STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Spring 2026

HYBRID ELECTRICITY PRICE FORECASTING IN SWEDISH BIDDING ZONES

Elias Paulsson (42849) and Oliver Westin Bergh (42859)

Abstract: Electricity prices spike sharply, cluster in volatility, and respond to weather, demand, and grid conditions in ways no single framework captures cleanly. This thesis examines whether machine-learning residual corrections add forecast value on top of a transparent GARCH-X baseline across Sweden's four bidding zones (SE1-SE4). Using hourly ENTSO-E, SMHI, and SCB data for 2024-2025, we evaluate GARCH-X and three residual learners (XGBoost, Random Forest, LSTM) in rolling 90-day out-of-sample backtests over the full 2025 calendar year at the intraday (h=1) and day-ahead (h=24) horizons.

The findings are conditional on a constrained exogenous-information design. Moving from a plain GARCH(1,1) benchmark to GARCH-X, with autoregressive lags and exogenous fundamentals, substantially improves accuracy in all four zones. Beyond that, machine-learning corrections add value selectively. At h=1, the best residual-learning hybrid significantly improves on GARCH-X in every zone under Newey-West-adjusted Diebold-Mariano comparisons, with the largest gains from XGBoost in SE3 and SE4 and smaller gains from Random Forest in SE1 and XGBoost in SE2. At h=24, ARX is best in SE1-SE3 while Random Forest is best in SE4; only the SE4 day-ahead gain is statistically significant. The practical message is diagnostic rather than universal: GARCH-X is the main econometric baseline, and residual-learning overlays are useful only where it leaves stable, forecastable residual structure. The findings offer actionable model guidance for producers, traders, and system operators operating across Sweden's structurally distinct bidding zones.

Keywords: electricity price forecasting; GARCH-X; residual learning; hybrid models; XGBoost; LSTM; machine learning; Swedish bidding zones

JEL: C22, C45, C52, C53, Q41

Supervisor:	Rickard Sandberg
Date submitted:	2026-06-07
Date examined:	2026-05-29
Discussants:	Juan Francisco Perez
Examiner:	Sampreet Goraya

Contents

1 Introduction

1.1 Context and Motivation

Electricity prices exhibit a unique combination of characteristics that make them particularly challenging to model and forecast. Unlike most financial assets, electricity cannot be economically stored at scale, and prices are therefore determined by real-time balances between supply and demand. This leads to pronounced within-day seasonality, strong persistence, sudden price spikes, and periods of elevated volatility. Accurate short-term electricity price forecasts are crucial for market participants, including producers, retailers, and system operators, as well as for policymakers concerned with market efficiency and system reliability.

Electricity markets are among the most volatile and complex commodity markets due to non-storability, real-time balancing requirements, and pronounced seasonal demand and supply fluctuations. Modelling and forecasting electricity prices, while accounting for time-varying uncertainty, is thus an economically and operationally relevant challenge for market participants, risk managers, regulators and policymakers (Weron, 2014; Lago et al., 2021).

The importance of electricity price forecasting has grown with market liberalisation and increased penetration of renewable generation. High shares of intermittent generation such as wind and solar add complexity and unpredictability to the price formation process, making traditional time-series approaches less effective without adjustments for nonlinear dynamics (Ganczarek-Gamrot et al., 2025). Ganczarek-Gamrot et al. (2025) provide recent comparative evidence from seven European markets that zones with higher shares of low-controllability renewable generation tend to display greater relative price variability and less stable forecasting performance.

Nordic and European electricity markets, including Sweden's four bidding zones (SE1–SE4), display significant regional heterogeneity in price dynamics and volatility patterns driven by zonal transmission constraints and differences in generation mix (Koopman et al., 2007). This heterogeneity underscores the need for models that not only capture temporal dynamics but also accommodate structural peculiarities of individual zones.

Traditional econometric models, and in particular models from the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) family, have been widely used to capture volatility clustering and persistence in electricity prices. Extensions of these models that incorporate exogenous variables (GARCH-X) allow economic fundamentals such as load, generation mix, and cross-border flows to influence price dynamics directly. While such models are transparent and theoretically grounded, they are inherently parametric and may struggle to fully capture nonlinear relationships and regime-dependent effects that

are prevalent in modern electricity markets.

At the same time, machine learning methods have gained increasing attention in energy economics due to their ability to model complex nonlinear patterns and interactions in high-dimensional data. Gradient boosting algorithms, such as XGBoost, have shown strong predictive performance in a wide range of forecasting applications. However, purely data-driven machine learning models often lack economic interpretability and do not explicitly model conditional volatility, which is a central feature of electricity prices.

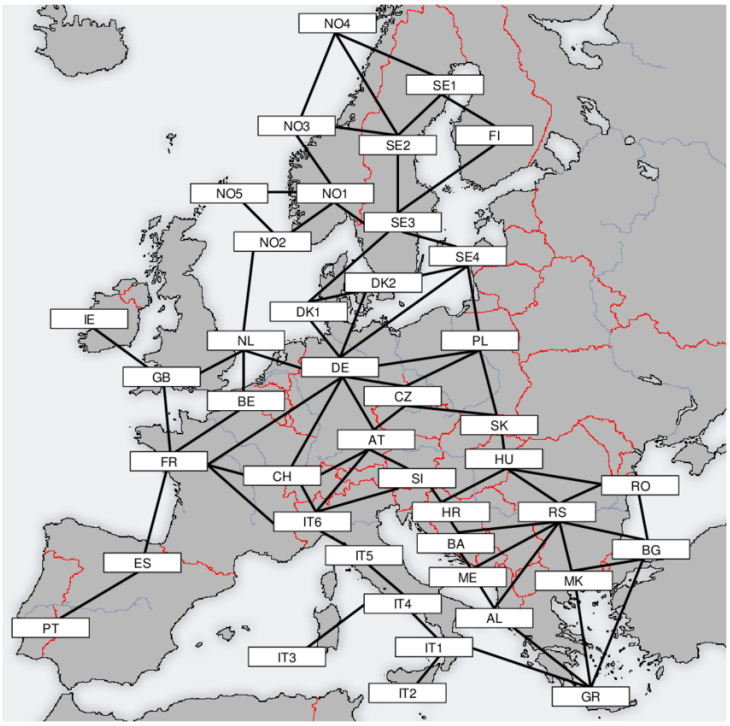


Figure 1: European bidding zone configuration and cross-border connections relevant to the Swedish market. Source: ENTSO-E.

1.2 Research Questions and Goals

Despite the extensive literature on electricity price forecasting, there remains no consensus on the most effective modelling framework, particularly when it comes to producing accurate short-horizon price forecasts in environments with strong heteroskedasticity and price spikes. Classical econometric methods such as GARCH and its variants have been applied to capture conditional heteroskedasticity in electricity prices (Koopman et al., 2007; Weron, 2014), but these models are parametric and may struggle with nonlinear dependencies and structural changes in modern markets. At the same time, machine learning (ML) approaches have demonstrated strong predictive performance in electricity price forecasting tasks, especially in capturing complex nonlinear interactions that parametric models may not easily detect (Lago et al., 2021; Henje Scott and Mellander, 2025).

Accordingly, this thesis addresses the following central research questions:

- **RQ1:** Under the constrained exogenous-information design, do machine-learning residual corrections (XGBoost, Random Forest, LSTM) deliver economically and statistically significant out-of-sample improvements over GARCH-X at the one-hour and 24-hour-ahead horizons?
- **RQ2:** How do price dynamics and time-varying volatility differ across Sweden's bidding zones, and how far do differences in residual structure and economically interpretable drivers account for cross-zone forecast performance?
- **RQ3:** Under the maintained information-set assumptions, which specification provides the most credible balance between forecast accuracy, robustness, and interpretability across horizons?

While the primary forecasting target in this thesis is the electricity price level, the analysis explicitly models volatility as a time-varying feature of forecast uncertainty that is directly relevant to economic decision-making and market design.

1.3 Contribution and Preview of Findings

This thesis asks a fairly narrow question: once a well-specified GARCH-X baseline is in place, do machine-learning residual corrections still add forecast value in Swedish electricity prices under a constrained (constant) exogenous-information design? To answer that question, the thesis builds a zone-specific GARCH-X framework with observable market fundamentals and time-varying volatility, and then evaluates XGBoost, Random Forest, and LSTM as residual correctors under a common backtest design.

The contribution is empirical rather than methodological. The thesis does not argue that residual-learning hybrids dominate in general. Instead, it asks where the extra layer earns its complexity: after the econometric baseline has absorbed the main linear dynamics, across zones with different market exposure, and under information assumptions that a forecaster could plausibly maintain in operation.

Within this constrained (constant) design, the main-horizon pattern is fairly consistent but not uniform. Over the full 2025 evaluation period, GARCH-X remains the main econometric baseline at $h = 1$, and the best residual-learning hybrid improves on it in every zone: Random Forest in SE1 and XGBoost in SE2–SE4. The largest short-horizon gains occur in SE3 and SE4. At $h = 24$ within the same constrained (constant) design, the ranking is more conservative: ARX is best in SE1–SE3, while Random Forest is best in SE4, and only the SE4 day-ahead residual-learning improvement is statistically significant in the best residual-learning comparison.

Three points guide the rest of the analysis:

- Machine-learning residual corrections are useful in this setting only when diagnostics

indicate that the GARCH-X baseline still leaves a stable residual pattern worth exploiting.

- In this study, that condition holds mainly at the one-hour horizon, with the largest gains in SE3 and SE4; at the 24-hour horizon under the constrained design, ARX is best in SE1–SE3 and only the SE4 Random-Forest hybrid provides a statistically significant improvement over GARCH-X.
- The cross-zone pattern is graded rather than binary: the strongest one-hour residual-learning gains occur in SE3 and SE4, but the full-year evidence also shows smaller, statistically significant short-horizon gains in SE1 and SE2.

