IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

Target Representation as a Key Factor for Generalization of Physics-Based Electricity Price Correction Models

Istifa Shania Putri*[1], Tien Fabrianti Kusumasari[2], Sinung Suakanto[3]
[1,2,3]Information Systems, Faculty of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia
e-mail: *istifashaniaputri@telkomuniversity.ac.id, tienkusumasari@telkomuniversity.ac.id, sinung@telkomuniversity.ac.id

Abstrak

Model komputer yang digunakan untuk memprediksi harga listrik, seperti Linear Optimal Power Flow (LOPF), memang akurat secara teori fisik tetapi sering kali meleset dari harga asli di pasar. Untuk mengatasi masalah ini, para ahli biasanya menambahkan lapisan koreksi berbasis kecerdasan buatan atau pembelajaran mesin. Namun, model koreksi ini sering kehilangan akurasi saat kondisi pasar berubah drastis karena titik batas kemampuan bertahan model serta penyebab utama pergeserannya belum diketahui. Melalui data pasar listrik Jerman tahun 2019 dan 2024 yang mencakup 17.544 jam pengamatan, penelitian ini memetakan batas ketahanan model tersebut melalui tiga temuan utama. Pertama, mengganti algoritma atau mengoptimalkan setelan internal (hyperparameter) belum menunjukkan perbaikan yang berarti; seluruh 16 kombinasi model yang diuji gagal ketika model yang dilatih pada tahun dengan harga listrik mahal dipaksa menebak harga pada tahun yang murah, dimana tingkat kesalahannya justru membengkak 2,2 hingga 4,0 kali lipat. Kedua, sumber kegagalan terletak pada cara target pembelajaran dirumuskan; pada konfigurasi model terbaik, menggantikan target berbasis selisih harga (EUR/MWh) dengan target berbasis rasio mengubah kedua skema pengujian menjadi positif (forward dan backward), termasuk perbaikan 34,1% pada backward. Ketiga, faktor penentu keberhasilan adalah jarak antar-kondisi pasar biaya bahan bakar, berkaitan erat dengan jarak harga bahan bakar; pada 30 pasang perpindahan, metode konvensional memburuk seiring bertambahnya jarak kondisi pasar (rezim). Penelitian ini menguji ketahanan model pada lonjakan pasar ekstrem tahun 2022 (puncak krisis gas Eropa, RMSE: 159,06 EUR/MWh). Hasilnya, metode target rasio lebih akurat (+8% sd +11%) daripada target selisih saat memprediksi masa krisis. Namun, model ini gagal jika dibalik (dilatih saat krisis tetapi diuji pada kondisi normal), menunjukkan adanya batas kemampuan model.

Kata kunci—Model hibrida; transferabilitas; volatilitas; ekstrapolasi; peramalan residu

Abstract

Physics-based electricity price forecasting models, such as Linear Optimal Power Flow (LOPF), provide strong theoretical accuracy but often fail to capture real market prices. To reduce this gap, machine learning-based correction models are commonly applied; however, their reliability under significant market regime shifts remains unclear. This study investigates the robustness boundaries of price correction models using German electricity market data from 2019 and 2024, consisting of 17,544 hourly observations. Three key findings are identified. First, model selection and hyperparameter optimization do not improve cross-regime robustness: all 16 tested configurations fail when models trained in high-price regimes are applied to low-price regimes, increasing forecasting errors by 2.2–4.0 times. Second, the main source of failure is the target formulation. For the best-performing configuration, replacing the conventional price-difference target (EUR/MWh) with a ratio-based target turns all four transfer scenarios positive, achieving up to a 34.1% improvement in backward transfer. Third, model degradation is

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

primarily driven by the distance between fuel-cost market regimes rather than temporal distance. Evaluation across 30 regime transitions shows that conventional correction approaches deteriorate as market condition differences increase. These findings reveal the limitations of current AI-based electricity price correction frameworks and highlight target redesign and regime-aware modeling as critical strategies for improving robustness under evolving market conditions.This study tests model resilience during the 2022 extreme market volatility (peak of the European gas crisis, RMSE: 159.06 EUR/MWh). The results show that the ratio target outperforms the difference target by +8% to +11% when predicting crisis periods. However, the model fails when trained on crisis data but tested under normal conditions, showing an explicit regime limit.

Keywords—Hybrid models; transferability; volatility; extrapolation; residual forecasting

1. INTRODUCTION

Electricity prices in the day-ahead market follow the merit order principle. Plants are ranked starting from the cheapest fuel cost. They are turned on sequentially until the total demand is met. As a result, the operational cost of the last active plant sets the market price [1]. Due to this mechanism, prices can be simulated computationally. This is often done using open-source PyPSA software with the Linear Optimal Power Flow (LOPF) method [2]. This method essentially works as an optimizer. It schedules plant operations and outputs hourly electricity price estimates as a simulated market reference.

It is worth stating plainly which markets this mechanism describes. Uniform marginal pricing of the kind simulated here is the clearing rule of the European internal electricity market, where day-ahead auctions across the participating countries are matched by a single algorithm, PCR EUPHEMIA, under the Single Day-Ahead Coupling arrangement, yielding one clearing price per bidding zone for every trading interval. Germany, the case examined in this paper, sits inside that arrangement

The issue is that real market prices depend on highly complex factors. It goes far beyond simple fuel costs. These include the cost of starting machines from a cold state. Some plants must also run continuously to supply district heating. Transmission network capacity limits and cross-border electricity trading also play major roles. Even the bidding strategies of market participants influence the final outcome. Because LOPF ignores most of these variables, its simulated prices deviate from actual prices following a fairly regular pattern [3].

