model established itself as a competent model, achieving the absolute lowest MAE in Germany (15.79) and Spain (13.78) among the foundation models.

The cross-model metrics prove that while zero-shot models are powerful, their generalized knowledge is insufficient for Day-Ahead electricity markets unless coupled with exogenous covariates that represents the actual reality of the load and generation potential.

6.3 Prediction intervals

In Day-Ahead electricity markets, point forecasts are inherently insufficient for utility operators and traders who must optimize bidding strategies under uncertainty. The accuracy of a models intervals are as important as its point forecast predictions.

Classical time series models assume that residual errors are independent and normally distributed. As established in the residual diagnostics, the errors exhibit severe excess kurtosis and volatility clustering. When models like MSTL-ARIMAX and ETS attempt to construct 95% prediction intervals using normal standard deviations, they are penalized by the Winkler Score for failing to capture fat-tailed events. This is empirically evident in the metrics: the classical time series models generated Winkler Scores ranging from 140 to well over 300 across the six evaluated bidding zones.

The modern algorithms bypassed the parametric assumptions by forecasting the distribution directly. By utilizing the Multi-Quantile Loss function, both CatBoost and NHITS learned the distribution of the market, shrinking their prediction intervals during stable periods and expanding them only during true volatile periods. This resulted in much lower Winkler scores ranging from 94 to 138.

The absolute best intervals were achieved by the multivariate foundation models. TimesFM 2.5 and Chronos V2 achieved the lowest Winkler scores in the metrics. These scores confirm that pre-trained foundation models when supplied with exogenous covariates, do not just predict the mean price accurately but they also map the true underlying uncertainty of the Day-Ahead electricity market. Although one must consider the risk of data memorization.

6.4 Statistical Significance: Diebold-Mariano

The Error metrics provide a clear hierarchy of predictive performance; they do not inherently prove that the variance between models is statistically significant. In the finite out-of-sample periods, a marginally lower RMSE could just be the result of random noise rather than a true improvement in predictive power. To confirm the comparative findings, the Diebold-Mariano (DM) test was executed across all model pairings for each bidding zone.

In the constructed tables (A8) (blue) indicates that the challenger model (column) outperforms the benchmark model (row), and (red) indicates that the challenger model (column) underperforms the benchmark model (row). The intensity of the color represents the DM test statistic. Furthermore, the p-values are shown through star markings in each cell.

When the classical time series models (Daily Naïve, Weekly Naïve, SARIMAX, MSTL-ARIMAX, ETS) act as the benchmark, the machine learning models (CatBoost, NHITS) the DM test statistics are positive at the $p < 0.01$ significance level. This proves that the transition to non-linear yields statistically significant predictive gains.

When univariate foundation models (Chronos t5-small, Lag-Llama) act as the benchmark rows against their multivariate counterparts (Chronos V2, TimesFM 2.5), the DM test statistics are positive. Proving that integrating exogenous covariates is an absolute necessity for Day-Ahead electricity price forecasting.

The introduction of zero-shot foundation models introduces another debate. Does the massive pretraining and generalized pattern recognition of foundational models outperform localized machine learning algorithms. The DM test reveals that the answer is highly specific to the bidding zone. In bidding zones subject to extreme volatility, such as Denmark and Germany we observe that CatBoost outperforms all foundation models, with statistically significant results. For NHITS it’s observed that it only outperforms the univariate foundation models in the German bidding zone, all results in this bidding zone are also statistically significant at the one percent level. In the Danish bidding zone, we observe that NHITS outperforms all foundation models except Time-GPT and Chronos V2, although it must be said that for those two models, there is no statistical significance at the five percent level. In contrast, the more stable bidding zones such as Norway and Spain, its observed that CatBoost outperforms all foundation models except Chronos V2. In Spain, NHITS outperforms only the univariate foundation models. In contrast NHITS manages to outperform all foundation models in the South Norwegian bidding zone, all with statistically significant results at the one percent level.

Foundation and machine learning models perform competitively across bidding zones, although for highly volatile and complex nonlinear relationships, machine learning clearly outperforms foundation models.

Conclusion

This thesis empirically evaluated the predictive performance of classical time series models, machine learning algorithms and zero shot foundation models across six heterogeneous European Day-Ahead electricity markets. The results prove that the ongoing integration of renewable sources has changed the market dynamics, resulting in classical time series models such as SARIMAX, ETS, MSTL-ARIMAX, uncapable of capturing the highly complex nonlinear relationships. The transition to nonlinear architectures is therefore necessary. Machine learning models such as CatBoost and NHITS demonstrated statistically significant predictive gains over classical models. Furthermore, the analysis revealed that even though state of the art foundation models are the most complex, this is not sufficient to offset the need for covariates, hence univariate zero shot forecasting is inadequate. However, when multivariate foundation models like Chronos V2 and Times FM are supplied with covariates, they deliver highly competitive forecasts. Nevertheless, because pre-trained foundation models carry an unquantifiable risk of data leakage, inflicted by the memorization bias, localized machine learning algorithms remain the most trustworthy. Concluding this thesis, multivariate foundation models perform competitively, but localized machine learning models remain the most robust and reliable tools for electricity price forecasting, not only because they excel in highly volatile bidding zones, but also because their localized training guarantees validity in their out-of-sample performance.

Appendix

A1: ADF and KPSS test

Table A1: ADF & KPSS test

Bidding Zone	ADF p-value	ADF Stat	KPSS p-value	KPSS Stat
DK_1	0.0	-8.1136	0.01	5.7599
DE_LU	0.0	-7.8446	0.01	6.0907
ES	0.0	-6.0456	0.01	5.9942
FR	0.0	-5.8697	0.01	5.7935
NO_2	0.0	-5.2383	0.01	6.2042
PL	0.0	-8.4709	0.01	10.6112

A2: VIF test

Table A2: Variance Inflation Factor (VIF) per Bidding Zone

Bidding Zone	Variable	VIF
DK_1	Forecasted Load	1.12
	Total_Wind_Forecast_MW	1.30
	Solar_Forecast_MW	1.19
	NTC_Import_GB_to_DK_1	63.74
	NTC_Export_DK_1_to_GB	63.40
	NTC_Export_DK_1_to_NL	4.47
	NTC_Import_NL_to_DK_1	4.38
	NTC_Export_DK_1_to_DE_LU	2.24
	NTC_Import_DE_LU_to_DK_1	1.87
	CO2_EUR_ton_Lag1	2.20
	TTF_EUR_MWh_Lag1	2.03
DE_LU	Forecasted Load	1.26
	Total_Wind_Forecast_MW	2.75
	Solar_Forecast_MW	1.31
	NTC_Export_DE_LU_to_DK_2	7.77
	NTC_Import_DK_2_to_DE_LU	6.78
	NTC_Export_DE_LU_to_CH	2.49
	NTC_Import_DK_1_to_DE_LU	2.46
	NTC_Export_DE_LU_to_DK_1	2.14
	NTC_Export_DE_LU_to_CZ	1.53
	NTC_Import_CZ_to_DE_LU	1.46
	NTC_Export_DE_LU_to_NL	1.46
	NTC_Import_NL_to_DE_LU	1.45
	NTC_Import_CH_to_DE_LU	1.21
	NTC_Import_AT_to_DE_LU	0.00
	NTC_Export_DE_LU_to_AT	0.00
	CO2_EUR_ton_Lag1	2.67
	TTF_EUR_MWh_Lag1	2.12
ES	Forecasted Load	1.31
	Total_Wind_Forecast_MW	1.13
	Solar_Forecast_MW	1.28
	NTC_Export_ES_to_FR	1.43
	NTC_Import_FR_to_ES	1.27
	NTC_Export_ES_to_PT	1.22
	NTC_Import_PT_to_ES	1.19
	CO2_EUR_ton_Lag1	2.17
	TTF_EUR_MWh_Lag1	2.04
FR	Forecasted Load	1.54
	Total_Wind_Forecast_MW	1.46
	Solar_Forecast_MW	1.14
	NTC_Export_FR_to_GB	1002.31
	NTC_Import_GB_to_FR	996.48
	NTC_Import_IT_NORD_to_FR	1.72
	NTC_Export_FR_to_CH	1.60
	NTC_Export_FR_to_IT_NORD	1.38
	NTC_Import_ES_to_FR	1.37
	NTC_Import_CH_to_FR	1.29
	NTC_Export_FR_to_ES	1.29
	CO2_EUR_ton_Lag1	2.51
	TTF_EUR_MWh_Lag1	2.20
NO_2	Forecasted Load	1.01
	Total_Wind_Forecast_MW	1.33
	Solar_Forecast_MW	
	NTC_Import_NL_to_NO_2	5.65
	NTC_Export_NO_2_to_NL	5.25
	CO2_EUR_ton_Lag1	2.35
	TTF_EUR_MWh_Lag1	2.17
PL	Forecasted Load	1.07
	Total_Wind_Forecast_MW	1.13
	Solar_Forecast_MW	1.18
	NTC_Export_PL_to_SK	5.16
	NTC_Import_SK_to_PL	3.87
	NTC_Import_CZ_to_PL	1.27
	NTC_Export_PL_to_CZ	1.10
	CO2_EUR_ton_Lag1	2.52
	TTF_EUR_MWh_Lag1	2.03

A3: Granger Causality test

Table A3: Granger Causality Test (Lag 24)