The thesis is best understood as a study of model choice under operationally realistic information constraints rather than a proposal for a new forecasting architecture. The contribution is in showing, under a realistic information set, where residual-learning hybrids earn their added complexity and where they do not. Keeping that question narrow is deliberate: it is what makes the results actionable in practice.

Practical takeaway. Machine-learning residual corrections are most credible at the one-hour horizon in this constrained design, especially in SE3 and SE4, where the gains are largest. For day-ahead forecasting under the constrained design, transparent mean-equation models remain the default choice in three of four zones; the SE4 Random-Forest result is a useful exception in this sample, not a general rule.

1.4 Thesis Structure

The remainder of this thesis is organised as follows. Section 2 reviews the relevant literature on electricity price forecasting and volatility modelling. Section 3 describes the data sources, variables, and preprocessing steps. Section 4 outlines the econometric and machine learning methods employed. Section 5 presents empirical results and model evaluation. Section 6 discusses the findings in relation to the research questions. Section 7 concludes with key insights, limitations, and directions for future research.

2 Literature Review

2.1 Electricity Price Forecasting: Overview

Electricity price forecasting (EPF) has been an active research area since the deregulation of power markets in the 1990s. Weron (2014) provides the field's most comprehensive review, organising approaches into four broad families: multi-agent models that simulate strategic bidding behaviour, fundamental models that derive prices from physical supply and demand, reduced-form statistical models that capture price dynamics directly, and

computational intelligence methods including machine learning. A central finding of that survey is that no single modelling class dominates across markets and horizons, and that careful benchmarking against simple reference models is essential for drawing valid conclusions about predictive performance.

More recently, Lago et al. (2021) benchmark a wide range of statistical and deep learning methods across multiple European markets and propose a set of best practices for EPF research. Their key methodological recommendation, that new algorithms should be evaluated over full calendar years and across multiple markets rather than short test windows, directly motivates the rolling out-of-sample design used in this thesis. They find that simple linear models with appropriate regularisation remain competitive with deep learning approaches in many settings, suggesting that the marginal benefit of model complexity is often limited in EPF. Nowotarski and Weron (2018) extend this literature to probabilistic EPF, providing guidelines for uncertainty quantification that are directly relevant to the time-varying volatility modelled by the GARCH-X framework here.

2.2 GARCH Models for Electricity Price Volatility

Electricity prices exhibit strong volatility clustering, with periods of relatively stable prices alternating with episodes of extreme variability. The modelling of time-varying volatility originates from the autoregressive conditional heteroskedasticity (ARCH) framework introduced by Engle (1982), later generalised by Bollerslev (1986) into the GARCH model. The core idea is intuitive: tomorrow's volatility is likely to be high if today's was high. Formally, the conditional variance evolves as a function of past forecast errors and past variance, so the model updates its uncertainty estimates in response to recent shocks.

In electricity markets, GARCH-type models are well suited to capturing the price spikes, heavy tails, and regime shifts that characterise these series.

For example, Koopman et al. (2007) apply periodic seasonal Reg-ARFIMA–GARCH models to electricity spot prices and document strong evidence of volatility clustering and seasonal patterns across European markets. Hickey et al. (2012) apply ARMAX–GARCH models to hourly MISO hub prices, illustrating how autoregressive mean dynamics and time-varying volatility can be combined in electricity-price applications. Weron and Misiorek (2008) and Misiorek et al. (2006) provide complementary benchmarking evidence that AR/ARX–GARCH model families remain competitive with non-linear alternatives in spot electricity markets, supporting the use of GARCH-X as a demanding baseline rather than a deliberately weak comparator.

In addition to capturing volatility, extensions of GARCH models that incorporate exogenous variables allow market fundamentals such as demand, generation mix, and transmission constraints to influence price dynamics directly. This is particularly relevant in electricity

markets, where price formation is closely tied to physical system conditions and network congestion.

2.3 Machine Learning in Electricity Price Forecasting

The application of machine learning to EPF has grown substantially since the mid-2010s. In line with the broader evidence summarised by Lago et al. (2021), this thesis treats machine learning as a complement to, rather than a replacement for, well-specified linear benchmarks. The importance of rolling window length for model stability in EPF is also discussed by Marcjasz et al. (2018), who show that the choice of calibration window has a significant impact on forecast accuracy in high-dimensional autoregressive models for day-ahead prices. The three machine learning methods employed in this thesis as residual correctors are reviewed in turn below.

2.3.1 XGBoost

Gradient boosting methods, and XGBoost in particular, are attractive in this thesis because the second-stage target is not the price level itself but the residual left by the GARCH-X filter. XGBoost, introduced by Chen and Guestrin (2016) and grounded in the gradient boosting framework of Friedman (2001), builds regularised trees sequentially so that later trees focus on errors left by earlier ones. This makes it well suited to testing whether interactions among load, wind and hydro shares, net flows, outages, calendar effects, and fitted GARCH-X volatility remain useful after the linear mean equation has been estimated. Empirical evidence in electricity price forecasting, including the benchmark study of Lago et al. (2021), highlights the strong predictive performance of boosting methods across multiple markets.

A directly related application is the hybrid linear regression plus XGBoost approach of Henje Scott and Mellander (2025), a student thesis rather than a peer-reviewed article but nonetheless a directly analogous reference, which forecasts long-term electricity spot prices in the Norwegian NO4 bidding zone using a two-stage model in which XGBoost corrects the residuals of a linear regression baseline. Their study finds that XGBoost improves overall forecasting accuracy by capturing nonlinear relationships in the data. The present thesis applies an analogous residual correction architecture at short horizons in the Swedish market. Whether equivalent nonlinear structure exists at one- and 24-hour-ahead horizons is an empirical question: the shorter training window and higher signal-to-noise challenge at short horizons may be insufficient for tree-based models to learn stable nonlinear patterns, in contrast to the multi-year training set used by Henje Scott and Mellander (2025) at a long-term horizon.

More broadly, Ganczarek-Gamrot et al. (2025) examine how price volatility across seven

European bidding zones affects the predictive performance of machine learning models including Random Forest, Support Vector Regression, k-Nearest Neighbour, and Decision Tree regressors. Using ENTSO-E data for 2023–2024, they find that bidding zones dominated by low-controllability renewable generation exhibit greater price volatility and reduced forecast accuracy, while zones with higher shares of controllable sources demonstrate more stable prices and better model performance. A key methodological finding directly relevant to this thesis is their result on training window length: for highly volatile markets, the best forecasts were obtained with very short training windows of one to six days, while for more stable markets longer windows of 60–100 days performed better. This suggests the 90-day rolling window used in the present study may be appropriate for the lower-volatility northern zones (SE1–SE2) but potentially suboptimal for the more volatile southern zones (SE3–SE4).

2.3.2 Random Forest

Random Forest, introduced by Breiman (2001), provides a useful non-boosted tree benchmark for the same residual-correction task. It averages many independently grown trees, each trained on a bootstrap sample and a random subset of features, so it can capture nonlinear relations among the same load, flow, outage, weather, calendar, and statistical-context variables without sequentially chasing residual errors. In electricity price forecasting, Ganczarek-Gamrot et al. (2025) include Random Forest among four machine learning regressors benchmarked across seven European bidding zones, finding that it performs comparably to other tree-based methods in stable market conditions but deteriorates alongside all ML models in high-volatility environments.

Relative to XGBoost, Random Forest does not build trees sequentially to minimise a residual objective, which makes it a useful architectural comparator: if Random Forest and XGBoost perform similarly, the sequential boosting structure of XGBoost adds little value; if XGBoost dominates, the residual-minimising property of gradient boosting is what drives the improvement.

2.3.3 Long Short-Term Memory Networks

Recurrent neural networks, and Long Short-Term Memory (LSTM) networks in particular (Hochreiter and Schmidhuber, 1997), have attracted growing attention in energy forecasting due to their ability to capture temporal dependencies across sequences of arbitrary length through learnable gating mechanisms. Unlike tree-based methods, which treat each observation as an independent input vector, LSTM networks process inputs as ordered sequences, allowing the model to condition predictions on the full history of the input signal rather than a fixed lag window. In electricity price forecasting, Lago et al. (2021) document competitive performance of deep learning architectures including recurrent networks across

multiple European day-ahead markets, though they find that well-specified linear models remain competitive in many settings. The relative advantage of recurrent architectures is most pronounced at longer forecast horizons where multi-step temporal dependencies accumulate, a property directly relevant to the 24-hour-ahead horizon evaluated in this thesis. Whether this advantage materialises in a residual correction setting, where the LSTM is trained only on the unexplained component of a well-specified ARX baseline rather than the full price process, is an empirical question addressed in Section 5.

2.4 Hybrid Econometric–Machine Learning Models

The broader literature on hybrid models that combine econometric baselines with machine learning corrections is mixed. The general principle–train a nonlinear model on the residuals of a well-specified linear model–has theoretical appeal: the linear model handles the predictable component of the series, leaving the machine learning model to capture remaining structure. In practice, the success of this approach depends critically on whether genuine nonlinear structure remains in the residuals after the baseline is fitted.