Machine learning models can easily learn these regular patterns. This gives rise to the idea of a hybrid system. The main physics model continues to run as usual. Then, a machine learning model is trained to predict and correct its deviations. This physics-machine learning hybrid approach has proven useful in power flow analysis [4]. It helps bridge simplified DC models toward more realistic AC conditions. Adding a correction layer on top of a main model is also quite common in statistical forecasting [5].

However, there is a fundamental question that is rarely explored. How long does that correction formula actually remain valid? This matters because power systems are highly dynamic environments. Between 2019 and 2024 in Germany, for instance, natural gas prices more than doubled. Carbon emission rates increased by two and a half times. The share of renewable energy also grew by roughly eight percent [6],[7]. A statistical correction learned from one historical condition might lose relevance when the market changes. Most previous studies only tested accuracy on data ranges similar to their training sets. Therefore, the long-term reliability of these correction models remains unanswered.

On the other hand, using tree ensembles like XGBoost to forecast electricity prices is mature. These models consistently outperform conventional linear methods [8]. This also includes

their combinations with neural networks [9]. Furthermore, explaining model behavior through feature importance is now an established reporting standard [10].

Nevertheless, most of those studies tested their models under relatively stable market conditions [11],[12],[13]. When data conditions change significantly, the main problem shifts away from mere accuracy. The focus moves to generalization, meaning how well a model performs outside its training data. The challenge of concept drift in energy forecasting is widely recognized in existing literature [14]. The limits of model transferability across electricity markets have also been well documented empirically [15].

This research takes a completely different perspective. Most past studies view reduced generalization as a data or architecture limitation. Their proposed solutions often involve deploying larger capacity models [16]. Other common approaches include stricter parameter optimization [17] or self-adapting mechanisms [18]. We demonstrate the exact opposite in this study. For residual correction cases, the root of generalization failure stems from a rarely questioned fundamental decision. It is the specific choice of the variable used as the learning target.

Based on this background, the current research formulates three distinct questions. First, can we fix a correction model's failure during market shifts just by changing algorithms or tuning hyperparameters? Second, if that fails, which design component actually serves as the main determinant? Third, what factors limit model transferability success, and how much price improvement is still reasonable to expect?

To answer these, this study conducts experiments that reveal four main findings. We show empirically that hyperparameter tuning does not restore performance when market conditions shift. We identify the target variable formulation as the primary cause of failure and propose a better design. We introduce a fuel cost-based market condition definition for robust evaluation. This allows testing across 30 regime shift scenarios instead of comparing two isolated years. Finally, we measure the correction layer baseline error as a threshold. This threshold clearly explains when the correction actually becomes beneficial.

2. METHODS

This section outlines the comprehensive research design structured across eight main stages. The complete sequence of these methodological steps is visually presented in Figure 1. The analytical process initially begins by gathering the German energy dataset for 2019 and 2024. Following this, the underlying physics model is configured using the PyPSA LOPF framework. This setup phase also includes establishing scenario assumptions and defining the target parameterization. The subsequent stage focuses heavily on data integration and pre-processing the main experimental features. Model calibration is then carefully executed by testing 16 distinct sets of hyperparameter combinations. The entire research workflow concludes with a rigorous evaluation based on cross-validation techniques. Finally, a verification process analyzes model robustness and generalization power against actual empirical market findings.

2.1 Data

This study relies on publicly available German electricity market data from 2019 and 2024. The dataset encompasses a total of 17,544 hourly observations spanning across both periods. Table 1 summarizes the core content and inherent characteristics of this compiled data. These two specific years were purposely selected due to their sharply contrasting market conditions. The year 2019 represents a baseline market characterized by cheap fuel and stable prices. Conversely, 2024 reflects a highly volatile market with expensive fuel and larger renewable energy shares. This significant contrast provides the fundamental basis needed to effectively evaluate model generalization capabilities. The initial data cleaning process involved

standardizing all temporal records to the UTC time zone. Duplicate hours occurring during daylight saving transitions were carefully removed to maintain chronological integrity. Finally, any brief gaps in the dataset were seamlessly filled using linear interpolation methods.