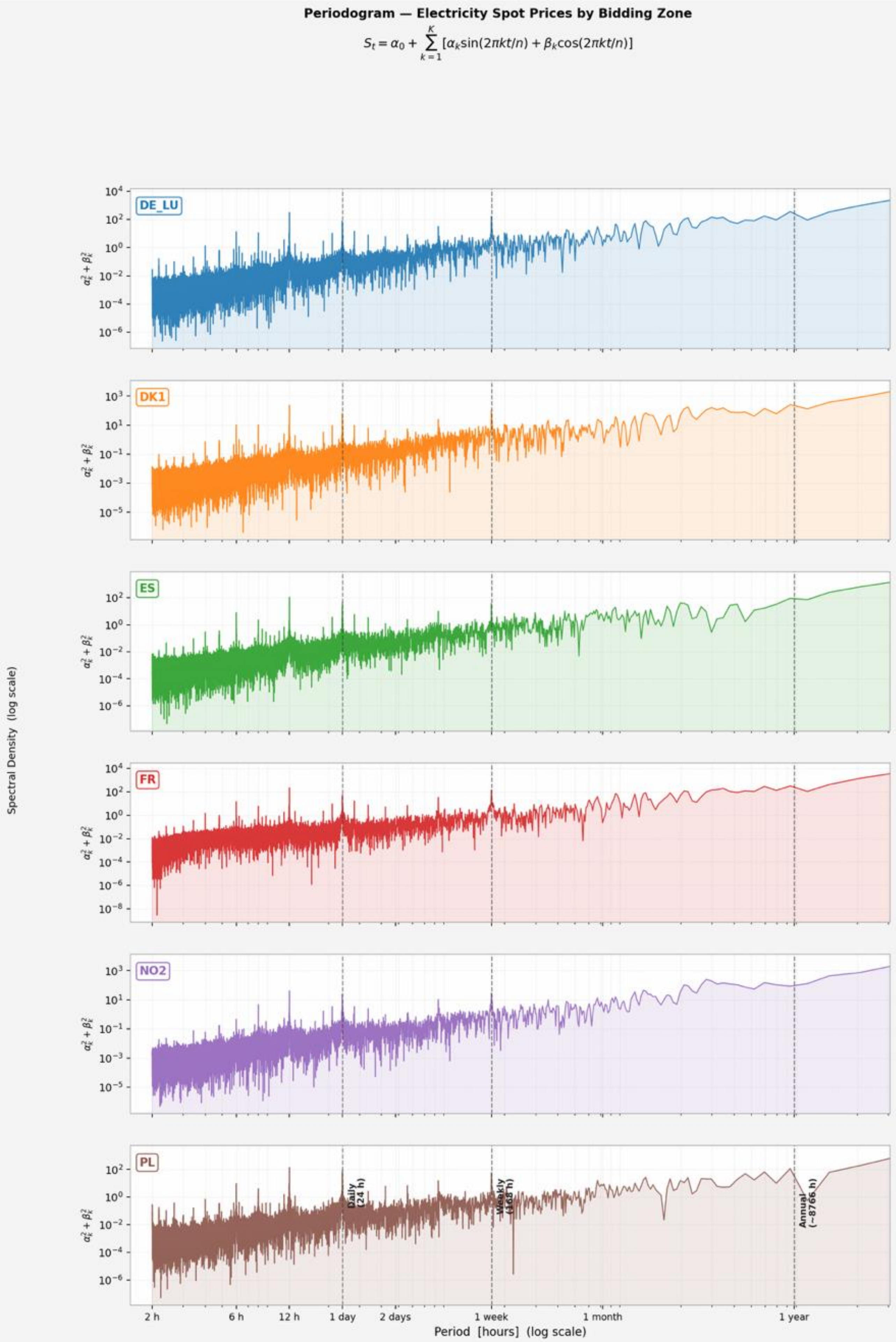
Bidding Zone	Exogenous Variable	Lag Hours	P-value
DK_1	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	0.0
	NTC_Export_DK_1_to_DE_LU	24	0.0
	NTC_Import_DE_LU_to_DK_1	24	0.0
	NTC_Export_DK_1_to_NL	24	0.0
	NTC_Import_NL_to_DK_1	24	0.0027
	CO2_EUR_ton_Lag1	24	0.0029
	TTF_EUR_MWh_Lag1	24	0.0
DE_LU	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	0.0
	NTC_Export_DE_LU_to_DK_1	24	0.0
	NTC_Import_DK_1_to_DE_LU	24	0.0
	NTC_Export_DE_LU_to_DK_2	24	0.0022
	NTC_Import_DK_2_to_DE_LU	24	0.004
	NTC_Export_DE_LU_to_CZ	24	0.0
	NTC_Import_CZ_to_DE_LU	24	0.0001
	NTC_Export_DE_LU_to_NL	24	0.3896
	NTC_Import_NL_to_DE_LU	24	0.0001
	NTC_Export_DE_LU_to_CH	24	0.0
	NTC_Import_CH_to_DE_LU	24	0.0
	NTC_Export_DE_LU_to_AT	24	Singular Matrix
	NTC_Import_AT_to_DE_LU	24	Singular Matrix
	CO2_EUR_ton_Lag1	24	0.6187
	TTF_EUR_MWh_Lag1	24	0.0
ES	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	0.0
	NTC_Export_ES_to_FR	24	0.0
	NTC_Import_FR_to_ES	24	0.0
	NTC_Export_ES_to_PT	24	0.0
	NTC_Import_PT_to_ES	24	0.0
	CO2_EUR_ton_Lag1	24	0.003
	TTF_EUR_MWh_Lag1	24	0.0
FR	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	0.0
	NTC_Export_FR_to_IT_NORD	24	0.0
	NTC_Import_IT_NORD_to_FR	24	0.0
	NTC_Export_FR_to_CH	24	0.4602
	NTC_Import_CH_to_FR	24	0.1878
	NTC_Export_FR_to_ES	24	0.0
	NTC_Import_ES_to_FR	24	0.0
	CO2_EUR_ton_Lag1	24	0.8071
	TTF_EUR_MWh_Lag1	24	0.0
NO_2	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	Singular Matrix
	NTC_Export_NO_2_to_NL	24	0.0112
	NTC_Import_NL_to_NO_2	24	0.9733
	CO2_EUR_ton_Lag1	24	0.0
	TTF_EUR_MWh_Lag1	24	0.0
PL	Forecasted Load	24	0.0
	Total_Wind_Forecast_MW	24	0.0
	Solar_Forecast_MW	24	0.0
	NTC_Export_PL_to_CZ	24	0.0271
	NTC_Import_CZ_to_PL	24	0.0
	NTC_Export_PL_to_SK	24	0.3999
	NTC_Import_SK_to_PL	24	0.9946
	CO2_EUR_ton_Lag1	24	0.4302
	TTF_EUR_MWh_Lag1	24	0.0

A4: Spectral analysis - Periodogram

It is well known that one of the most important characteristics of electricity markets is that they are highly seasonal. This seasonality arises from both sides of the bidding process, supply and demand. For example, the hydro plants in Norway are highly dependent on the temperature change that comes from seasonal fluctuations, because it makes the snow melt, letting the water fall down the mountain and hence, producing electricity. For the demand side this is more prevalent in different timeframes also, because of the habits of humans and the demand for electricity changing across the year. The results of this, is that we observe mainly 3 different frequencies of seasonal pattern. The highest frequency pattern is the daily one. This is driven by the human and industrial activity, We typically observe that prices peak during morning hours and evening hours, which is when people go to work and get back from work, with the lowest demand being at midday and at night. Another cycle that can be observed in the prices is the week effect. The seasonality here is the weekend effect. Where because of the change in industrial activity in weekends, the prices are generally lower (Weron, 2014b). The last seasonal pattern is the annual cycle. This is mainly driven by the changes in weather seasons, temperature, macroeconomic trends etc. Because the seasonal patterns play such a big role in electricity prices, it is crucial to make sure that your models understand these patterns. First, we have to be able to quantify that this pattern exists in our data. For this we will use spectral analysis via Fourier decomposition (Rob Hyndman, G. Athanasopoulos, 2021). The intuition is that any seasonal pattern can be expressed as a frequency, making it possible to explain and model complex seasonality as a linear combination of sine and cosine functions, which vary in amplitude. Mathematically this looks as follows.

$$S_t = \alpha_0 + \sum_{k=1}^{K} \left[\alpha_k \sin\left(\frac{2\pi kt}{n}\right) + \beta_k \cos\left(\frac{2\pi kt}{n}\right) \right]$$

We view this as a multiple linear regression which we must fit to the data, hence the dependent variable is the time series, and the independent variables are the sine and cosine functions. This means that α_k and β_k are interpreted as the coefficients that determine the strength of the wave, k represents the waves that are stacked on top of each other, because price movements are not smooth and do not have only one peak, we add the term k which stacks different frequencies on top of each other, and n is the length of the seasonal period per unit of time. With this we can now plot the so called periodogram, which plots the spectral density, which can be interpreted as the variance explained against the different frequencies. When evaluating the periodogram, if we observe the strong seasonality we expect, this means that extra seasonal components will have to be added, since the way for example the ARIMA model handles seasonality won't be sufficient hereby dummies, and other season decompositions such as MSTL must be added.