When baseline residuals are dominated by stochastic noise, the machine learning correction risks learning noise rather than signal. This is a well-recognised failure mode of residual correction methods in time-series forecasting (Lago et al., 2021): in competitive short-term markets with efficient price discovery, the information advantage of nonlinear corrections diminishes rapidly. The residual signal diagnostics confirm this pattern: at $h = 24$, residual correlations are weaker and less stable than at $h = 1$ (roughly 0.03–0.27 across models; Table 11), and sign-flip rates are mostly above 50%. At $h = 1$, tree-based residual corrections show clearer signal, with correlations around 0.25–0.43 and the largest values in SE3 and SE4, where the forecast gains are also largest. Empirical evidence further suggests that this problem is amplified in high-volatility environments (Ganczarek-Gamrot et al., 2025). The present findings contribute empirical evidence of this failure mode in the context of Swedish electricity markets and document a specific instance of regime-shift vulnerability–a late-December 2025 stress episode–where the residual correction layer failed to generalise to out-of-support price states.

This highlights a key limitation of residual-correction frameworks. Their performance depends not only on model flexibility, but critically on the presence of stable and exploitable structure in the baseline residuals. In that respect, the present thesis contributes less by proposing a new hybrid design than by showing the conditions under which a familiar hybrid logic is and is not worth using in practice.

2.5 Cross-Zone Heterogeneity in the Swedish Market

Sweden's four electricity bidding zones (SE1–SE4) were introduced in 2011 to reflect transmission constraints between the northern, central, and southern parts of the country. Prices in the northern zones (SE1–SE2) are generally lower and less volatile due to surplus hydropower capacity, while southern zones (SE3–SE4) are more demand-intensive and closely integrated with Continental European markets through interconnectors to Denmark, Germany, and Poland (Koopman et al., 2007).

The zonal heterogeneity documented in this thesis–particularly the higher baseline RMSE and larger low-to-high volatility spread in SE3 and SE4–is consistent with this market structure. Ganczarek-Gamrot et al. (2025) provide recent European evidence that bidding zones with higher shares of low-controllability generation sources, primarily wind and solar, tend to display greater market clearing price variability and weaker forecasting performance. This evidence is used here as external market context rather than as a direct estimate from the thesis's explanatory matrix, since the generation-share variables available for Sweden are national rather than zone-specific. The cross-zone heterogeneity also motivates the zone-specific modelling strategy: Ziel and Weron (2018) show that univariate zone-by-zone frameworks are competitive with–and at times superior to–multivariate pooled approaches for day-ahead EPF when zones display structurally distinct price dynamics.

3 Data and Variable Construction

3.1 Data Source

All electricity market data used in this thesis are obtained from the ENTSO-E Transparency Platform, which is the official data portal of the European Network of Transmission System Operators for Electricity (ENTSO-E). The platform was established under Regulation (EU) No 543/2013 with the aim of improving transparency, efficiency, and integration in European electricity markets. It provides harmonised and publicly available data reported by national transmission system operators (TSOs) across Europe (ENTSO-E, 2026; European Commission, 2013).

ENTSO-E data are widely used in academic research and policy analysis due to their comprehensive coverage, standardised reporting procedures, and regulatory backing. The platform constitutes the authoritative source for historical electricity market data within the European Union.

All electricity-market series in this thesis are retrieved programmatically through the ENTSO-E Transparency Platform API (ENTSO-E, 2026). Weather series are obtained from the SMHI open-data service (SMHI, 2026), and SCB population data are used

to construct the weights required for population-weighted weather aggregation at the bidding-zone level (Statistics Sweden, 2026). The final sample covers the period from January 2024 to December 2025 at an hourly frequency and includes all four Swedish electricity bidding zones: SE1, SE2, SE3, and SE4.

3.2 Electricity Prices

Hourly electricity prices are measured using day-ahead market clearing prices, referred to as "Energy Prices" in ENTSO-E terminology. These prices represent the outcome of the day-ahead auction, in which electricity for delivery in each hour of the following day is traded. Prices are expressed in EUR/MWh.

Day-ahead prices are the dominant reference price in European electricity markets and form the basis for forward contracts, risk management, and operational decision-making. As such, they are the most relevant price measure for analysing electricity price dynamics and time-varying uncertainty at high frequencies. Prices are reported separately for each bidding zone and are matched to the corresponding hour using CET/CEST timestamps to ensure consistency with load, generation, and transmission data.

3.3 Electricity Load

Electricity demand is measured by actual total load (MW) at the bidding-zone level, reported hourly by the Swedish transmission system operator. Load reflects realised electricity consumption and captures both economic activity and exogenous influences such as weather conditions and seasonal patterns.

Load is included as a fundamental determinant of electricity prices and time-varying uncertainty, as fluctuations in demand directly affect marginal generation costs and congestion in the transmission network.

3.4 Electricity Generation Mix

Electricity generation data are obtained from ENTSO-E's "Generation by Production Type" dataset and aggregated at the national level for Sweden. The model retains only the generation-share variables that enter the empirical specification:

- Hydro water reservoir
- Nuclear
- Wind (onshore)

For each hour, the retained generation shares are constructed by dividing generation from the relevant production type by total national generation. The remaining production

categories are used only as part of this denominator and are not retained as separate regressors. This keeps the downloaded and exported model data aligned with the variables that enter the forecasting models. The retained shares capture changes in the composition of the Swedish generation mix and are included to account for technology-specific differences in marginal costs, intermittency, and supply flexibility.

National-level generation shares are applied uniformly across bidding zones. This modelling choice reflects the highly integrated nature of the Swedish power system, where zonal price differences are primarily driven by transmission constraints rather than zonal production portfolios. Similar approaches are commonly adopted in the Nordic electricity market literature.

3.5 Transmission and Net Flows

Transmission data are obtained from ENTSO-E's physical flow dataset, which reports hourly electricity flows between bidding zones and neighbouring countries. For each Swedish bidding zone, net transmission flow is constructed by summing all incoming and outgoing flows, where imports are treated as positive and exports as negative.

Net flows capture congestion, spatial arbitrage limitations, and the degree of market integration. These factors are particularly important for explaining price differentials and time-varying uncertainty across Swedish bidding zones, especially during periods of high demand or supply stress.

3.6 System Outages

Information on generation outages is obtained from ENTSO-E's "Unavailability of Production Units" dataset. For each reported outage, unavailable capacity (MW) is computed as the difference between installed and available capacity. Outages are then distributed evenly across all hours within the reported start and end time interval.

Outages are classified into planned and unplanned events based on the reported status and reason fields. Planned outages typically reflect scheduled maintenance, while unplanned outages capture unexpected failures and operational disruptions. Hourly planned and unplanned unavailable capacity are included as separate variables to account for asymmetric supply-side shocks.

3.7 Data Cleaning and Processing

All datasets are merged into a zone-specific hourly panel for SE1, SE2, SE3 and SE4. To ensure temporal consistency, the data are reindexed to a complete hourly time grid. Duplicate timestamps are averaged, and missing observations are handled using a combination

of economically motivated imputation strategies.

Missing load, flow, temperature, and generation-share observations are imputed using economically motivated procedures that preserve local temporal structure. Outage variables are set to zero when no unavailability is reported, because outages are reported by exception. Missing electricity prices are not imputed; observations without valid prices are excluded from the final model-ready panel and recorded in the curation audit. The resulting datasets form hourly panels suitable for econometric modelling of prices with heteroskedastic errors and time-varying volatility.

3.8 Summary of Variables

Table 1 summarises the complete variable set that is used for each bidding zone.

Table 1: Variable definitions and data sources.

Variable	Description	Unit	Source
price_eur_mwh	Day-ahead spot electricity price	EUR/MWh	ENTSO-E
y	Inverse-hyperbolic-sine transformed price	-	Derived from price_eur_mwh
load_mw	Actual total system load	MW	ENTSO-E
domestic_flow_mw	Net domestic transmission flow proxy	MW	ENTSO-E
foreign_flow_mw	Net cross-border flow proxy	MW	ENTSO-E
share_hydro	Hydro generation share	[0,1]	ENTSO-E
share_nuclear	Nuclear generation share	[0,1]	ENTSO-E
share_wind	Wind generation share	[0,1]	ENTSO-E
outage_planned_mw	Planned outage level (raw in appendix, z-scored in models)	MW / z-score	ENTSO-E
outage_unplanned_mw	Unplanned outage level (raw in appendix, z-scored in models)	MW / z-score	ENTSO-E
temp_celsius	Ambient temperature	°C	Weather merge

3.9 Summary Statistics

Table 2 reports summary statistics for the pooled four-zone sample and by individual zone. All statistics are computed on the post-cleaning dataset. Price statistics are reported for both raw EUR/MWh values and the inverse-hyperbolic-sine transformed series y_t used throughout the empirical analysis (Uniejewski et al., 2018).

Table 2: Descriptive statistics, pooled sample (SE1–SE4).

Variable	N	Mean	Std. dev.	P05	Median	P95	Min	Max
price_eur_mwh	69545	34.625	40.716	-0.110	21.630	111.730	-74.110	707.520
y	69545	3.305	1.770	-0.110	3.768	5.409	-4.999	7.255
load_mw	69545	3733.204	3518.671	978.000	2009.000	11327.600	161.000	16747.000
domestic_flow_mw	69545	7.897	3174.853	-5303.004	45.839	4376.332	-8397.820	7209.330
foreign_flow_mw	69545	-890.437	1049.061	-2814.328	-706.003	375.687	-4090.898	2877.024
share_hydro	69545	0.458	0.148	0.233	0.455	0.742	0.000	0.922
share_nuclear	69545	0.222	0.130	0.000	0.274	0.363	0.000	0.767
share_wind	69545	0.264	0.148	0.056	0.247	0.529	0.000	0.772
temp_celsius	69545	6.890	8.848	-7.784	6.914	20.095	-34.158	30.071

Together, these variables provide a comprehensive representation of demand conditions, supply composition, system constraints, and exogenous shocks, forming a robust empirical

Table 3: Zone-level summary statistics by bidding zone. *Note*: retained generation mix shares (wind, hydro, nuclear) are national-level averages and do not vary across bidding zones in this dataset. Zone heterogeneity is captured through zone-specific load and price series.