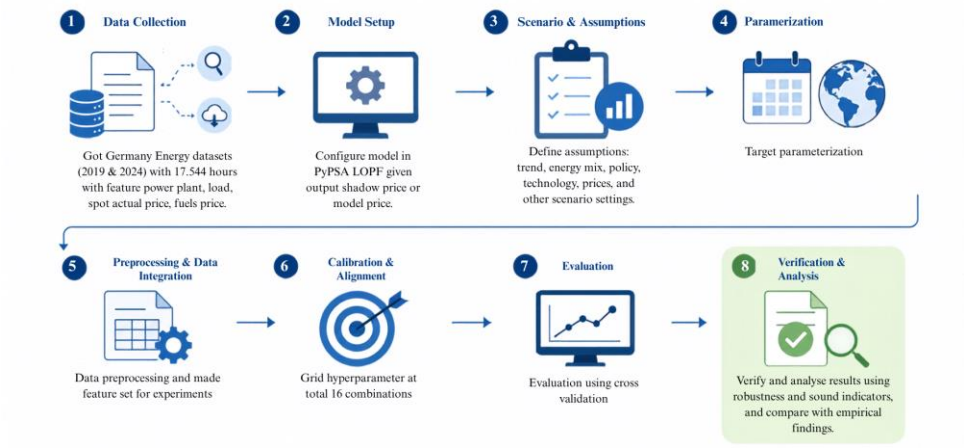


Figure 1 Methodological steps

We compare 2019 and 2024 instead of a continuous timeline because they represent the widest gap in recent German energy conditions. This maximum-contrast pair tests how well the correction layer handles big changes in gas costs, carbon prices, and renewable shares. Section 3.4 further proves this by testing across 30 different fuel cost regimes, making the two-year framing just a starting point. Section 3.5 tests our model against genuine extreme volatility using 2022 data from the peak European gas crisis. We harmonized CO2 prices to EEX annual auction averages, while 2022 natural gas and coal prices come from World Bank monthly averages.

Table 1 Summary of research data and comparison of conditions for both years

Description	2019	2024
Number of observation hours	8,760	8,784
Data source for generation, load, spot price	Energy-Charts	Energy-Charts
Gas, coal, carbon price (monthly)	Commodity market	Commodity market
Average spot price (EUR/MWh)	37.7	78,5
Average gas price (EUR/MWh fuel)	14.6	34.6
Wind and solar share	33.2%	41.8%
Hours with negative prices	211	457
Hours with prices above 150 EUR/MWh	0	361

2. 2 Physical Model

In this research, the underlying electrical grid is primarily modeled as a single-bus system. This approach treats the entire German territory as one massive unified convergence point. All electricity supply and demand meet here without considering geographical or regional transmission limits. Power plants within this specific node are systematically grouped based on their technology types. Each technology cluster is then assigned a single reference fuel cost per Megawatt-hour (MWh). This mathematical model is subsequently solved using the LOPF approach independently for every hour. The entire computational simulation process is actively executed by utilizing the open-source PyPSA framework. Next, two distinct simulation scenarios are implemented to thoroughly evaluate the model's operational performance. Scenario A0 operates on the assumption that fuel cost values remain constant all year.

2. 3 How the correction layer works

The correction layer is essentially a standard regression model in its most basic form. It is initially trained using data from a specific timeframe called the source period. The trained model is subsequently tested on a different timeframe called the target period. Throughout this process, the model learns to predict price differences measured in EUR/MWh. This predicted difference is then directly added to the physics model's generated price [5]. We define this conventional method as the difference target in our current research framework. We actively utilize this specific method as the primary comparative baseline for our evaluation. The exact mathematical formulation for this approach is clearly demonstrated within Equation 1. Here, the variable *e(t)* represents the calculated residual correction value at a specific hour. The actual price variable accurately denotes the real market price that occurred at that time. Meanwhile, the model price represents the simulated price output generated by the LOPF physics model.

$$e(t) = actual\ price(t) - shadow\ price(t) \quad (1)$$

This predicted difference is directly added to the physics model to obtain final predictions. However, implementing this regression model faces significant challenges when using decision tree ensemble algorithms. These specific algorithms generate predictions by averaging target values learned during the training process. This characteristic creates an inherent weakness regarding the model's overall predictive output boundaries. The output values can never exceed the maximum or minimum bounds of the training data. This fundamental limitation clearly demonstrates a strict inability to perform data extrapolation.

As a concrete example of this limitation, consider a model trained on 2019 data. During this specific timeframe, the price error margin only hovered around 13 EUR/MWh. Suppose the same model is applied to 2024 market conditions needing larger adjustments. If the new market requires a 44 EUR/MWh correction, the model cannot predict it. This problem will persist regardless of the duration of the model's training process. It also remains true regardless of how carefully the hyperparameter tuning is conducted. This fundamental machine learning limitation is often recognized as a major operational obstacle. Various power flow analysis studies have documented this exact constraint as a primary challenge.

2. 4 Target parameterization

Target parameterization simply defines what the algorithm actually needs to predict during its training phase. Implementing this specific strategy ultimately makes the overall machine learning process much more adaptive. This research actively explores two distinct target formulations as its primary core experimental variables. The approach relies upon residual forecasting and variance stabilization concepts to dampen extreme price volatility. The first formulation naturally serves as our conventional evaluation baseline for this particular study. It calculates the absolute difference target measured strictly in standard EUR/MWh unit increments. This exact value comes directly from subtracting the physics model output from actual market prices.