Periodogram — Electricity Spot Prices by Bidding Zone
$S_t = \alpha_0 + \sum_{k=1}^{K} [\alpha_k \sin(2\pi kt/n) + \beta_k \cos(2\pi kt/n)]$
DE_LU
DK1
ES
FR
NO2
PL
$\alpha_k^2 + \beta_k^2$
Spectral Density (log scale)
Daily (24 h)
Weekly (168 h)
Annual (~8766 h)
2 h
6 h
12 h
1 day
2 days
1 week
1 month
1 year
Period [hours] (log scale)

A5: SARIMAX features

Daily_Avg_Price_Lag1	The 24 hour mean price of the previous day
Peak_Avg_Price_Lag1	The mean price during peak hours (08:00 - 19:00)
OffPeak_Avg_Price_Lag1	The mean price during off peak hours

A6: Residual Diagnostics

Each models residual diagnostics table is presented below

A6.1: ETS residual diagnostics

Table A6.1: Residual Diagnostics: ETS

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−0.94	55.87	0.937	0	0	0	−0.37	3.99	0
DE_LU	−1.05	56.64	0.935	0	0	0	−0.03	2.48	0
FR	−1.60	44.86	0.933	0	0	0	−0.01	0.90	0
PL	−4.43	61.41	0.928	0	0	0	−0.51	5.24	0
ES	−1.07	47.04	0.943	0	0	0	−0.09	0.59	0
NO_2	−2.60	36.83	0.936	0	0	0	−0.83	5.66	0

Evaluating the residual diagnostics for the ETS framework reveals severe structural limitations in its ability to model the complex dynamics of the European Day-Ahead electricity market. As observed the residuals systematically violate the core econometric assumptions of white noise and normal distribution across all bidding zones. Its observed that the model is incapable of filtering our temporal dependence, as shown in the LB test for 24 and 168 hours. This indicates that the univariate framework is unable to extract all autoregressive information, ultimately failing to capture the prominent seasonality of the European electricity market. When evaluating the distributional properties its observed that the Jarque-Bera test rejects normality for all bidding zones. Additionally, we observe non-zero mean errors, indicating that the model systematically over predicts the actual prices. We also observe significant excess kurtosis ranging from 0,59 to 5,66. This fat tailed asymmetric distribution highlights that the model is unable to adapt to price spikes created by the merit order effect, although this is likely because of its univariate nature.

Ultimately, these diagnostics empirically confirm that, the univariate ETS framework is inadequate for modelling price dynamics in the European Day-Ahead electricity market.

A6.2 SARIMAX residual diagnostics

Table A6.2: Residual Diagnostics: SARIMAX

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	2.32	37.27	0.898	0	0	0	−0.24	5.41	0
DE_LU	2.51	36.54	0.878	0	0	0	0.05	8.72	0
FR	6.21	30.45	0.904	0	0	0	0.01	0.84	0
PL	2.46	30.42	0.857	0	0	0	1.37	10.89	0
ES	1.89	27.14	0.908	0	0	0	0.30	1.47	0
NO_2	−5.19	25.32	0.870	0	0	0	0.16	1.85	0

Residual diagnostics across all six bidding zones expose limitations in the SARIMAX framework. The Ljung-Box test rejects no autocorrelation at 24- and 168-hour lags, revealing that unextracted price signals remain in the error terms. The ARCH-LM test robustly rejects homoskedasticity, confirming the model's inability to capture volatility clustering caused by persistent exogenous shocks. This inadequacy is compounded by severe non-normality, with positive skewness and extreme excess kurtosis reaching 412.35 in France. Hence, the model produces massive, errors during non-linear pricing extremes. Ultimately, although functional during stable regimes, SARIMAX's linear architecture is fundamentally unsuited for the extreme, asymmetric volatility characteristic of European Day-Ahead electricity markets.

A6.3 MSTL-ARIMAX residual diagnostics

Table A6.3: Residual Diagnostics: MSTL-ARIMAX

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−2.36	37.18	0.919	0	0	0	−0.17	4.72	0
DE_LU	−0.97	32.68	0.904	0	0	0	0.13	8.95	0
FR	−0.96	29.29	0.914	0	0	0	0.01	2.01	0
PL	1.04	31.65	0.890	0	0	0	0.07	4.10	0
ES	1.61	28.80	0.926	0	0	0	0.21	2.65	0
NO_2	−5.15	31.24	0.933	0	0	0	−1.48	7.74	0

When running the diagnostics on the MSTL-ARIMAX model we observe the following. Across all bidding zones, the models errors are not white noise. We observe this in the Ljung-box test, where the p-value is 0 for all of them, at lag 24 and 168, confirming there is information that is being left behind. These results hints at that the model is accurate enough during a stable market, but the errors explode when shocks occur. Germany for instance shows this the most, with an excess kurtosis of 8,95, showing the fat tails of the distribution. This indicates that when the model misses, it misses by a massive amount. We also observe in Norway severe downward bias in the errors at -5,15, this in combination with a skewness of -1,48. This tells us that the model is failing to anticipate the downward crashes in price caused by hydropower. In general, we can conclude, that the classical linear model MSTL-ARIMAX is too rigid for the highly complex and volatile electricity markets.

A6.4 CatBoost residual diagnostics

Table A6.4: Residual Diagnostics: CatBoost

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−0.62	25.69	0.871	0	0	0	−0.35	19.47	0
DE_LU	2.34	18.38	0.783	0	0	0	1.50	74.26	0
FR	−0.99	22.19	0.882	0	0	0	−0.31	2.29	0
PL	3.18	19.61	0.793	0	0	0	1.20	19.45	0
ES	−1.33	17.48	0.856	0	0	0	−0.05	2.34	0
NO_2	−0.80	20.69	0.910	0	0	0	0.20	4.61	0

The diagnostics reveals that even though the model is generally good at predicting the price, we still observe that this tree-based architecture is not able to capture all available information and use it correctly. We observe that results are biased, especially in Germany and Poland. We also observe that the model cannot completely filter through the autocorrelation, the Ljung-Box test at lag 24 and 168 rejects the H_0. We also observe a skewness close to zero for most of the bidding zones, but very high excess kurtosis specially in Denmark, Germany and Poland. This indicates that while CatBoost gets most errors close to zero, it still completely misses the massive price spikes that can occur, resulting in some outliers.

A6.5 NHITS residual diagnostics

Table A6.5: Residual Diagnostics: NHITS

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−1.20	26.64	0.849	0	0	0	0.56	8.39	0
DE_LU	−0.01	25.42	0.836	0	0	0	1.81	24.37	0
FR	−0.98	26.59	0.865	0	0	0	−10.09	323.35	0
PL	0.49	23.37	0.797	0	0	0	0.40	9.79	0
ES	−0.90	20.08	0.874	0	0	0	−2.67	33.59	0
NO_2	−1.79	15.73	0.863	0	0	0	0.04	6.00	0

To validate the statical integrity of the NHITS model, diagnostic test were conducted on the out-of-sample residuals. Unlike the classical time series models, NHITS significantly reduced residual bias, with mean errors centering close to zero across most zones. However, the Ljung-Box, ACF, and ARCH-LM tests confirm persistent autocorrelation and volatility clustering. This indicates that although the NHITS models hierarchical addition successfully captures market trends, some high frequency hourly stochasticity remains uncaptured by the model.

A6.6 Time-GPT residual diagnostics

Table A6.6: Residual Diagnostics: Time-GPT

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−2.23	27.95	0.852	0	0	0	−0.38	11.68	0
DE_LU	−1.92	21.79	0.785	0	0	0	−0.34	43.69	0
FR	−1.27	25.75	0.878	0	0	0	0.08	2.24	0
PL	−1.26	20.43	0.770	0	0	0	0.16	6.86	0
ES	−2.14	20.47	0.856	0	0	0	−0.29	2.53	0
NO_2	−0.37	23.34	0.904	0	0	0	0.09	2.51	0

When evaluating the errors of the forecast we observe that the model lacks the ability to completely filter out the autocorrelation in the errors. We also observe that the model is consistently biased in the errors, overpredicting the price for all models. However, when looking at the skewness of the errors we observe that they are mostly within a normal range. The excess kurtosis clearly shows that the volatility in Denmark and Germany gives some outliers that the model cannot predict, because of the merit order effect.

A6.7 Chronos t5-small residual diagnostics

Table A6.7: Residual Diagnostics: Chronos t5-small

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	2.46	40.68	0.917	0	0	0	0.82	10.39	0
DE_LU	1.03	38.76	0.897	0	0	0	1.03	17.36	0
FR	2.19	29.41	0.901	0	0	0	0.24	1.56	0
PL	−0.41	30.38	0.849	0	0	0	0.19	7.15	0
ES	3.04	29.00	0.911	0	0	0	0.22	2.76	0
NO_2	1.33	21.96	0.913	0	0	0	0.27	3.02	0

Evaluating the residual diagnostics of the Chronos 5-small foundation model, reveals the consequences of its unique LLM-style tokenization architecture. While the model brings a zero-shot approach, the residuals confirm that its univariate nature and discrete binning system is not sufficient to capture the market dynamics. We observe that the model is unable to capture the temporal dependencies, as shown by the LB test for 24 and 168 hours. This indicates that relying solely on past prices within a 512 context window and no covariates, leaves critical information behind. Furthermore, we observe these limitations in the most volatile zones, which are prone to price spikes, Denmark and Germany. In these zones we see an excess kurtosis of 10,39 and 17,36. This empirically validates the theoretical flaws of its bidding systems. This confirms that while Chronos t5-small performs adequately in stable zones like Norway, its univariate architecture and binning system are incapable of mapping the price dynamics.

A6.8 ChronosV2 residual diagnostics

Table A6.8: Residual Diagnostics: Chronos V2

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	0.22	27.20	0.890	0	0	0	1.62	33.59	0
DE_LU	0.73	21.75	0.844	0	0	0	2.08	64.19	0
FR	−0.24	22.51	0.895	0	0	0	0.32	4.03	0
PL	0.61	19.32	0.804	0	0	0	1.35	14.70	0
ES	0.14	16.67	0.864	0	0	0	0.04	2.89	0
NO_2	−0.17	19.91	0.912	0	0	0	0.12	3.16	0

Evaluating the residual diagnostics for Chronos V2 reveals the impact of integrating covariates and transitioning to a encoder-transformer architecture. Compared to its univariate predecessor, Chronos V2 demonstrates a massive reduction in systematic bias and overall error variance. Despite this, the model fails to capture the temporal dynamics, leaving the LB test to reject its H_0 for 24 and 168 hours. We also observe that in the stable zones, the error is close to normally distributed, like in France, Spain and Norway, whereas in volatile zones we observe distributions with a high amount of excess kurtosis, indicating that the model is unable to predict price spikes. Ultimately, these diagnostics confirm that even with its higher performance, it struggles understanding the dynamics of the European Day-Ahead markets.