Zone	N	Mean price	Price std.	Negative-price share	Avg. load (MW)	Avg. wind share	Avg. hydro share	Avg. nuclear share
SE1	17352	21.047	27.759	0.063	1250.552	0.264	0.458	0.223
SE2	17327	20.786	28.074	0.078	1725.020	0.264	0.458	0.222
SE3	17429	41.204	41.665	0.055	9469.646	0.265	0.458	0.222
SE4	17437	55.313	49.820	0.056	2465.460	0.265	0.458	0.223

Table 4: ADF stationarity tests (Dickey and Fuller, 1979) and ARCH–LM tests for y_t by zone

	SE1	SE2	SE3	SE4
ADF statistic	-29.3^{***}	-28.7^{***}	-31.2^{***}	-30.8^{***}
ARCH–LM (χ^2, lag 5)	142.4^{***}	138.6^{***}	96.3^{***}	104.1^{***}

Note: *** significant at the 1% level. ADF test with AIC-selected lags and a constant term. ARCH–LM test (lag 5) applied to the transformed price series y_t directly, prior to model estimation, to assess whether conditional heteroskedasticity is present in the raw data and thereby motivate the GARCH variance specification. Post-estimation ARCH–LM diagnostics on GARCH-X standardised residuals are reported separately in Table 7.

foundation for the forecasting framework employed in this thesis. For transparency on how the cleaned panel, lag selection, rolling estimation, forecast generation, and thesis artifacts connect in practice, Appendix E provides a concise end-to-end workflow overview.

4 Model and Methodology

Table 5: Model family overview and linked empirical hypotheses.

Model	Role in thesis	Mean specification	Volatility specification	Economic hypothesis
GARCH(1,1)	Volatility benchmark	Constant mean	Symmetric GARCH(1,1)	Pure volatility model underperforms richer specifications.
GARCH-X	Main econometric baseline	ARX mean with selected lags and exogenous drivers	Student-t GARCH(1,1)	Contemporaneous system information improves forecasts relative to pure GARCH.
GARCH-X + XGB	Hybrid nonlinear residual correction	GARCH-X plus XGBoost residual learner	Inherited baseline volatility scale	Nonlinear residual structure remains after ARX-GARCH filtering.
GARCH-X + Random Forest	Hybrid tree ensemble	GARCH-X plus RF residual learner	Inherited baseline volatility scale	Flexible interactions may help under heterogeneous market regimes.
GARCH-X + LSTM	Sequence-based hybrid	GARCH-X plus LSTM residual learner	Inherited baseline volatility scale	Sequential residual dynamics may matter especially at longer horizons.

Table 5 links each model family to a distinct empirical question. XGBoost tests for nonlinear interactions among fundamentals; Random Forest benchmarks that test by removing the sequential boosting logic; and the LSTM, implemented with weekly retraining, probes whether temporal dependence survives the ARX filter. Read together, the three contenders cover the main structural alternatives to a linear residual.

4.1 Overview and Forecasting Objective

The primary objective of this thesis is to produce accurate short-horizon forecasts of hourly day-ahead electricity prices in Sweden's four bidding zones (SE1–SE4). The analysis

focuses on two forecasting horizons that are of direct relevance for market operations: $h = 1$ (one-hour-ahead) provides a short-horizon forecast benchmark, while $h = 24$ (24-hour-ahead) corresponds to the day-ahead horizon forecast.

While the main forecasting target is the electricity price level, the modelling framework explicitly accounts for time-varying volatility through a GARCH component. This allows the model to capture volatility clustering and changing forecast uncertainty, which are central empirical features of electricity prices.

4.2 Transformation of Electricity Prices

Hourly electricity prices can take both positive and negative values and are characterised by extreme spikes. Direct modelling of prices in levels can therefore lead to numerical instability and disproportionate influence of rare extreme observations. To stabilise the variance and reduce the impact of outliers while preserving sign information, let p_t denote the day-ahead price in EUR/MWh and define the transformed series $y_t = \text{asinh}(p_t)$:

$$y_t = \text{asinh}(p_t), \tag{1}$$

The transformation behaves similarly to the logarithm for large absolute values while remaining well-defined at zero and for negative prices (Uniejewski et al., 2018). All models are estimated and primarily evaluated in the transformed price space $y_t = \text{asinh}(p_t)$, and transformed-space RMSE is the main ranking criterion throughout the thesis. Price-space diagnostics based on the inverse transformation $p_t = \sinh(y_t)$ are reported only as supplementary metrics and can be numerically unstable when rare extreme forecast errors are back-transformed. For numerical reporting, extreme transformed predictions are clipped before the inverse transformation only to avoid overflow in supplementary price-space diagnostics; transformed-space RMSE remains the basis for all model rankings. Table 5 summarises the role of each model and its linked empirical hypothesis in the thesis design. One terminological point is worth noting upfront: throughout the thesis, GARCH-X denotes the main econometric baseline that includes exogenous fundamentals, while plain GARCH(1,1) is the simpler benchmark used to assess how much those fundamentals actually contribute. Keeping this distinction clear matters because the two-step improvement–from GARCH(1,1) to GARCH-X, then from GARCH-X to a hybrid–is central to interpreting the results.

4.3 Baseline Model: GARCH-X with Exogenous Regressors

For each bidding zone, the baseline model is specified as an autoregressive model with exogenous regressors and conditionally heteroskedastic errors (GARCH-X). The transformed

price y_t is decomposed into a conditional mean μ_t and an innovation term ε_t:

$$y_t = \mu_t + \varepsilon_t. \tag{2}$$

4.3.1 Conditional Mean Specification

The conditional mean captures how today's electricity price depends on its own recent history and on observable market conditions:

$$\mu_t = \varphi_0 + \sum_{i \in \mathcal{L}} \varphi_i \, y_{t-i} + \boldsymbol{\gamma}' \mathbf{x}_t, \tag{3}$$

The first sum is the autoregressive part, past prices, weighted by coefficients φ_i, capturing price persistence. The second term, $\boldsymbol{\gamma}'\mathbf{x}_t$, is where economic fundamentals enter. Specifically, $\mathbf{x}_t$ includes electricity load, net transmission flows, generation mix shares (hydro, nuclear, wind), and planned and unplanned outage capacity. Seasonality is represented using sine and cosine transformations of hour-of-day and day-of-week, which provide a smooth periodic representation without requiring long autoregressive lag structures.

The set of lags $\mathcal{L}$ is chosen separately for each bidding zone rather than imposed uniformly, because the price dynamics in SE1 (hydro-dominated, sparsely connected north) differ structurally from those in SE3 and SE4 (demand-heavy south with stronger cross-border coupling). Through a data-driven procedure, candidate lags are screened using autocorrelation and cross-correlation evidence, then retained only if selected by at least two of three information criteria (AIC, BIC, HQIC). This keeps the model compact and reduces the risk of overfitting within the 90-day rolling windows. The resulting zone-specific lag sets are summarised in Table 6; the full selection algorithm is documented in Appendix A.1.

$$\begin{aligned}
\mathcal{L}_{\text{SE1}} &= \{1,\, 3,\, 16,\, 23,\, 27\}, \\
\mathcal{L}_{\text{SE2}} &= \{1,\, 2,\, 3\}, \\
\mathcal{L}_{\text{SE3}} &= \{1,\, 2,\, 3,\, 5,\, 16,\, 23,\, 26\}, \\
\mathcal{L}_{\text{SE4}} &= \{1,\, 2,\, 3,\, 8,\, 13,\, 23,\, 26\}.
\end{aligned} \tag{4}$$

All four zones retain strong dependence on recent hours, but the southern zones still require richer lag structures than the north. The difference is less about isolated distant target lags and more about denser short-run and intraday timing effects, especially in SE3 and SE4. SE1 combines short-run persistence with late-day lags, whereas SE2 is well described by a parsimonious 1–3 hour structure. The full lag structure, including the selected exogenous-variable lags reported in the appendix, is shown in Table 6 (compact

overview) and Table 16 (full appendix listing).

Table 6: Selected lag summary by bidding zone.

Zone	Target lags	Exogenous variables with lags	Total exogenous lag terms
SE1	1,3,16,23,27	5	9
SE2	1,2,3	4	8
SE3	1,2,3,5,16,23,26	5	6
SE4	1,2,3,8,13,23,26	6	9

4.3.2 Conditional Variance Specification

Electricity prices exhibit volatility clustering: periods of relative calm interrupted by bursts of extreme variability that a constant-variance assumption cannot capture. The innovation term is assumed to follow a GARCH(1,1) process:

$$\varepsilon_t = \sigma_t z_t, \quad z_t \sim t_\nu(0,1), \tag{5}$$

where σ_t is the time-varying conditional volatility and t_ν denotes the standardised Student-t distribution with ν degrees of freedom. The conditional variance evolves as:

$$\sigma_t^2 = \omega + \alpha\,\varepsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2. \tag{6}$$

Intuitively, this says that today's volatility is a weighted combination of three things: a long-run baseline (ω), the size of yesterday's shock (ε_{t-1}^2), and yesterday's volatility itself (σ_{t-1}^2). A market that was calm yesterday is likely to be calm today, and a market that just experienced a large surprise is likely to remain turbulent for some time. The α parameter governs how quickly volatility responds to new shocks, while β governs how persistent that elevated volatility is. Together they capture the volatility clustering observed in electricity prices. The Student-t specification accommodates heavy tails arising from price spikes and extreme events. Note that α and β belong exclusively to the variance equation (6) and are distinct from the mean equation parameters φ and $\boldsymbol{\gamma}$ in equation (3).