Meanwhile, the second formulation completely shifts the analytical focus toward a relative numerical scale. It utilizes a ratio target approach rather than relying heavily on absolute price values. Under this new setup, the model learns the exact division between actual and simulated prices. Comparing these two distinct forms provides valuable relevance for our overall methodological evaluation. It specifically tracks the critical transition from absolute measurement units into a completely dimensionless metric. Adopting this relative scale detaches the target value from original price magnitudes almost entirely. For instance, a 1.2 ratio will always maintain a constant and stable proportional meaning. This fact remains perfectly true whether average market prices sit at 37 or 78 EUR/MWh. This specific characteristic deliberately ensures that learned machine learning patterns remain valid and highly robust. It significantly helps the model preserve its core generalization power across different market environments. The model stays firmly resilient even if market price levels shift drastically during testing periods. Table 2 then shows

ISSN (print): 1978-1520, ISSN (online): 2460-7258

five rows of the assembled dataset so that the shape of the learning problem is concrete. The final two columns of that table are the two competing learning targets. s(x) = max(x, f) jika x ≥ 0, dan s(x) = min(x, −f) jika x < 0

$$r(t) = clip\,(actual\ price(t)/s(model\ price\ (t)), -c, c),$$
$$s(x) = max(x, f)\ if\ x \geq 0, and\ s(x) = min(x, -f)\ if\ x < 0 \quad (2)$$

To stop division by zero or massive spikes, we use two safety steps in Equation 2. First, if the physics model price gets too close to zero, we replace it with a small floor value (f = 1 EUR/MWh) while keeping its positive or negative sign [19]. The function *s(x)* introduces a denominator small floor value that prevents division-by-zero errors when the physics model outputs near-zero values, while still keeping the original price sign Second, we cap the final ratio between -5 and 5 so that a few extreme hours do not ruin the model's training. Our data checks prove these rules are rarely needed but protect the model when necessary. The floor rule triggered in only 0.02% of hours for 2019/2024 and 0% for 2022. The cap rule active rate was 0.14% for 2019/2024 and rose to 2.5% during the wild 2022 crisis. This shows the safeguards stabilize the math without changing the overall results.

Table 2 Five rows of the assembled dataset, Germany, 3 January 2024, scenario A0

Timestamp (UTC)	Load (MW)	Residual load (MW)	VRE share	LOPF price (EUR/MWh)	Actual price (EUR/MWh)	Difference target (EUR/MWh)	Ratio target
2024-01-03 00:00	44,335	4,822	0.891	23.0	−1.2	−24.2	−0.05
2024-01-03 06:00	58,239	10,810	0.814	23.0	40.6	+17.6	1.76
2024-01-03 10:00	64,448	11,653	0.819	23.0	54.9	+31.9	2.39
2024-01-03 14:00	61,490	17,890	0.709	68.5	67.2	−1.3	0.98
2024-01-03 17:00	64,586	27,867	0.569	68.5	83.0	+14.5	1.21

2.5 Feature set

This study evaluates four feature sets for electricity price correction. FS1 uses absolute raw inputs, including actual load and standard time variables. FS2 converts these inputs into dimensionless representations, such as percentiles, to improve robustness against extreme price variations. FS3 incorporates historical trends from the previous 24 hours and fuel price information; however, it introduces data leakage because statistical features are computed using the full-year dataset, allowing future information to influence past predictions. For example, January forecasts may implicitly include statistical patterns from December, making FS3 unsuitable for real-time forecasting. To address this limitation, FS4 is designed based on strict causal constraints, where only past information is accessible to the model. Full-year statistics are replaced with a 30-day rolling backward average, and constant fuel price features are removed due to their limited contribution to model learning. Table 3 summarizes all predictors provided to the correction layer for each feature set. Actual market prices are excluded from the input matrix, while LOPF-generated prices remain valid inputs because they are available before market clearing.

2.6 Design of Experiment

The experimental design evaluates four factors: target parameterization, feature set, algorithm, and evaluation scheme. Two target formulations are compared: price difference and ratio, while four feature configurations are examined: FS1 (original), FS2 (dimensionless), FS3 (history and fuel features), and FS4 (causal features). Four machine learning models are evaluated: Random Forest, XGBoost, Neural Network, and Kernel Ridge, using five testing schemes: CV 2019, CV 2024, forward, backward, and chronological evaluation. The CV schemes apply 5-fold cross-validation within each year as homogeneous baselines, while forward and backward evaluate cross-regime transfer between 2019 and 2024. The chronological scheme trains on January–September 2024 and tests on October–December 2024 to represent real-world forecasting conditions. Model performance is measured using RMSE and reported as percentage improvement over the baseline physics model, where positive values indicate improved predictions and negative values indicate performance degradation.

Table 3 Predictor variables of the correction layer

Variable	Definition	FS 1	FS2	FS3	FS 4
residual_load_mw	Load minus wind and solar generation	✓			
load_pct / c_load_norm	System load, percentile or trailing z-score		✓	✓	✓
rl_pct, rl_norm	Residual load, percentile and z-score		✓	✓	✓
c_rl_norm, c_rl_pos	Residual load, trailing-window z-score and position				✓
mp_pct, mp_rel	LOPF model price, percentile and ratio to mean		✓	✓	
c_mp_rel, c_mp_pos	LOPF model price, trailing ratio and position				✓
vre_share	Wind and solar share of total generation	✓	✓	✓	✓
month, hour_of_day	Calendar month and hour index	✓			
hr_sin, hr_cos, mo_sin, mo_cos	Cyclical encoding of hour and month		✓	✓	✓
weekday	Day of week	✓	✓	✓	✓
season_winter … season_autumn	Season, one-hot encoded	✓	✓	✓	✓
rl lag 1, lag 24; mp lag 1, lag 24	Lagged residual load and model price			✓	✓
rl roll24, mp roll24	24-hour trailing means			✓	✓
gas_rel, coal_rel	Monthly fuel cost relative to annual mean			✓	
c_gas_rel, c_coal_rel	Monthly fuel cost relative to previous month				✓

Hyperparameter configurations (Table 4) are empirically defined based on dataset characteristics, with C1–C4 representing increasing model capacity. Random Forest and XGBoost configurations vary maximum depth and estimator numbers, while Neural Network settings adjust hidden layers and alpha regularization, enabling systematic evaluation of model robustness under market regime shifts rather than relying on incidental tuning outcomes.