A6.9 TimesFM residual diagnostics

Table A6.9: Residual Diagnostics: TimesFM

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−3.26	34.28	0.907	0	0	0	−0.73	10.76	0
DE_LU	0.02	28.35	0.590	0	0	1	20.64	1388.34	0
FR	1.22	27.11	0.912	0	0	0	0.24	3.10	0
PL	−0.89	22.19	0.817	0	0	0	1.35	14.11	0
ES	0.47	19.63	0.891	0	0	0	−0.11	1.17	0
NO_2	−7.84	25.65	0.931	0	0	0	−0.54	3.33	0

We see much the same picture as earlier, with persistent autocorrelation and non-normal error distributions across the board. However, there are two significant differences. First, TimesFM drastically reduces the Mean Error in interconnected and stable zones (DE_LU, FR, ES) compared to the classical time series models, though it worsens the bias in the NO_2 zone.

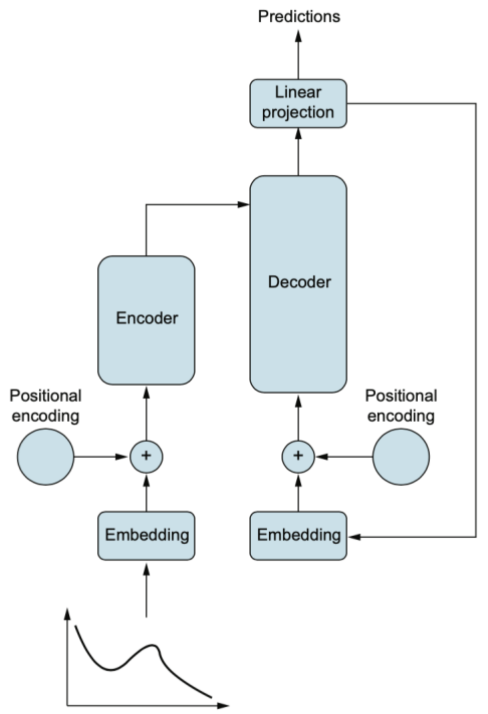
A6.10 Lag-llama residual diagnostics

Table A6.10: Residual Diagnostics: Lag-Llama

Country	Mean Error	Std. Dev	Lag-1 ACF	LB p-val (24h)	LB p-val (168h)	ARCH-LM (p)	Skewness	Excess Kurtosis	Jarque-Bera (p)
DK_1	−9.33	53.14	0.906	0	0	0	−0.31	2.54	0
DE_LU	−15.50	56.02	0.896	0	0	0	−0.37	3.02	0
FR	−15.06	48.02	0.908	0	0	0	−0.40	0.83	0
PL	−25.92	54.07	0.880	0	0	0	−0.06	3.43	0
ES	−14.38	47.89	0.913	0	0	0	−0.29	0.65	0
NO_2	−11.97	31.97	0.899	0	0	0	−0.86	2.78	0

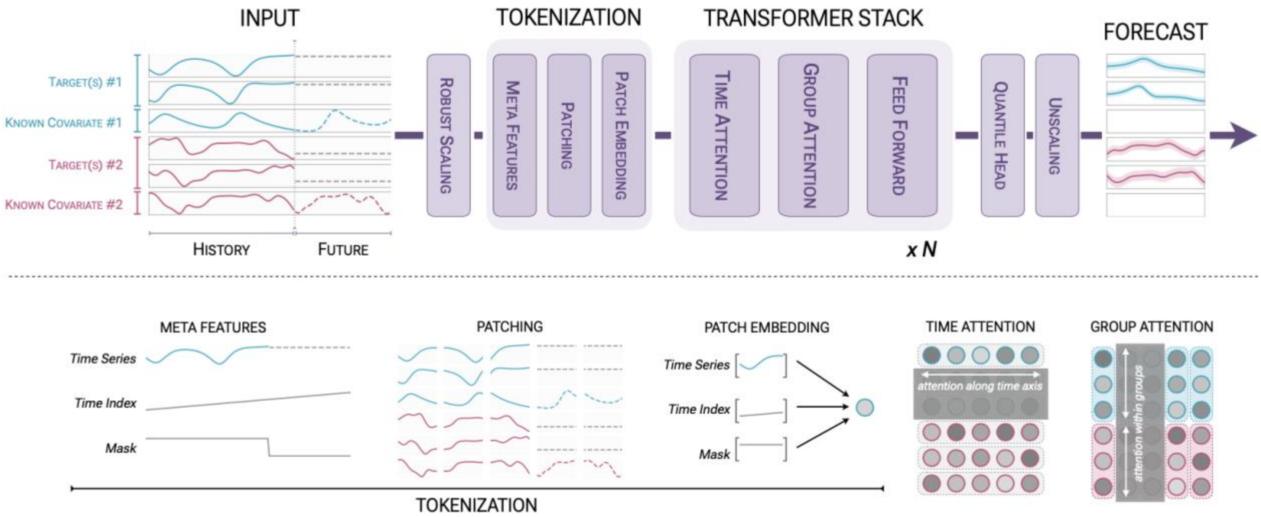
Evaluating the residual diagnostics of the Lag-llama model highlights the severe limitations of using a univariate framework, in a highly complex and covariate heavy scenario. The most prominent flaw shown in the diagnostics is directional bias. The model consistently manages to overpredict, additionally its has high standard deviation, this is a result of the low number of Montecarlo simulations, which was clearly insufficient to properly center the predicted distribution. Interestingly, when examining the distributional shape of the errors, the excess kurtosis is notably lower than of the Chronos V2 model. This suggest that the models design of a student's t-distribution help the excess kurtosis down. These diagnostics confirm that regardless of theoretical sophistication of lag-llama probabilistic output, its univariate nature leaves it structurally insufficient, to model the European Day-Ahead electricity market.

A7: Architecture of Time-GPT model



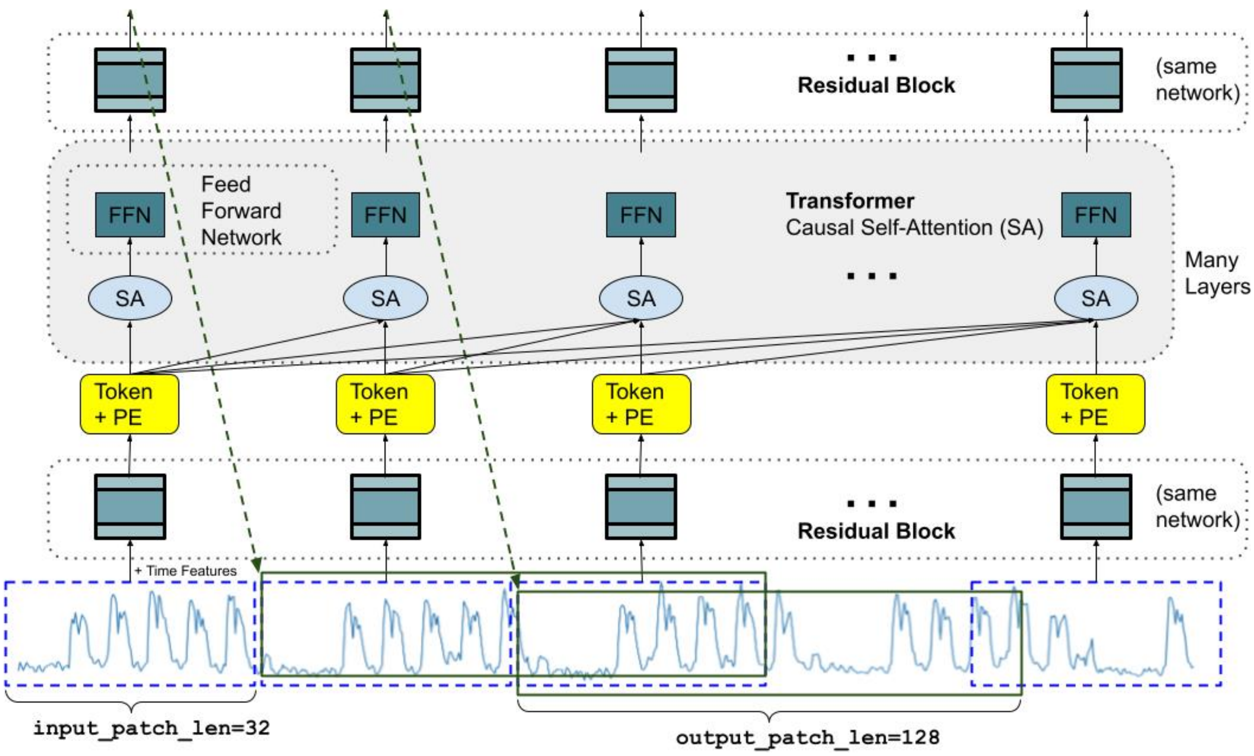
Source: (Peixeiro, 2026)