4.3.3 Estimation and Standardisation

Model parameters are estimated separately for each bidding zone by maximum likelihood. Before estimation, all exogenous regressors in $\mathbf{x}_t$ are standardised within each rolling training window, that is, rescaled to have zero mean and unit variance. This is necessary because the regressors operate on very different scales: electricity load is measured in megawatts and runs into the thousands, while generation shares are fractions between zero and one. Without standardisation, the optimiser would effectively treat large-scale

variables as more important simply because their numbers are bigger. Standardisation is recomputed from scratch at each re-estimation step, using only the data in the current 90-day window, so no future information leaks into the model.

4.4 Benchmark Models

4.4.1 Benchmark: GARCH(1,1) without Exogenous Regressors

To isolate the combined contribution of autoregressive price dynamics and exogenous fundamentals, a parsimonious GARCH(1,1) benchmark without exogenous regressors or AR terms is estimated. This model retains the same volatility structure as the baseline model but replaces the richer conditional mean with a constant, eliminating both the structural information set and the zone-specific lag structure. The benchmark therefore provides a lower-bound reference against which the full GARCH-X specification can be assessed, rather than a clean decomposition of the separate value of AR lags and fundamentals.

The conditional mean is a constant, so the model takes the form:

$$y_t = \varphi_0 + \varepsilon_t, \tag{7}$$

where $\varepsilon_t = \sigma_t z_t$ with $z_t \overset{\text{iid}}{\sim} t_\nu$. The conditional variance follows a standard GARCH(1,1) process with Student-t innovations, allowing for volatility clustering and heavy-tailed forecast errors. Parameters are estimated separately for each bidding zone using maximum likelihood within each rolling estimation window.

This benchmark serves two purposes. First, it quantifies the combined predictive content of the zone-specific autoregressive lag structure and the exogenous regressors included in the GARCH-X baseline, since the GARCH(1,1) lacks both. Second, it provides a lower-bound reference for evaluating whether modelling time-varying volatility alone, without price dynamics or structural fundamentals, is sufficient for useful short-horizon electricity price forecasts. The design does not separately identify the contribution of fundamentals relative to an AR-only mean equation; this limitation is revisited in Section 6.5.

A simple seasonal naive benchmark is not included as a main model in the final comparison. This is a limitation of the benchmark set rather than a claim that such benchmarks are uninformative. The main purpose of the benchmark design is to evaluate the incremental value of adding an ARX mean equation, a GARCH volatility component, and residual-learning corrections within a common rolling framework. The absence of a pure seasonal naive benchmark means that the reported comparisons should be interpreted as comparisons within this econometric–hybrid model class, not as an exhaustive benchmark exercise.

4.4.2 Ablation Benchmark: ARX without GARCH

To separate the contribution of the conditional mean equation from the contribution of the GARCH variance specification, the main forecast tables also include an ARX (no GARCH) ablation at $h = 1$ and $h = 24$. This model uses the same zone-specific autoregressive lags, exogenous regressors, rolling standardisation, and constrained future-exogenous design as GARCH-X, but estimates the mean equation with homoskedastic errors and no conditional variance recursion. It is therefore not a pure AR-only model: it retains the structural fundamentals in the mean equation. Its role is practical rather than fully decompositional: it separates the richer mean specification from the conditional-variance specification, showing how much of the GARCH(1,1)-to- GARCH-X improvement is associated with the ARX mean and how much remains attributable to time-varying volatility.

4.4.3 Main Econometric Baseline: GARCH-X

The GARCH-X model is the main econometric baseline because it combines economic interpretability with a flexible volatility structure. It captures the dominant linear relationships between prices and market fundamentals while producing time-varying estimates of forecast uncertainty. The subsequent residual-learning hybrids ask whether any nonlinear and forecast-relevant structure remains in the residuals of this econometric baseline. The asymptotic properties of QMLE in GARCH-X models with persistent covariates are studied by Han and Kristensen (2014); the reference motivates the use of this model class but does not remove the need for finite-sample diagnostic checks in the rolling application considered here.

The GARCH-X baseline refers to an ARX conditional mean with exogenous regressors and a GARCH(1,1) conditional variance. The conditional mean follows equation (3), with the exogenous information set described in Section 4.3.1. The conditional variance is specified as a GARCH(1,1) process with Student-t innovations, as defined in equations (5)–(6). All exogenous regressors are standardised within each rolling training window, as described in Section 4.3.3.

For multi-step forecasting, the future path of the exogenous regressors is governed by the information-set hierarchy defined in Section 4.7: the constrained (constant) design is the main thesis specification, the seasonal-naive design is documented as an alternative feasible information-set design, and the oracle benchmark is reserved for appendix use only.

4.5 Hybrid Contender Models

Before turning to the individual machine-learning specifications, it is worth stating the residual-correction logic in plain terms. GARCH-X is a forecasting model that uses past prices and observable market fundamentals to predict tomorrow's price. It captures the

dominant linear dynamics, but its forecasts are not perfect; the gap between the GARCH-X prediction and the realised price is the forecast residual. The hybrid contenders introduced below ask a machine-learning model to look at the same inputs and predict that residual. If the residual contains exploitable structure, the machine-learning correction is added to the GARCH-X forecast and accuracy improves. If the residual is mostly noise, the correction layer has nothing to learn and may even degrade the forecast by treating noise as signal. Whether the residuals leave enough exploitable structure is an empirical question, and one that the diagnostics in Section 5 address directly. The three contenders below approach this task with different architectures: XGBoost (gradient-boosted trees), Random Forest (independently grown trees), and an LSTM network (sequence-aware deep learning).

4.5.1 Contender: GARCH-X with XGBoost Component

The GARCH-X model is deliberately linear, it assumes that the relationship between market fundamentals and prices can be captured by weighted sums. This is a useful simplification, but electricity prices are known to behave nonlinearly: the effect of an extra gigawatt of wind generation on price is not the same when the grid is relaxed as when it is congested. XGBoost is added as a second layer to catch whatever nonlinear structure the econometric baseline leaves behind.

XGBoost (Extreme Gradient Boosting) works by building a sequence of shallow decision trees, where each tree is trained specifically on the errors of the previous ones (Chen and Guestrin, 2016). The result is an ensemble that progressively corrects its own mistakes and can represent complex, nonlinear relationships without requiring the modeller to specify their functional form in advance. Three properties make it particularly suited to this application. First, it can capture threshold effects and interactions between variables, for instance, the joint effect of high load *and* low wind on residual prices. Second, it incorporates built-in regularisation that penalises complexity, reducing the risk of overfitting in the 90-day rolling windows, which contain exactly 2,160 hourly observations before validation splits and missing-value exclusions. Third, it handles variables measured on very different scales without requiring additional preprocessing.

From a forecasting perspective, the case for a gradient-boosted residual correction rests on the hypothesis that the GARCH-X conditional mean, while capturing linear dynamics and calendar seasonality, leaves exploitable nonlinear structure in its residuals. XGBoost is therefore included not just because it often performs well, but because it is well suited to testing whether nonlinear interactions among load, generation mix, transmission flows, and volatility remain once the econometric filter has removed the dominant linear component. Lago et al. (2021) document strong performance of gradient boosting methods across multiple European day-ahead markets, attributing this to their capacity to model complex feature interactions that linear models miss. Henje Scott and Mellander (2025) demonstrate

that a linear baseline plus XGBoost residual correction improves forecast accuracy in the Norwegian NO4 zone, suggesting that long-term price dynamics contain stable nonlinear patterns that the linear model cannot capture. Whether equivalent nonlinear structure is present at the short-horizon and 24-hour-ahead horizons studied here is, however, an empirical question–one that the empirical results in Section 5 address directly.

The choice of XGBoost over alternative machine learning methods–such as random forests or neural networks–reflects several practical considerations. Random forests, while similarly non-parametric and robust to overfitting, are not designed to minimise a sequential residual and therefore have less natural motivation as a correction layer. Neural networks require substantially larger training samples to generalise reliably and are more sensitive to hyperparameter choices, making them less suitable for a 90-day rolling window with roughly 2,160 hourly observations before lag-induced sample losses. Among boosting approaches, XGBoost has emerged as a strong benchmark across structured forecasting tasks (Lago et al., 2021), and its implementation is stable and computationally efficient for the rolling backtest design used here. Hyperparameters for the tree-based residual learners are kept deliberately conservative and fixed across zones. The purpose is to test whether robust residual signal remains after GARCH-X filtering, not to maximise each learner through an extensive tuning exercise within every rolling window.

Formally, let $\hat{\mu}_t$ and $\hat{\sigma}_t$ denote the fitted conditional mean and volatility from the GARCH-X model. The residual that XGBoost is asked to predict is the one-step-ahead transformed-price error, $e_t = y_t - \hat{\mu}_t$. Equivalently:

$$e_t = y_t - \hat{\mu}_t. \tag{8}$$

XGBoost is trained to predict the target e_t using the standardised exogenous regressors $\mathbf{x}_t$, the fitted conditional mean $\hat{\mu}_t$, and the fitted conditional volatility $\hat{\sigma}_t$ as features in the residual learner. Including $\hat{\sigma}_t$ as a feature allows XGBoost to behave differently in calm versus turbulent market periods, a form of regime awareness that a static correction would lack.

All three residual-learning hybrids share the same forecast construction logic: the GARCH-X conditional mean forecast is taken as a starting point, and the machine learning model adds a predicted correction:

$$\hat{y}_{t+h}^{\text{Hybrid}} = \hat{\mu}_{t+h} + \hat{e}_{t+h}^{\text{ML}}, \tag{9}$$

where $\hat{e}_{t+h}^{\text{ML}}$ is replaced by $\hat{e}_{t+h}^{\text{XGB}}$, $\hat{e}_{t+h}^{\text{RF}}$, or $\hat{e}_{t+h}^{\text{LSTM}}$ depending on the model. The GARCH component remains responsible for time-varying volatility throughout – the machine learning layer only adjusts the conditional mean.