Table 4 Hyperparameter values of the sixteen testing algorithm configurations

Algorithm	Code	Hyperparameter set
Random Forest	RF_C1	n_estimators=300, min_samples_leaf=5, max_features=1,0
	RF_C2	n_estimators=200, min_samples_leaf=20, max_depth=8, max_features=0,5
	RF_C3	n_estimators=700, min_samples_leaf=1, max_features=1.0
	RF_C4	n_estimators=500, min_samples_leaf=2, max_features=sqrt
XGBoost [12]	XGB_C1	n_estimators=400, max_depth=6, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8
	XGB_C2	n_estimators=200, max_depth=3, learning_rate=0.1, subsample=0.8, reg_lambda=5,0
	XGB_C3	n_estimators=1000, max_depth=9, learning_rate=0.01, subsample=0.8, colsample_bytree=0.8
	XGB_C4	n_estimators=600, max_depth=8, learning_rate=0.03, min_child_weight=5, reg_alpha=0.1
Neural Network	MLP_C1	hidden_layer_sizes=(64, 32), alpha=1e-4, learning_rate_init=1e-3, max_iter=400
	MLP_C2	hidden_layer_sizes=(32), alpha=1e-2, learning_rate_init=1e-3, max_iter=400
	MLP_C3	hidden_layer_sizes=(128, 64, 32), alpha=1e-5, learning_rate_init=1e-3, max_iter=600
	MLP_C4	hidden_layer_sizes=(100), alpha=1e-4, learning_rate_init=1e-3, max_iter=500
Kernel Ridge	KR_C1	n_components=300, gamma=0,1, alpha=1
	KR_C2	n_components=150, gamma=0,05, alpha=10
	KR_C3	n_components=600, gamma=0,5, alpha=0.1
	KR_C4	n_components=300, gamma=0,2, alpha=0.5

3. RESULTS AND DISCUSSION

3.1 Model price results

Before evaluating the performance of the correction layer, the output price characteristics of the physics model need to be analyzed first. The data in Table 5 shows the fundamental findings that determine the interpretation of all subsequent experimental results.

Table 5 Price characteristics of LOPF physics model output, scenario A0

Item	2019	2024
Number of distinct model price values	5	6
Number of distinct actual price values	3,880	6,102
Model price range (lowest to highest) (EUR/MWh)	0.9 – 32.2	0.6 – 81.5
Actual price range (lowest to highest) (EUR/MWh)	−90.0 – 121.5	−135.4 – 936.3
Hours with negative prices: actual (model)	211 (0)	457 (0)
Hours with prices above 150 EUR/MWh: actual (model)	0 (0)	361 (0)
Physics model RMSE (EUR/MWh)	14.72	44.62

Figure 2 traces a single day, 3 January 2024, and it shows what the correction layer can and cannot do. Between 00:00 and 05:00 the market cleared at roughly −1.3 EUR/MWh. The LOPF formulation cannot return a negative price at all, so it sat at 23.0 EUR/MWh throughout the block, wrong by about 24 EUR/MWh every hour and wrong in a direction it has no way of

expressing. Every correction model crossed into negative territory here. XGBoost landed on −1.2 EUR/MWh at 00:00, within 0.1 EUR/MWh of the market.

The rest of the day is less flattering. From 09:00 to 12:00 the market held between 53.9 and 57.0 EUR/MWh while LOPF stayed pinned at 23.0; the corrections lifted their estimates only into the 15 to 29 EUR/MWh band, closing roughly a third of the gap. In the evening the error changed sign. XGBoost answered an actual 83.0 EUR/MWh at 17:00 with 106.7. Summed over the 24 hours the physics model records an RMSE of 19.7 EUR/MWh and every corrected series is worse, ranging from 20.8 to 29.8. On this day, correction costs accuracy.

Adding a correction layer sometimes makes the Root Mean Square Error (RMSE) look slightly worse during normal hours (which make up 90% of the time) because the basic physics model is already quite accurate then. However, overall, this correction is a massive improvement. The reason is that it fixes huge, extreme errors that only happen in the worst 10% of hours. Because RMSE is a calculation that heavily penalizes large errors, fixing these massive spikes in the worst 10% of hours far outweighs the tiny drop in accuracy during normal hours. Furthermore, we shouldn't only look at RMSE to judge performance. The correction layer is essential because it can predict negative prices which something the basic physics model is structurally unable to do, but is crucial for making real-world operational decisions in the power sector.