A8: Architecture of Chronos V2 model



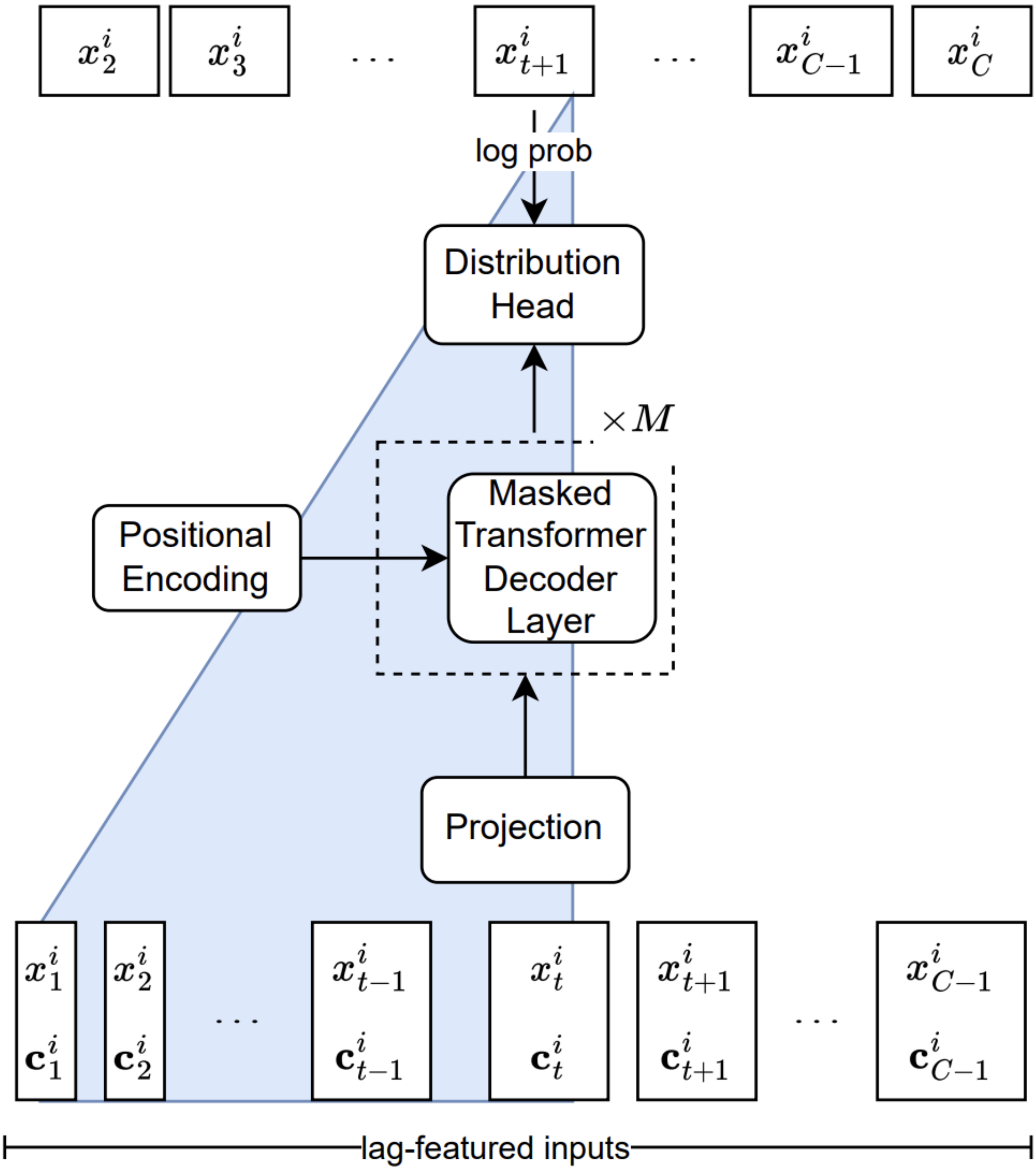
Source: (Ansari et al., 2025)

A9: Architecture of the TimesFM model



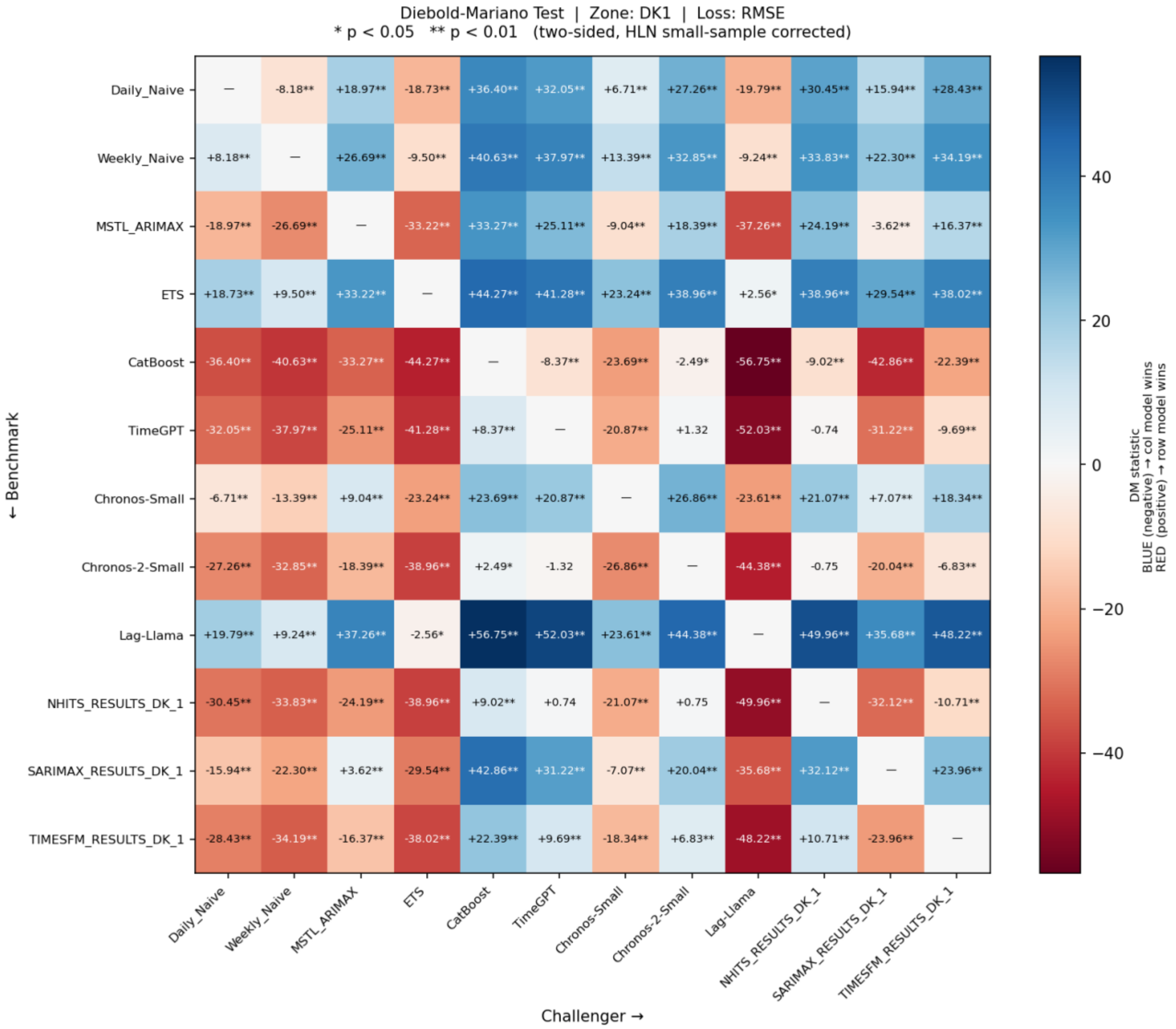
Source: (Google, 2024)