4.5.2 Contender: GARCH-X with Random Forest Component

To assess whether the results for the XGBoost residual correction are specific to the gradient boosting architecture or reflect a more general failure of tree-based corrections at short horizons, an alternative residual correction based on a Random Forest regressor is also estimated.

Random Forest builds its ensemble differently from XGBoost. Rather than fitting trees sequentially to correct previous errors, it grows B trees independently on random subsamples of the data and averages their predictions (Breiman, 2001). This averaging reduces variance but the trees have no awareness of each other's mistakes. If XGBoost and Random Forest perform similarly, the sequential boosting structure adds little; if XGBoost dominates, it is the residual-minimising property that drives the improvement.

The Random Forest correction is trained on the same feature set as XGBoost (Section 4.5.1): standardised exogenous regressors $\mathbf{x}_t$, the fitted conditional mean $\hat{\mu}_t$, and the fitted conditional volatility $\hat{\sigma}_t$. This makes Random Forest a natural benchmark for XGBoost: if the two models perform similarly, the sequential residual-minimising logic of boosting adds little beyond the general flexibility of tree methods; if XGBoost dominates, that sequential correction mechanism is likely doing the real work. The same rolling 90-day training window is used, and the corrected forecast is constructed analogously:

$$\hat{y}^{\mathrm{RF}}_{t+h} = \hat{\mu}_{t+h} + \hat{e}^{\mathrm{RF}}_{t+h}. \tag{10}$$

The Random Forest hyperparameters are set to commonly used defaults, namely 500 trees, one third of features per split, and no minimum leaf size constraint, to provide a conservative but representative tree-ensemble benchmark in this architecture test. This mirrors the conservative tuning strategy used for XGBoost and keeps the Random Forest comparison focused on architecture rather than model-specific hyperparameter search. Performance is reported alongside the GARCH-X baseline and the XGBoost hybrid in Table 9.

4.5.3 Contender: GARCH-X with LSTM Network Component

Tree-based methods treat each observation as an independent input vector, they have no memory of what came before. An LSTM network is fundamentally different: it reads the input as a sequence and can learn which parts of the recent history are relevant for the current prediction, retaining useful signals across many time steps through learnable gating mechanisms (Hochreiter and Schmidhuber, 1997).

The motivation for including an LSTM is architectural. If tree-based corrections fail at the 24-hour-ahead horizon because they cannot see patterns that unfold over multiple steps, a

sequence-aware model may partially compensate. Under the constrained (constant) design (Section 4.7), the GARCH-X baseline carries exogenous variables forward as constants from the forecast origin, which may leave systematic residual dynamics across the forecast horizon that an LSTM, reading the input as a 48-hour sequence, could exploit.

The LSTM residual network uses a single layer with 32 hidden units and a sequence length of 48 hours, meaning the model looks back two days of hourly observations when forming each prediction.[1] A two-day window is long enough to capture the daily seasonality cycle twice over, while staying within the 90-day training window. The 32 hidden units control how much the network can remember from that sequence, a deliberately modest capacity chosen to avoid overfitting on the small rolling samples.

Because retraining a neural network from scratch at every forecast origin would be computationally prohibitive, the LSTM is retrained at most once per week within each rolling backtest stream. This means the network weights are held fixed between retraining points, while the GARCH-X baseline and tree-based corrections continue to update at every step. The corrected forecast follows equation (9) with $\hat{e}_{t+h}^{\mathrm{LSTM}}$.

4.6 Rolling Out-of-Sample Forecasting Design

Forecast performance is evaluated using a rolling 90-day out-of-sample backtest over the full 2025 calendar year, with each model re-estimated on the preceding 90 days of data within the 2024–2025 sample. For each bidding zone, this window contains exactly 2,160 hourly observations before any missing-value exclusions, and the estimated models are then used to forecast ahead. The LSTM applies the same rolling window on the weekly retraining cadence described above. The window then moves forward and the process repeats, mimicking how a forecaster would operate in real time with only historical data available. The forecast evaluation focuses on horizons $h \in \{1, 24\}$.

To keep computation manageable without sacrificing coverage, forecast origins are stepped forward in increments of 6 hours rather than 1.

The common 90-day window is a design choice rather than an estimated optimum. The EPF literature shows that more volatile markets can sometimes benefit from shorter calibration windows, while more stable markets often favour longer windows (Marcjasz et al., 2018; Ganczarek-Gamrot et al., 2025). This thesis uses one window length across all zones to keep the rolling comparison stable, transparent, and computationally feasible. If shorter windows would adapt more quickly in SE3 and SE4, the reported residual-learning gains in those zones are likely conservative rather than inflated by window tuning.

[1] Full training details: Adam optimiser (learning rate 0.001), batch size 64, up to 50 epochs, early-stopping patience 4, dropout rate 0.10, weight decay 10^{-5}, gradient clipping at 1.0.

A technical note on forecast error autocorrelation at the 24-hour-ahead horizon: when using a direct multi-step forecasting design, forecast errors at horizon h are mechanically serially dependent, with the usual non-overlapping case often represented as an MA($h-1$) process (Diebold and Mariano, 1995; Marcellino et al., 2006). Because forecast origins in this thesis are stepped forward every six hours, adjacent $h = 24$ forecasts also overlap in calendar time, so the dependence structure is more complex than in a non-overlapping design. Ljung–Box tests at $h = 24$ are therefore not interpreted as misspecification tests. Diagnostic inference based on residual autocorrelation is confined to the $h = 1$ horizon, where the overlap issue is substantially weaker.

4.7 Information Set for Future Exogenous Variables

Forecasting more than one step ahead raises a practical problem: the model requires future values of non-deterministic exogenous variables, but these values are not known at the forecast origin. The main empirical comparison therefore uses a constrained carry-forward information set, under which non-deterministic exogenous regressors are held fixed at their last observed values. Two additional future-path designs were considered during the information-set design stage: a seasonal-naive path and an oracle path. These designs are documented for transparency, but only the constrained design is used for the quantitative model rankings and headline conclusions.

To ensure the comparison is internally consistent, the same exogenous-information design is applied both when training the models and when generating forecasts, so no design benefits from stronger information assumptions than the ones stated above.

4.7.1 Baseline Specification: Constrained (Constant) Design

Under the constrained (constant) design, each non-deterministic exogenous regressor is held fixed at its last observed value at the forecast origin. Formally, for any such regressor x_t and forecast origin t, the path used by the GARCH-X baseline satisfies $x_{t+k|t} = x_t$ for $k = 1, \ldots, h$.

4.7.2 Alternative Feasible Design: Seasonal-Naive Exogenous Paths

Under the seasonal-naive design, each non-deterministic exogenous series repeats its most recently observed 24-hour profile cyclically, preserving the most recent daily shape while avoiding realised future information.

4.7.3 Infeasible Oracle Information Set (Conceptual Upper Bound)

The oracle benchmark instead uses realised future exogenous values. It is included only as an appendix upper bound and is never interpreted as an operational forecasting design.

4.8 Forecast Evaluation Metrics

Forecast accuracy is assessed in the transformed price space using the root mean squared error (RMSE) and mean absolute error (MAE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \qquad \text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t|. \tag{11}$$

RMSE penalises large errors more heavily than small ones, a single large miss counts for more than several small ones, while MAE treats all errors proportionally to their size. Together they give a fuller picture of forecast accuracy: a model that performs well on both is reliable across both typical and extreme price episodes. The primary ranking criterion is RMSE in transformed price space.

Evaluating in transformed rather than raw price space serves two purposes. First, it avoids numerical instability from extreme price spikes, which can dominate raw-space error measures. Second, it ensures consistency with the estimation framework, since all models are fitted to transformed prices. Price-space RMSE and MAE, obtained after applying the inverse transformation $p_t = \sinh(y_t)$, are retained as secondary diagnostics, but are not used for model ranking because rare extreme misses are amplified disproportionately in raw price space. Forecast performance is reported separately for each bidding zone and horizon.

4.9 Diagnostic Checks

4.9.1 Autocorrelation function

Model adequacy is assessed using the autocorrelation function (ACF) of the standardised residuals $\hat{z}_t = \hat{\varepsilon}_t/\hat{\sigma}_t$, with $y_t = \operatorname{asinh}(p_t)$. If the baseline model has absorbed the systematic structure in prices, these residuals should be close to white noise. This diagnostic is also directly relevant to the residual-learning hybrids, because residual correction is plausible only in zones where forecast-relevant structure remains after econometric filtering.

Two further diagnostics are used for completeness. First, residual volatility clustering is assessed formally with the ARCH–LM tests reported below. Second, the estimated conditional volatility series is plotted over the 2025 evaluation year so that the model's volatility response can be checked visually against turbulent and calm market periods in the same calendar year used for the main forecast comparisons.