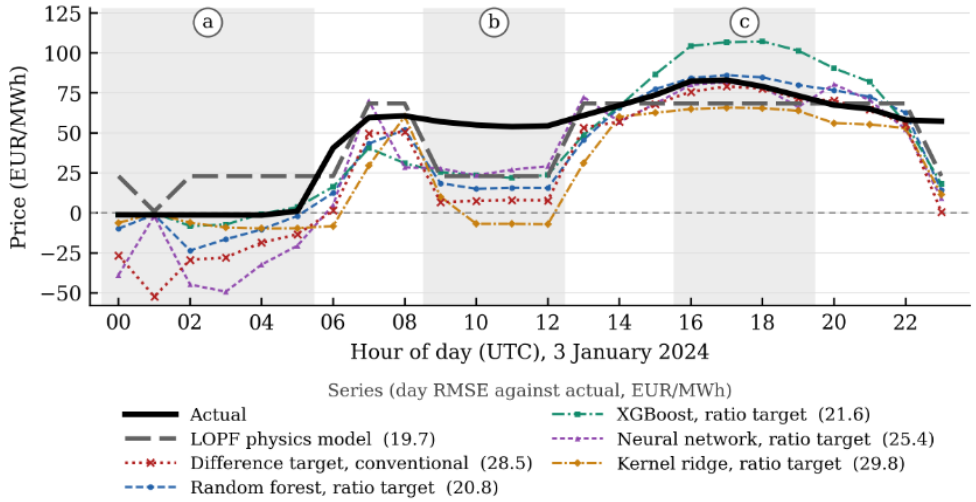


Figure 2 Comparison actual price, LOPF price and price after correction by each algorithm

3.2 Hyperparameter optimization analysis

Table 6 presents the testing results of 16 hyperparameter settings across two prediction targets. These findings directly answer the first research question about model optimization impacts. Using the conventional method, which targets the difference, all 16 settings completely failed during backward testing. In contrast, forward testing produced very similar outcomes across all configurations. The maximum performance difference observed among them was merely 2.7%. This shows that tweaking hyperparameters or changing algorithms only yields marginal improvements below 3%. Such minor gains mean little compared to error surges exceeding 100% during cross-year testing.

3.3 Target parameterization analysis

Table 7 presents the testing outcomes for our best model configuration. This setup utilized a Random Forest with 500 trees (RF_C4). These results help answer the second research question regarding learning target effectiveness. The most obvious performance difference appeared during the backward cross-time testing phase. Using the standard difference target actually caused a massive spike in errors. The RMSE climbed all the way to 39.40 EUR/MWh. For comparison,

the basic physics baseline model only sat at 14.72 EUR/MWh. This outcome clearly shows that the conventional correction method failed. It ended up making the predictions almost three times worse

Table 6 RMSE improvement range (%)

Testing scheme	Difference target (Conventional method)	Ratio target (New approach)
Cross-validation 2019	+53.1 to +69.2	+41.7 to +69.8
Cross-validation 2024	+37.1 to +64.1	+34.9 to +60.1
Forward: train 2019, test 2024	+8.7 to +11.4	+0.1 to +23.5
Backward: train 2024, test 2019	−295.4 to −118.3	−46.1 to +34.1

On the other hand, using the new ratio target drastically recovered the system's performance. The RMSE dropped down to 9.69 EUR/MWh, which gave us a 34.1% accuracy improvement over the baseline model. We saw consistent benefits from the ratio target during forward testing as well. The accuracy upgrade increased significantly from +10.5% to +18.8% in that phase. A statistical test using a 95% confidence interval verified that this improvement holds real meaning. The model also proved to be completely free from future data leakage. We confirmed this by randomly shuffling the actual prices in the target period. The correction deviation from this random test was incredibly small, hitting just 8.5×10^(-14) EUR/MWh.

Table 7 Best configuration results (Random Forest, 500 trees) in EUR/MWh

Testing scheme	Physics model	Difference target	Ratio target	Ratio target, causal features
Cross-validation 2019	14.72	4.85 (+67%)	5.02 (+65.9%)	—
Cross-validation 2024	44.62	16.42 (+63.2%)	19.27 (+57.0%)	—
Forward: 2019 to 2024	44.85	40.16 (+10.5%)	36.42 (+18.8%)	37.36 (+16.7%)
Backward: 2024 to 2019	14.72	39.40 (−167.7%)	9.69 (+34.1%)	10.90 (+25.9%)
Chronological Oct–Dec 2024	71.36	48.03 (+32.7%)	49.73 (+30.3%)	57.01 (+20.1%)

3.4 Fuel cost dynamics as the primary determinant of market conditions

We often assume that different calendar years automatically reflect different market conditions. Yet, a closer examination reveals that this common assumption can actually be misleading. Monthly error dispersion throughout 2024 alone varied extremely, ranging from 13.76 to 86.42 EUR/MWh. This created a massive 6.28-fold difference within just a single calendar year. In contrast, the inter-annual error gap between 2019 and 2024 was only 3.42-fold. These findings clearly show that intra-year variability strongly dominates over inter-year changes. Therefore, merely using calendar years to indicate changing market conditions is biased and imprecise.

Among the tested candidate metrics, natural gas prices became the primary focus. Its role is crucial because gas plants often operate as marginal generation units. Under the merit order mechanism, these specific plants dictate the final electricity price. Fundamentally, gas price fluctuations directly reflect the volatility of prevailing market conditions. This hypothesis proved accurate during our empirical evaluation. Only natural gas prices maintained a stable correlation with model error dispersion. It showed a strong correlation coefficient (+0.862) across the 24-month analysis period. This strong correlation also held true during the internal 12-month analysis in 2024.