A10: Architecture of the Lag-llama model

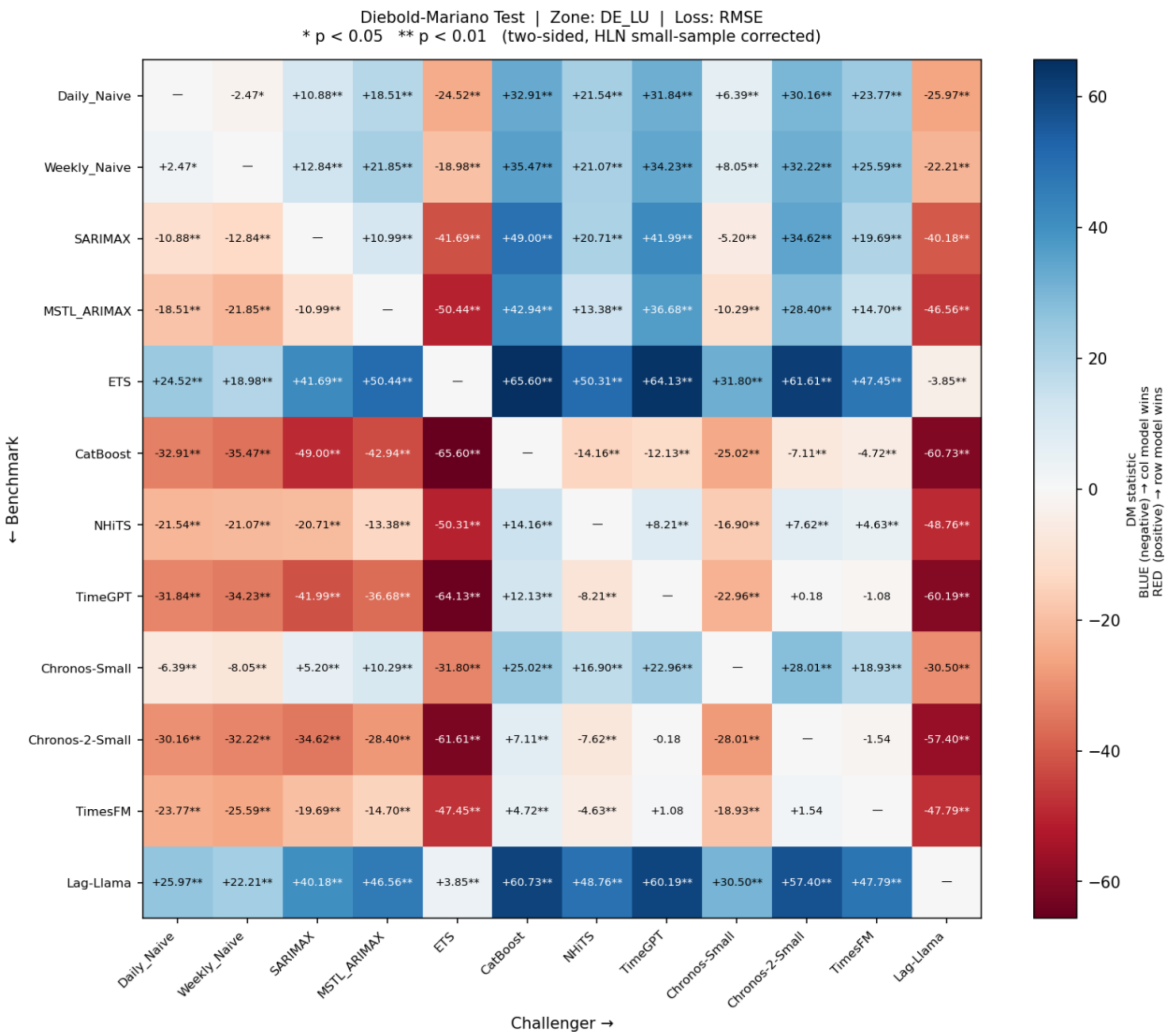


A11: Diebold-Mariano test

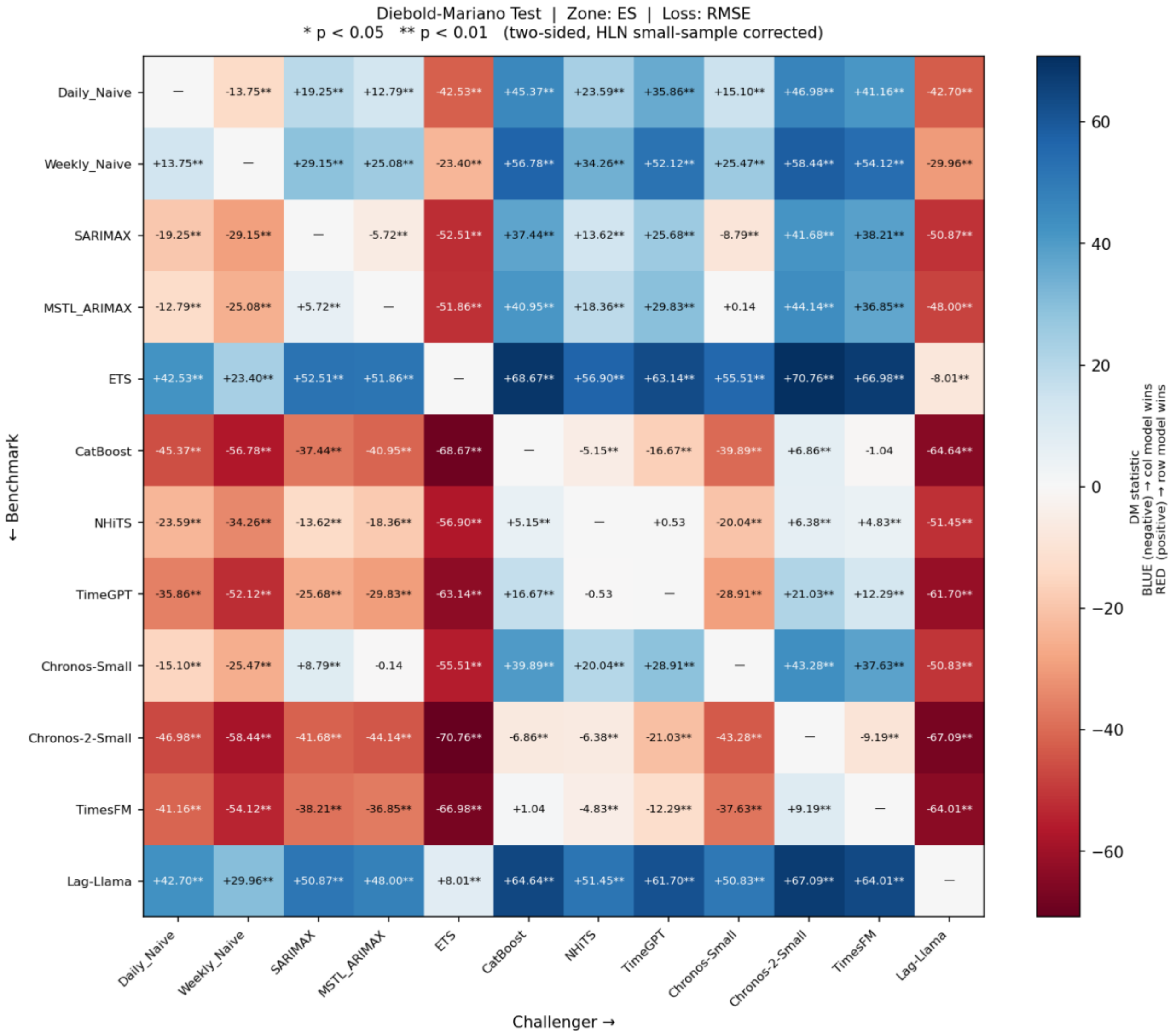
A11.1 Diebold-Mariano test - DK_1



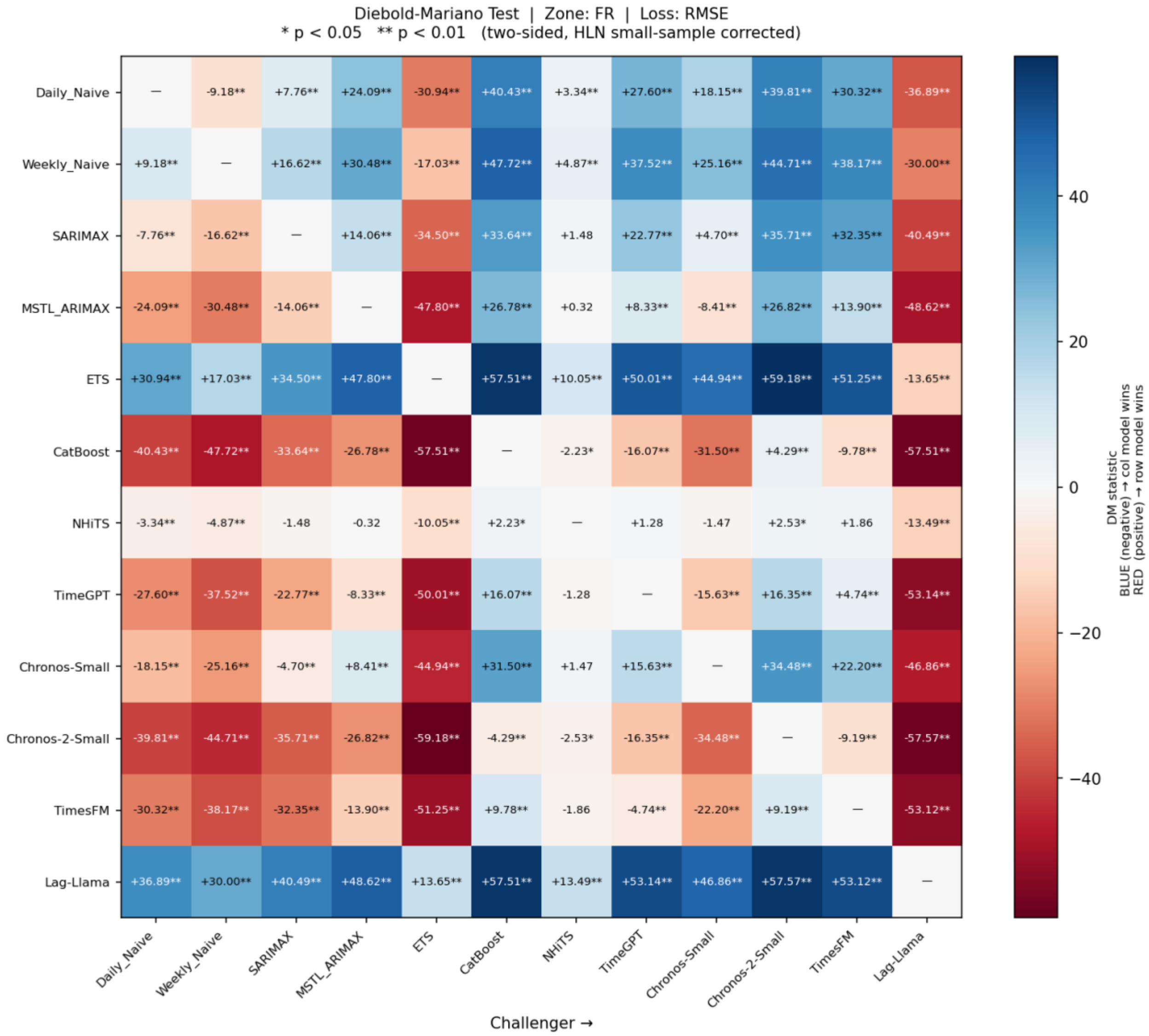
A11.2 Diebold-Mariano test - DE_LU



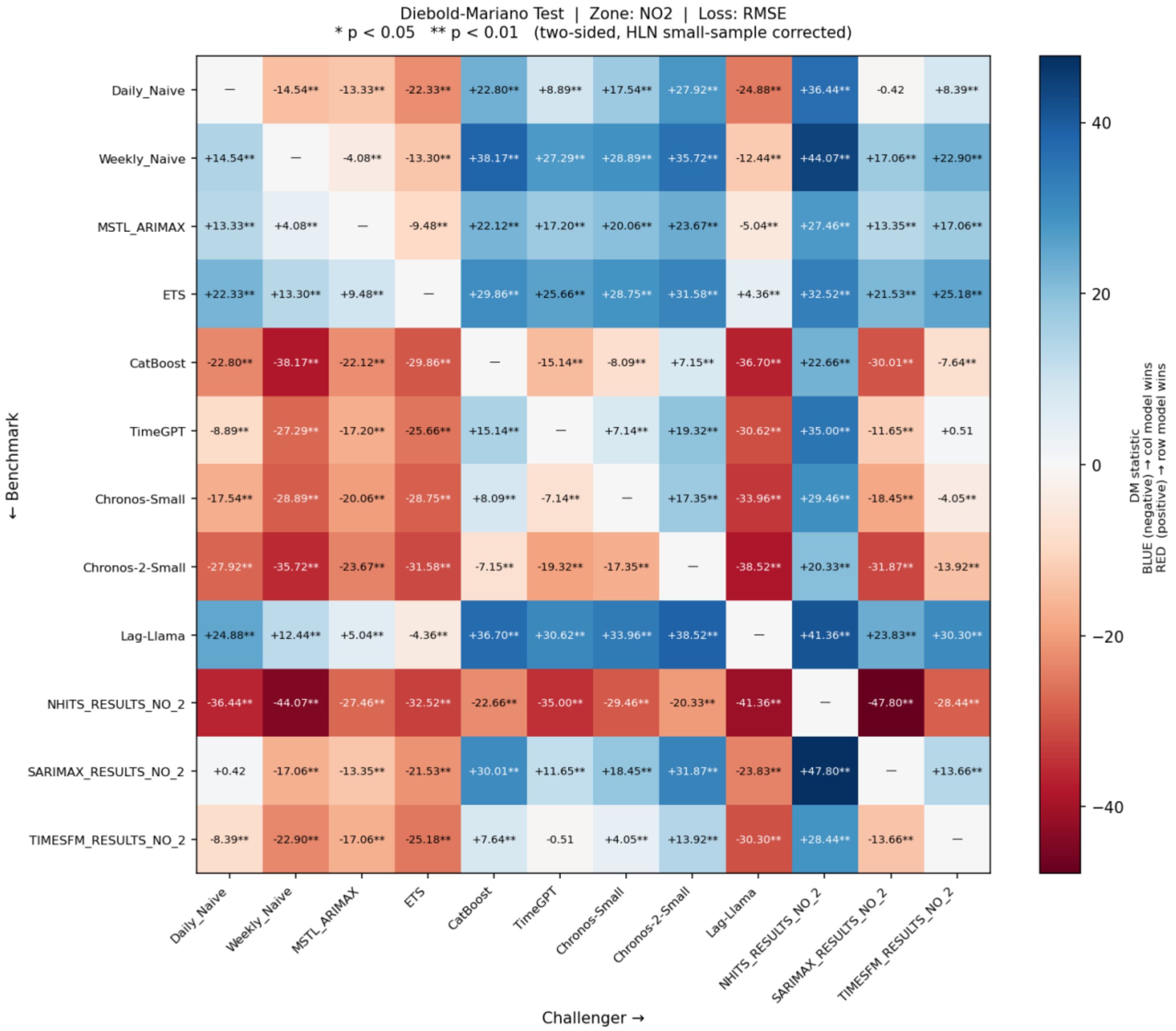
A11.3 Diebold-Mariano test - ES



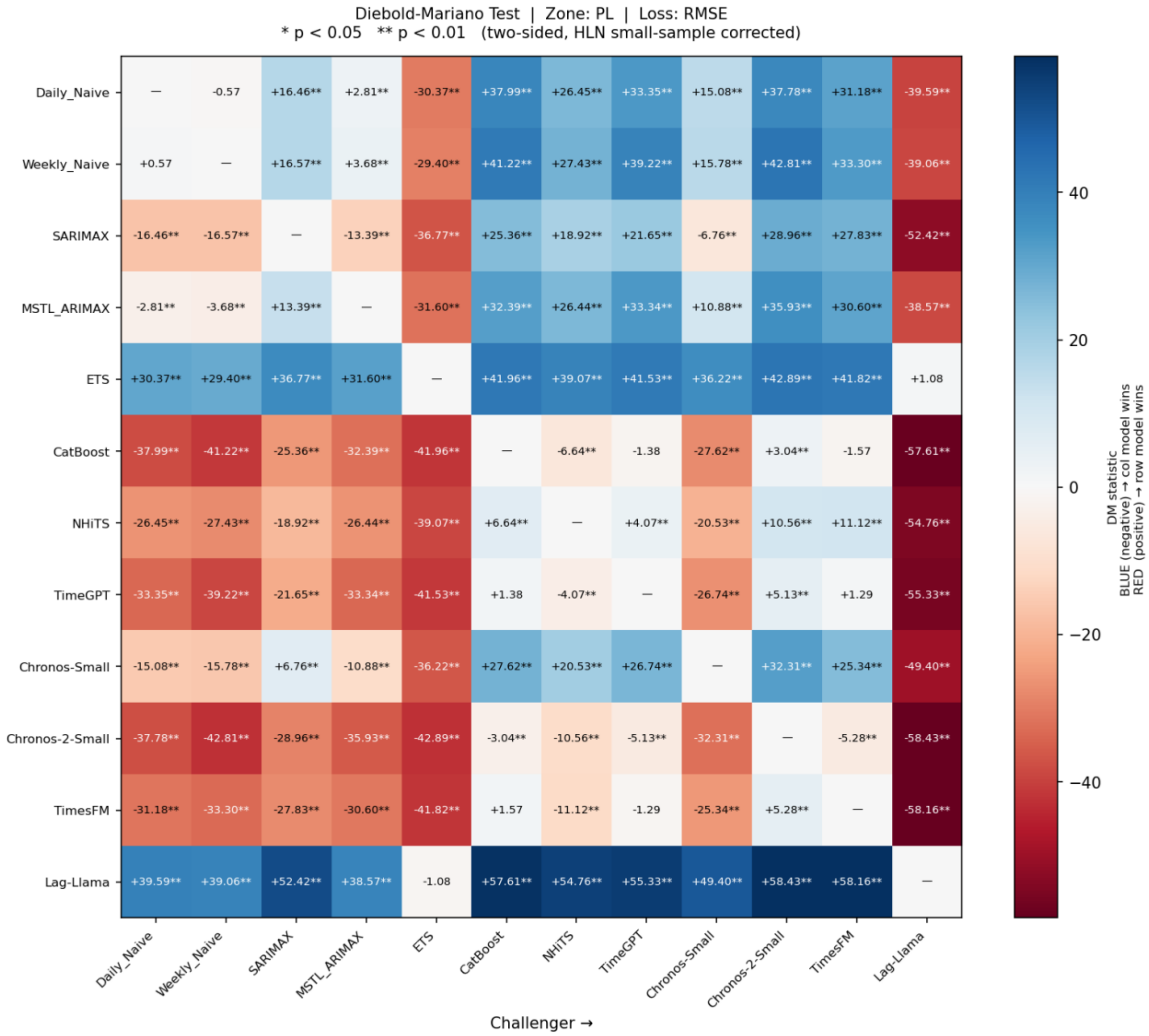
A11.4 Diebold-Mariano test - FR



A11.5 Diebold-Mariano test - NO_2



A11.6 Diebold-Mariano test - PL



References

ACER (2023) *Cross-zonal capacities and the 70% margin available for cross-zonal electricity trade (MACZT)*. Available at: https://arxiv.org/html/2501.02963v1 (Accessed: 2 April 2026).

ACER/CEER (2022) *Annual Report on the Results of Monitoring the Internal Electricity and Natural Gas Markets in 2021*. ACER/CEER. Available at: https://www.acer.europa.eu/sites/default/files/documents/Publications/MMR_2021_Energy_Retail_Consumer_Protection_Volume.pdf (Accessed: 6 April 2026).

'ACER-CEER 2024 Market Monitoring Report' (2024) *CEER*. Available at: https://www.ceer.eu/publication/acer-ceer-2024-market-monitoring-report-energy-retail-and-consumer-protection/ (Accessed: 29 April 2026).