4.9.2 Ljung–Box and ARCH–LM Diagnostics

A formal test of residual white noise is provided by the Ljung–Box Q-statistic (Ljung and Box, 1978) and the ARCH–LM test (Engle, 1982). Both ask a simple question:

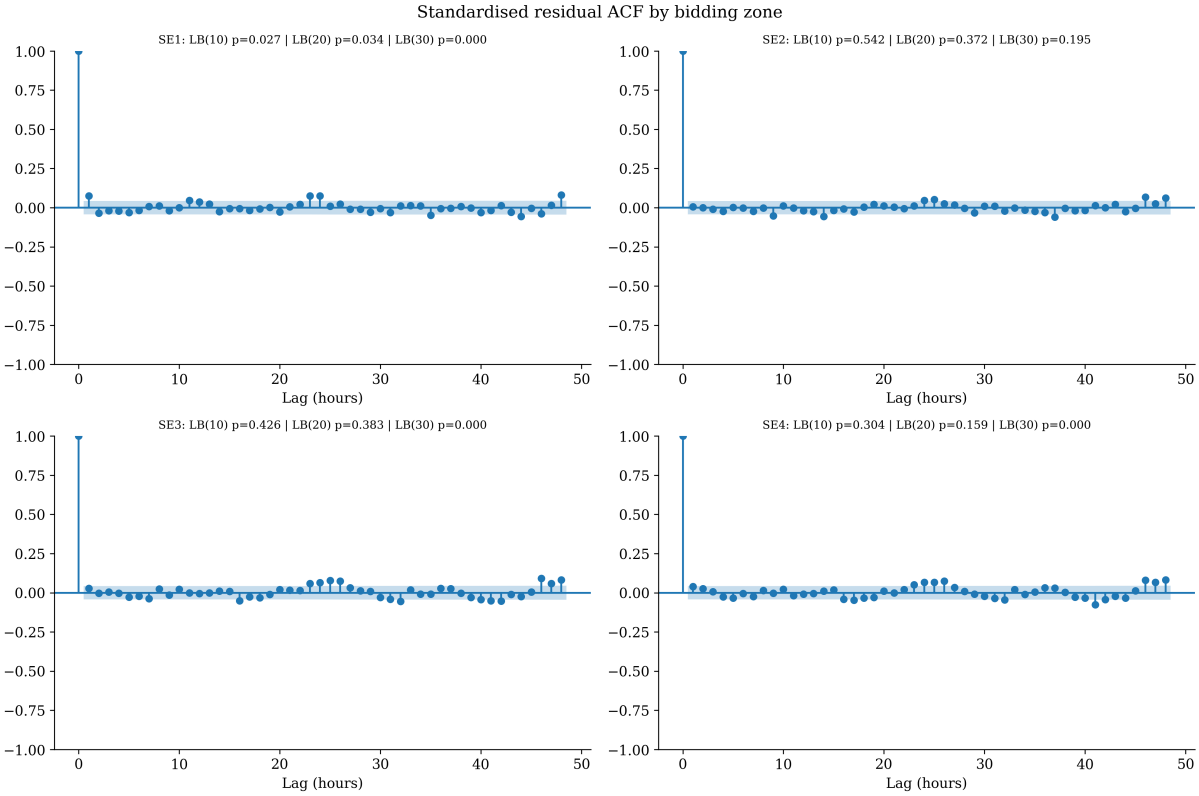


Figure 2: Autocorrelation function of standardised GARCH-X residuals ($\hat{z}_t = \hat{\varepsilon}_t/\hat{\sigma}_t$, with $y_t = \operatorname{asinh}(p_t)$) at lags 1–40, computed over the diagnostic estimation sample January 2024–September 2025. Dashed horizontal lines indicate 95% confidence bounds under the white-noise hypothesis.

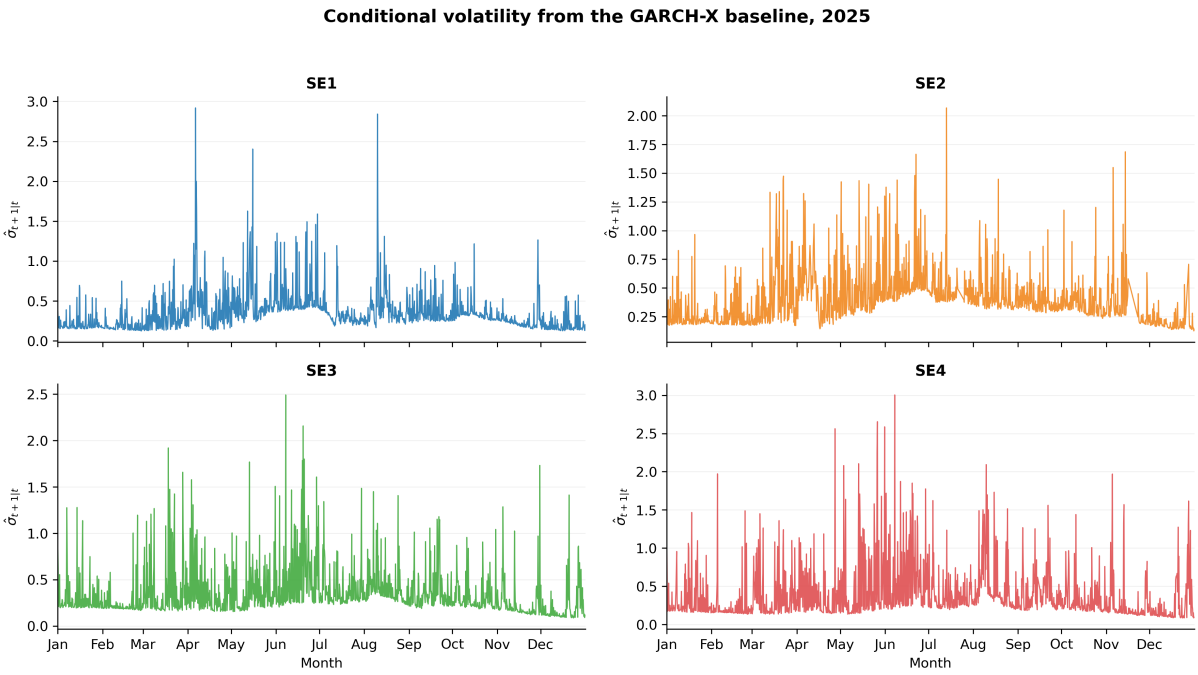


Figure 3: Estimated one-hour-ahead conditional volatility ($\hat{\sigma}_{t+1|t}$) from the GARCH-X baseline by bidding zone, plotted over the full 2025 evaluation year. Spikes correspond to periods of elevated price volatility; the series illustrates that the GARCH component responds to market stress and calms during stable periods.

after the model has done its job, does anything systematic remain in the residuals? The Ljung–Box $Q(24)$ statistic tests whether the first 24 autocorrelations of the standardised residuals are jointly zero, in other words, whether the residuals look like white noise. The ARCH–LM(12) test asks a narrower version of the same question specifically about volatility: do the squared residuals still exhibit clustering after the GARCH filter has been applied?

Table 7: Ljung–Box $Q(24)$ and ARCH–LM(12) diagnostics for standardised residuals at $h = 1$ by bidding zone. Results are computed over the diagnostic estimation sample January 2024–September 2025. p-values above 0.05 indicate no evidence against the white-noise null.

	GARCH(1,1) residual ACF		GARCH-X residual ACF		GARCH-X ARCH-LM(12)	
Zone	$Q(24)$	p	$Q(24)$	p	χ^2	p
SE1	916.7	<0.001	20.8	0.653	0.92	>0.999
SE2	696.9	<0.001	22.4	0.554	1.48	>0.999
SE3	452.7	<0.001	70.6	<0.001	23.5	0.024
SE4	417.8	<0.001	67.4	<0.001	41.1	<0.001

Table 7 reveals a stark and interpretively important pattern. The plain GARCH(1,1) residuals are severely autocorrelated in all four zones, with $Q(24)$ statistics ranging from 418 (SE4) to 917 (SE1) and p-values indistinguishable from zero. This confirms that a

constant-mean specification leaves substantial systematic price dynamics unfiltered, the model simply does not know enough about what drives prices to remove the predictable component.

The GARCH-X model produces a zone-differentiated outcome. In SE1 and SE2, standardised residuals pass the Ljung–Box test convincingly ($Q(24) \approx 20$–22, $p = 0.55$–0.65), and the ARCH–LM statistic is effectively zero ($p > 0.999$). Together, these diagnostics indicate that both the conditional mean and conditional variance are close to correctly specified in this sample. In SE3 and SE4, however, $Q(24)$ statistics of 70.6 and 67.4 reject the white-noise null ($p < 0.001$), and mild residual heteroskedasticity persists ($p = 0.024$ in SE3 and $p < 0.001$ in SE4).

These diagnostics map directly onto the thesis's central question. SE3 and SE4 retain the clearest forecast-relevant residual structure after econometric filtering, while SE1 and SE2 are closer to the white-noise benchmark in-sample. Section 5 then asks whether the residual learners can exploit any remaining structure out of sample, and whether the zones where the baseline falls short most clearly are also the zones where machine-learning corrections add the most value.

5 Empirical Results

The empirical analysis proceeds in four steps. First, it examines the GARCH parameter estimates and residual diagnostics to assess what structure remains after the econometric baseline. Second, it evaluates the gain from moving from the plain GARCH(1,1) benchmark to the ARX ablation and the GARCH-X baseline. Third, it tests whether XGBoost, Random Forest, and LSTM residual corrections improve on GARCH-X at the main horizons. Fourth, it decomposes performance by volatility regime and uses an event window to illustrate when residual-learning corrections become unstable.

5.1 GARCH Parameter Estimates and Volatility Dynamics

Table 8 reports the estimated GARCH(1,1) parameters for each bidding zone, together with an unconstrained persistence audit in which the stationarity restriction on $\hat{\alpha}_1 + \hat{\beta}_1$ is relaxed. The audit confirms that the constrained estimates sit at a near-integrated volatility boundary rather than at a clearly interior stationary solution. In SE1, SE3, and SE4 the unrestricted likelihood improves when the persistence restriction is relaxed, while SE2 shows an unreliable unrestricted optimum with a large gradient norm. The table should therefore be read as diagnostic evidence that volatility persistence is boundary-driven in this sample, not as a clean formal endorsement of an IGARCH data-generating process. Values of $\hat{\alpha}_1 + \hat{\beta}_1$ close to one imply very persistent, near-integrated volatility, but this

is treated here as a diagnostic feature of the fitted process rather than as an assumption that the true variance dynamics are exactly IGARCH. The main forecast conclusions do not depend on a strict covariance-stationarity interpretation of the variance process. The rolling backtest models impose the standard GARCH positivity and stationarity constraints ($\omega > 0$, $\hat{\alpha}_1, \hat{\beta}_1 \geq 0$, $\hat{\alpha}_1 + \hat{\beta}_1 < 1$). Given the near-boundary unconstrained optimum, the constrained rolling estimates have $\hat{\alpha}_1 + \hat{\beta}_1$ at or very close to the upper bound in all zones, consistent with the near-integrated pattern shown in the diagnostic table.