Based on these findings, this study redefines market condition indicators using fuel cost levels. First, the value originates outside the German power system, ensuring its statistical independence. Second, it exists as a continuous numerical variable suitable for statistical correlation analysis. Third, its value is known well in advance through the established futures market. This predictability makes it highly applicable for forecasting future electricity prices. To

validate this, we grouped the 17.544 observation hours into six hierarchical clusters. These distinct clusters ranged from the lowest to the highest natural gas price levels. The model's performance was then cross-tested to observe its adaptability across environments. We wanted to see how well it handled shifting from one condition to another. Testing from one source cluster to different target clusters mathematically generated 30 possible combinations. This setup represented 30 distinct cross-market condition transitions (6 x 5 cluster pairs).

Figure 3 shows that the conventional difference-based target experiences significant performance degradation, with errors increasing as the fuel cost gap between training and testing conditions widens. In contrast, the proposed ratio-based target maintains stable performance across different market conditions, showing limited sensitivity to regime shifts. This improvement does not result from increased model complexity but from a more effective target formulation. By representing the correction as a ratio rather than an absolute difference, the model becomes less dependent on market price levels and better generalizes across varying operating conditions.

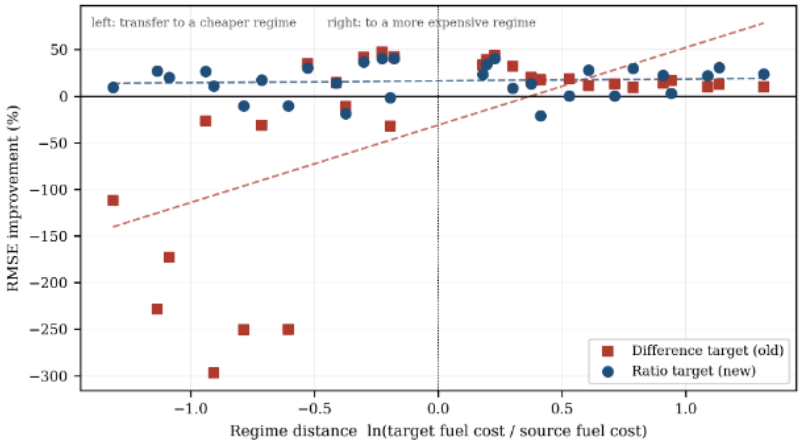


Figure 3 RMSE improvement of fuel cost regime transfers

3.5 Robustness under extreme volatility: the 2021–2022 European gas crisis

To check if our model stays reliable during a real market shock, we extended our test to 2022, the peak of the European gas crisis. We rebuilt Scenario A0 for 2022 using the same previously setup. For consistency, CO2 prices across all years use EEX annual auction averages. However, due to data limits, 2022 gas and coal prices use World Bank monthly averages, while 2019 and 2024 keep their original data. Reflecting the severe crisis, the physics-only model error (RMSE: 159.06 EUR/MWh) is 3.5 times higher than 2024 and 10.8 times higher than 2019.Using our locked Random Forest setup from Section 3.2, we compared the difference and ratio targets across five transfer schemes between 2022 and the other years. Table 8 shows the final RMSE improvements over the physics baseline.

Table 8 RMSE improvement (%) for schemes involving 2022, Random Forest 500 trees

Scheme	**Difference target**	**Ratio target**
CV 2022 (in-distribution)	+66.1%	+60.3%
Forward 2019→2022	+2.5%	+8.0%
Forward 2024→2022	+8.6%	+11.0%
Backward 2022→2019	–935.9%	–139.2%
Backward 2022→2024	–177.6%	–114.9%

Both feature sets show a clear pattern: the ratio target beats the difference target in all four transfer tests involving 2022. This includes moving into the crisis period, where the ratio target improves accuracy by +8% to +11%. However, both targets fail when moving out of the 2022 crisis back into normal years. Still, the ratio target drops far less severely than the difference

target. A model trained on extreme crisis data simply cannot adjust back to normal conditions, but the ratio format handles this shift much better. Therefore, our robustness claim holds during a crisis, but with a clear limit. Finally, testing different cap bounds shows that moving into the crisis is highly stable. In contrast, moving from crisis to normal is sensitive to the cap; a tighter cap $c = 3$ reduces the performance drop. We kept our default $c = 5$ to follow our original protocol and report this sensitivity as a transparent limitation.

4. CONCLUSIONS

This study evaluates a machine learning correction layer for the LOPF electricity price model. It assesses the model's reliability across shifting market conditions over varying time periods. Results show that the adaptation failure across different years has a deeper root. Merely retuning 16 hyperparameter configurations on various algorithms cannot fix this core issue. The fundamental cause of this generalization failure lies in the conventional learning target. It traditionally relies on an absolute difference formulation measured in EUR/MWh. This standard approach forces algorithms, like decision tree ensembles, into a blind spot. They end up predicting numerical values outside their established training data range. Transforming the learning target into a ratio successfully recovered performance across four testing schemes. Backward testing accuracy saw a massive recovery, shifting from a -167.7% drop to a +34.1% gain. Forward testing accuracy also improved, rising consistently from +10.5% to +18.8%. This proposed setup proved highly resilient when paired with pure causal features. These specific features strictly avoided any future data leakage during the training phase. Under this setup, the model still achieved solid improvements in both directions. It gained +25.9% in backward testing and +16.7% in forward testing.