Ansari, A.F. *et al.* (2024a) 'Chronos: Learning the Language of Time Series'. arXiv. Available at: https://doi.org/10.48550/arXiv.2403.07815.

Ansari, A.F. *et al.* (2024b) 'Chronos: Learning the Language of Time Series'. arXiv. Available at: https://doi.org/10.48550/arXiv.2403.07815.

Ansari, A.F. *et al.* (2025) 'Chronos-2: From Univariate to Universal Forecasting'. arXiv. Available at: https://doi.org/10.48550/arXiv.2510.15821.

Bandara, K., Hyndman, R.J. and Bergmeir, C. (2021) 'MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns'. arXiv. Available at: https://doi.org/10.48550/arXiv.2107.13462.

Bessembinder, H. and Lemmon, M.L. (2002) 'Equilibrium Pricing and Optimal Hedging in Electricity Forward Markets', *The Journal of Finance*, 57(3), pp. 1347–1382.

Böckers, V. and Heimeshoff, U. (2014) 'The extent of European power markets', *Energy Economics*, 46, pp. 102–111. Available at: https://doi.org/10.1016/j.eneco.2014.09.004.

Bommasani, R. *et al.* (2022) 'On the Opportunities and Risks of Foundation Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2108.07258.

Cevik, S. and Zhao, Y. (2025) 'Shocked: Electricity Price Volatility Spillovers in Europe', *IMF Working Papers*, 2025(007). Available at: https://doi.org/10.5089/9798400296901.001.A001.

Challu, C. *et al.* (2022a) 'N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting'. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.12886.

Challu, C. *et al.* (2022b) 'N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting'. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.12886.

Cludius, J. *et al.* (2014) 'The merit order effect of wind and photovoltaic electricity generation in Germany 2008–2016: Estimation and distributional implications', *Energy Economics*, 44, pp. 302–313. Available at: https://doi.org/10.1016/j.eneco.2014.04.020.

Contreras, J. *et al.* (2003) 'ARIMA models to predict next-day electricity prices', *IEEE Transactions on Power Systems*, 18(3), pp. 1014–1020. Available at: https://doi.org/10.1109/TPWRS.2002.804943.

Das, A. *et al.* (2024) 'A decoder-only foundation model for time-series forecasting'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.10688.

Duan, N. (1983) 'Smearing Estimate: A Nonparametric Retransformation Method', *Journal of the American Statistical Association*, 78(383), pp. 605–610. Available at: https://doi.org/10.2307/2288126.

EMBER (2024) *European Electricity Review 2024*. Available at: https://ember-energy.org/app/uploads/2024/10/European-Electricity-Review-2024.pdf (Accessed: 7 April 2026).

EPEXSPOT (2024) *Annual Trading Results of 2024*. Available at: https://www.epexspot.com/sites/default/files/download_center_files/2025-01-28_EPEX%20SPOT_Annual%20Power%20Trading%20Results%202024_finaldraft_0.pdf (Accessed: 7 April 2026).

European Commisson (2018) *Integration of the Iberian Peninsula into the internal energy market*, *European Commission - European Commission*. Available at: https://ec.europa.eu/commission/presscorner/detail/fr/memo_18_4622 (Accessed: 29 April 2026).

Fama, E.F. (1970) 'Efficient Capital Markets: A Review of Theory and Empirical Work', *The Journal of Finance*, 25(2), pp. 383–417. Available at: https://doi.org/10.2307/2325486.

Garza, A., Challu, C. and Mergenthaler-Canseco, M. (2024a) 'TimeGPT-1'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.03589.

Garza, A., Challu, C. and Mergenthaler-Canseco, M. (2024b) 'TimeGPT-1'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.03589.

Google (2024) *A decoder-only foundation model for time-series forecasting*. Available at: https://research.google/blog/a-decoder-only-foundation-model-for-time-series-forecasting/ (Accessed: 29 April 2026).

Greene, W. (2018) *Econometric analysis*. Eighth edition. New York, NY: Pearson.

International Energy Agency (2020) *Electricity Market Report*. OECD. Available at: https://doi.org/10.1787/f0aed4e6-en.

International Energy Agency (2021) *France 2021 Energy Policy Review*. OECD (IEA Energy Policy Reviews). Available at: https://doi.org/10.1787/2c889667-en.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) *An Introduction to Statistical Learning: with Applications in R. Springer.*

Karakatsani, N.V. and Bunn, D.W. (2008) 'Forecasting electricity prices: The impact of fundamentals and time-varying coefficients', *International Journal of Forecasting*, 24(4), pp. 764–785. Available at: https://doi.org/10.1016/j.ijforecast.2008.09.008.

Keppler, J.H., Phan, S. and Pen, Y.L. (2016) ‘The Impacts of Variable Renewable Production and Market Coupling on the Convergence of French and German Electricity Prices’, *The Energy Journal*, 37(3), pp. 343–360. Available at: https://doi.org/10.5547/01956574.37.3.jkep.

Ketterer, J.C. (2014) ‘The impact of wind power generation on the electricity price in Germany’, *Energy Economics*, 44, pp. 270–280. Available at: https://doi.org/10.1016/j.eneco.2014.04.003.

Knittel, C.R. and Roberts, M.R. (2005) ‘An empirical examination of restructured electricity prices’, *Energy Economics*, 27(5), pp. 791–817. Available at: https://doi.org/10.1016/j.eneco.2004.11.005.

Kwiatkowski, D. *et al.* (1992) ‘Testing the null hypothesis of stationarity against the alternative of a unit root’, *Journal of Econometrics*, 54(1–3), pp. 159–178. Available at: https://doi.org/10.1016/0304-4076(92)90104-Y.

Lago, J. *et al.* (2021) ‘Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark’, *Applied Energy*, 293, p. 116983. Available at: https://doi.org/10.1016/j.apenergy.2021.116983.

Lago, J., De Ridder, F. and De Schutter, B. (2018) ‘Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms’, *Applied Energy*, 221, pp. 386–405. Available at: https://doi.org/10.1016/j.apenergy.2018.02.069.

Lucia, J.J. and Schwartz, E.S. (2002) ‘Electricity Prices and Power Derivatives: Evidence from the Nordic Power Exchange’, *Review of Derivatives Research*, 5(1), pp. 5–50. Available at: https://doi.org/10.1023/A:1013846631785.

Mashlakov, A. *et al.* (2021) ‘Assessing the performance of deep learning models for multivariate probabilistic energy forecasting’, *Applied Energy*, 285, p. 116405. Available at: https://doi.org/10.1016/j.apenergy.2020.116405.

NEMO, C. (2020) *euphemia-public-description-2020*, *EUPHEMIA Public Description*. Available at: https://www.nemo-committee.eu/assets/files/euphemia-public-description-2020.pdf (Accessed: 28 March 2026).

Nogales, F.J. *et al.* (2002) ‘Forecasting next-day electricity prices by time series models’, *IEEE Transactions on Power Systems*, 17(2), pp. 342–348. Available at: https://doi.org/10.1109/TPWRS.2002.1007902.

NordPool (2024a) *Day-ahead market*. Available at: https://www.nordpoolgroup.com/en/the-power-market/Day-ahead-market/ (Accessed: 29 April 2026).

NordPool (2024b) *Successful go-live of the Nordic Flow-based Market Coupling project*. Available at: https://www.nordpoolgroup.com/en/message-center-container/newsroom/exchange-message-list/2024/q4/successful-go-live-of-the-nordic-flow-based-market-coupling-project/ (Accessed: 7 April 2026).

Norway Insights (2025) *EMBER*. Available at: https://ember-energy.org/countries-and-regions/norway (Accessed: 7 April 2026).

Peixeiro, M. (2026) *Time Series Forecasting Using Foundation Models*. Shelter Island: Manning Publications.

Prokhorenkova, L. *et al.* (2018) 'CatBoost: unbiased boosting with categorical features', *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html (Accessed: 2 April 2026).

Rob Hyndman, G. Athanasopoulos (2021) *Forecasting: Principles and Practice (3rd ed)*. Available at: https://otexts.com/fpp3/ (Accessed: 29 April 2026).

Seitaridis, M.I., Thomaidis, N.S. and Biskas, P.N. (2021) 'Fundamental Responsiveness in European Electricity Prices', *Energies*, 14(22), p. 7623. Available at: https://doi.org/10.3390/en14227623.

Sensfuß, F., Ragwitz, M. and Genoese, M. (2008) 'The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany', *Energy Policy*, 36(8), pp. 3086–3094. Available at: https://doi.org/10.1016/j.enpol.2008.03.035.

Shojaie, A. and Fox, E.B. (2022) 'Granger Causality: A Review and Recent Advances', *Annual Review of Statistics and Its Application*, 9(Volume 9, 2022), pp. 289–319. Available at: https://doi.org/10.1146/annurev-statistics-040120-010930.

SMARD | The electricity market in 2023 (2023). Available at: https://www.smard.de/page/en/topic-article/209972/211754/the-electricity-market-in-2023 (Accessed: 29 April 2026).

Verbeek, M. (2017) *A guide to modern econometrics*. Fifth edition. Hoboken, NJ: Wiley.

Weron, R. (2014a) 'Electricity price forecasting: A review of the state-of-the-art with a look into the future', *International Journal of Forecasting*, 30(4), pp. 1030–1081. Available at: https://doi.org/10.1016/j.ijforecast.2014.08.008.

Weron, R. (2014b) 'Electricity price forecasting: A review of the state-of-the-art with a look into the future', *International Journal of Forecasting*, 30(4), pp. 1030–1081. Available at: https://doi.org/10.1016/j.ijforecast.2014.08.008.

Wooldridge, J. M. (20215) *Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning*. 2015th edn.

Yıldız, C. (2023) 'A TWO STAGE MODEL FOR DAY-AHEAD ELECTRICITY PRICE FORECASTING: INTEGRATING EMPIRICAL MODE DECOMPOSITION AND CATBOOST ALGORITHM', *Konya Journal of Engineering Sciences*, 11(4), pp. 1047–1060. Available at: https://doi.org/10.36306/konjes.1290652.

Zachmann, G. (2008) 'Electricity wholesale market prices in Europe: Convergence?', *Energy Economics*, 30(4), pp. 1659–1671. Available at: https://doi.org/10.1016/j.eneco.2007.07.002.

Zachmann, G. *et al.* (2023) 'The design of the European electricity market - Current proposals and ways ahead'.

Zareipour, H. *et al.* (2006) ‘Application of Public-Domain Market Information to Forecast Ontario’s Wholesale Electricity Prices’, *IEEE Transactions on Power Systems*, 21(4), pp. 1707–1717. Available at: https://doi.org/10.1109/TPWRS.2006.883688.

Zhang, F., Fleyeh, H. and Bales, C. (2022) ‘A hybrid model based on bidirectional long short-term memory neural network and Catboost for short-term electricity spot price forecasting’, *Journal of the Operational Research Society*, 73(2), pp. 301–325.

Ziel, F. and Weron, R. (2018) ‘Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks’, *Energy Economics*, 70, pp. 396–420. Available at: https://doi.org/10.1016/j.eneco.2017.12.016.