Fuel cost variation is identified as the main driver of market shifts, having a greater impact than temporal distance. The difference-based target degrades as regime gaps increase, while the ratio-based target remains stable across conditions. This demonstrates that the proposed correction layer improves robustness and cross-regime generalization. Extending our test to the 2022 European gas crisis, the target ratio yields a small benefit when entering a crisis but fails when exiting it. However, it fails when transferring data from a crisis back to a normal period. Future work should test this limit using advanced physics models and extra market indicators.

REFERENCES

[1] J. Trebbien, L. R. Gorjão, A. Praktiknjo, B. Schäfer, and D. Witthaut, "Understanding electricity prices beyond the merit order principle using explainable AI," Energy AI, vol. 13, p. 100250, Mar. 2023 [Online]. Available: https://doi.org/10.1016/j.egyai.2023.100250

[2] M. Frysztacki, J. Hörsch, V. Hagenmeyer, and T. Brown, "The strong effect of network resolution on electricity system models with high shares of wind and solar," Appl. Energy, vol. 291, p. 116726, Mar. 2021 [Online]. Available: https://doi.org/10.1016/j.apenergy.2021.116726

[3] A. Papavasiliou, Optimization Models in Electricity Markets. Cambridge, U.K.: Cambridge University Press, 2024 [Online]. Available: https://doi.org/10.1017/9781009416627

[4] M. Nazari-Heris, S. Asadi, B. Mohammadi-Ivatloo, M. Abdar, H. Jebelli, and M. Sadat-Mohammadi, Eds., Application of Machine Learning and Deep Learning Methods to Power System Problems, Power Systems. Cham, Switzerland: Springer, 2021 [Online]. Available: https://doi.org/10.1007/978-3-030-77696-1

[5] S. Ghimire, R. C. Deo, D. Casillas-Pérez, and S. Salcedo-Sanz, "Two-step deep learning framework with error compensation technique for short-term, half-hourly electricity price forecasting," Appl. Energy, vol. 353, p. 122059, Oct. 2023 [Online]. Available: https://doi.org/10.1016/j.apenergy.2023.122059

[6] B. Zakeri, I. Staffell, P. E. Dodds, M. Grubb, P. Ekins, J. Jääskeläinen, et al., "The role of natural gas in setting electricity prices in Europe," Energy Rep., vol. 10, pp. 2778–2792, Sep. 2023 [Online]. Available: https://doi.org/10.1016/j.egyr.2023.09.069

[7] M. Liebensteiner, F. Ocker, and A. Abuzayed, "High electricity price despite expansion in renewables: How market trends shape Germany's power market," Energy Policy, vol. 198, p. 114448, Dec. 2024 [Online]. Available: https://doi.org/10.1016/j.enpol.2024.114448

[8] A. N. Alkawaz, A. Abdellatif, J. Kanesan, A. S. M. Khairuddin, and H. M. Gheni, "Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model," IEEE Access, vol. 10, pp. 108021–108033, 2022 [Online]. Available: https://doi.org/10.1109/access.2022.3213081

[9] W. Shi, Y. Wang, Y. Chen, and J. Ma, "An effective Two-Stage Electricity Price forecasting scheme," Electr. Power Syst. Res., vol. 199, p. 107416, Jun. 2021 [Online]. Available: https://doi.org/10.1016/j.epsr.2021.107416

[10] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting," Appl. Energy, vol. 353, p. 122079, Oct. 2023 [Online]. Available: https://doi.org/10.1016/j.apenergy.2023.122079

[11] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark," Appl. Energy, vol. 293, p. 116983, Apr. 2021

[12] G. Kapoor and N. Wichitaksorn, "Electricity price forecasting in New Zealand: A comparative analysis of statistical and machine learning models with feature selection," Appl. Energy, vol. 347, p. 121446, Jun. 2023 [Online]. Available: https://doi.org/10.1016/j.apenergy.2023.121446

[13] S. Haben, M. Voß, and W. Holderbaum, Core Concepts and Methods in Load Forecasting. Cham, Switzerland: Springer, 2023

[14] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," Knowl.-Based Syst., vol. 245, p. 108632, Mar. 2022 [Online]. Available: https://doi.org/10.1016/j.knosys.2022.108632

[15] H. Yang and K. R. Schell, "Real-time electricity price forecasting of wind farms with deep neural network transfer learning and hybrid datasets," Appl. Energy, vol. 299, p. 117242, Jun. 2021 [Online]. Available: https://doi.org/10.1016/j.apenergy.2021.117242

[16] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An Introduction to Statistical Learning, Springer Texts in Statistics. Cham, Switzerland: Springer, 2023 [Online]. Available: https://doi.org/10.1007/978-3-031-38747-0

[17] E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, Eds., Hyperparameter Tuning for Machine and Deep Learning with R. Singapore: Springer, 2023 [Online]. Available: https://doi.org/10.1007/978-981-19-5170-1

[18] E. Sarmas, V. Marinakis, and H. Doukas, Artificial Intelligence for Energy Systems, Learning and Analytics in Intelligent Systems. Cham, Switzerland: Springer, 2025

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016