The degrees-of-freedom estimates remain low across zones, indicating pronounced heavy tails and supporting the Student-t specification over a Gaussian alternative. The evidence points to a common cross-zone pattern of highly persistent conditional volatility, with the caveat that persistence estimates near or beyond the stationarity boundary are numerically delicate. Complementary in-sample variance-calibration diagnostics (QLIKE and Mincer–Zarnowitz regressions) are reported in Appendix G and support the GARCH(1,1) specification.

As an additional diagnostic, an asymmetric GJR-GARCH(1,1,1) variance specification was estimated alongside the symmetric GARCH(1,1) on a representative 90-day in-sample window for each zone, matching the rolling window length used in the backtest design. The asymmetry coefficient $\hat{\gamma}$ is effectively zero in SE1 ($\hat{\gamma} \approx 0$) and SE2 ($\hat{\gamma} \approx 0.015$) but positive and economically meaningful in SE3 ($\hat{\gamma} \approx 0.209$) and SE4 ($\hat{\gamma} \approx 0.277$), so AIC favours GJR over symmetric GARCH only in the two southern zones ($\mathrm{AIC}_{\mathrm{GJR}} - \mathrm{AIC}_{\mathrm{GARCH}} = -7.0$ in SE3 and -11.7 in SE4, against $+2.5$ in SE1 and $+2.7$ in SE2). The symmetric GARCH(1,1) specification is retained in the main model set for three reasons: to keep a single variance specification consistent across all four zones, to avoid embedding asymmetric leverage structure directly in the variance equation when the residual-learning layer is precisely the component intended to capture remaining nonlinear structure, and because the rolling out-of-sample comparison in this thesis is focused on the value of ML residual corrections rather than on the choice between symmetric and asymmetric GARCH families. Estimating a GJR variant within the hybrid pipeline is left as a natural robustness extension; the southern-zone asymmetry is, however, a candidate explanation for why the largest one-hour residual-learning gains arise in SE3 and SE4.

Table 8: Unconstrained GARCH persistence audit by bidding zone.

Zone	ω	α	β	$\alpha+\beta$	ν	Log-likelihood	LR vs. IGARCH	p-value	Converged	Gradient norm
SE1	0.0015	0.8064	0.5183	1.3247	2.7283	88315.7159	689.7916	< 0.0001	Yes	0.1836
SE2†	0.0114	1.7638	0.1847	1.9485	2.2946	87603.5588	-3838.6042	1.0000	Yes	4397.3993
SE3	0.0022	1.6758	0.4262	2.1020	2.4290	69402.4536	2291.3486	< 0.0001	Yes	0.6641
SE4	0.0044	1.6785	0.4520	2.1305	2.3370	55750.4402	1711.5158	< 0.0001	Yes	0.2371

Note: Parameters are estimated over the diagnostic estimation sample January 2024–September 2025 using a custom Student-t ARX–GARCH likelihood audit with the stationarity constraint relaxed. These *unconstrained diagnostic* values are descriptive only and are distinct from the rolling backtest parameters, which impose $\hat{\alpha}_1 + \hat{\beta}_1 < 1$ and generate all forecasts reported in Section 5.
The SE2 unrestricted fit reports a large optimiser gradient norm (approximately 4397) and should be read as a diagnostic optimiser audit rather than as a stable parameter estimate. It is retained only to document that the relaxed likelihood does not converge to a clean interior solution for SE2; it is not used in any forecast comparison. † The SE2 unrestricted fit does not yield a stable interior optimum (gradient norm ≈ 4397).

5.2 Plain GARCH Benchmark vs. GARCH-X Baseline

Before evaluating the residual-learning hybrids, it is instructive to assess the combined contribution of the autoregressive price lags and the exogenous regressors by comparing the plain GARCH(1,1) benchmark against the full GARCH-X baseline. Because the GARCH(1,1) uses only a constant conditional mean, any RMSE improvement of the GARCH-X model reflects the joint gain from adding both the zone-specific lag structure and the structural information set. The comparison therefore establishes the value of the full ARX mean structure, but it does not by itself separate short-run price persistence from the marginal contribution of fundamentals.

At the one-hour-ahead horizon ($h = 1$), the GARCH-X baseline reduces RMSE by roughly 72.6–74.1% relative to the plain GARCH(1,1) benchmark across all four bidding zones (SE1: 0.392 vs. 1.510; SE2: 0.417 vs. 1.596; SE3: 0.424 vs. 1.633; SE4: 0.498 vs. 1.821). The improvement is large and statistically significant in all zones ($p < 0.001$ by Diebold–Mariano test), confirming that the combined effect of autoregressive price dynamics and market fundamentals carries substantial predictive content relative to a constant-mean volatility benchmark at short horizons.

At the 24-hour-ahead horizon, the absolute improvement from adding AR lags and exogenous regressors is smaller than at $h = 1$ but still economically meaningful. RMSE reductions relative to the GARCH(1,1) benchmark now range from roughly 11.2% in SE1 and SE2 (SE1: 1.370 vs. 1.542; SE2: 1.428 vs. 1.607) to about 21.1% in SE3 and 21.0% in SE4 (SE3: 1.333 vs. 1.688; SE4: 1.426 vs. 1.805). The direct multi-step GARCH-X specification therefore remains informative even under the constrained (constant) design, although the incremental benefit of the full ARX mean structure is smaller than at short horizons.

The ARX (no GARCH) ablation in Table 9 separates the mean-equation gain from the

conditional-variance component. At $h = 1$, most of the improvement over the constant-mean GARCH(1,1) benchmark already appears when the ARX mean equation is added with homoskedastic errors, while the GARCH-X variance component adds a smaller but still useful gain. At $h = 24$, ARX and GARCH-X are much closer, and in some zones the ARX ablation is slightly better in transformed-space RMSE. This reinforces that the main forecast gain comes from the richer mean equation, while volatility modelling matters most for short-horizon risk dynamics and diagnostic coherence.

Taken together, these results confirm that the GARCH-X baseline is a meaningful improvement over the plain GARCH(1,1) benchmark. The result should be read together with the ARX ablation rather than as a standalone estimate of the value of exogenous fundamentals. Having established this stronger baseline, the following sections ask whether machine learning residual corrections can add further value on top of it, and under what conditions.

5.3 Benchmark and Contender Models Forecast Performance under the Constrained (Constant) Design

All metrics, rankings, and comparisons in this section use the constrained (constant) exogenous-information design defined in Section 4.7, evaluated over the full 2025 calendar-year out-of-sample period at $h = 1$ and $h = 24$. Model rankings are based on RMSE in transformed price space; price-space metrics are reported as secondary diagnostics.

Table 9 presents rolling out-of-sample RMSE and MAE in transformed price space for all four bidding zones and both forecast horizons. Percentage improvement figures are defined as positive when the hybrid model outperforms the baseline (i.e. lower error) and negative when it underperforms. Unless otherwise noted, model rankings in this section are based on RMSE in transformed price space; price-space metrics are reported as a diagnostic and can rank close contenders differently.

All main-text tables and figures in this subsection use the main thesis information set, i.e. the constrained (constant) design defined in Section 4.7. Non-deterministic exogenous regressors are therefore carried forward from the forecast origin; neither the seasonal-naive design nor the oracle benchmark is mixed into the model comparisons reported below.

Table 9: Rolling out-of-sample forecast accuracy under the constrained (constant) design at the main horizons (1h and 24h).

Zone	h	Model	N	$RMSE_y$	MAE_y	Price RMSE	Price MAE	Price MAPE (%)	Price sMAPE (%)
SE1	1	ARX (no GARCH)	1434	0.444	0.314	11.177	4.883	30.426	37.242
SE1	1	GARCH(1,1)	1434	1.510	1.107	30.214	14.234	82.185	86.787
SE1	1	GARCH-X	1434	0.392	0.223	7.763	3.033	21.930	27.078
SE1	1	GARCH-X + LSTM	1434	0.394	0.230	7.583	3.123	22.443	27.928
SE1	1	GARCH-X + Random Forest	1434	0.378	0.218	7.721	3.016	21.260	27.050
SE1	1	GARCH-X + XGB	1434	0.380	0.220	7.515	3.030	21.709	27.251
SE1	24	ARX (no GARCH)	1427	1.351	1.003	28.032	13.381	90.471	85.637
SE1	24	GARCH(1,1)	1427	1.542	1.131	30.799	14.368	79.434	88.374
SE1	24	GARCH-X	1427	1.370	1.002	29.342	14.031	102.088	84.400
SE1	24	GARCH-X + LSTM	1427	1.436	1.062	28.986	14.145	104.993	88.594
SE1	24	GARCH-X + Random Forest	1427	1.403	1.027	29.305	13.744	92.857	85.318
SE1	24	GARCH-X + XGB	1427	1.392	1.035	28.153	13.641	95.274	87.006
SE2	1	ARX (no GARCH)	1381	0.478	0.336	11.045	4.611	32.088	42.162
SE2	1	GARCH(1,1)	1381	1.596	1.210	29.513	13.789	84.664	94.408
SE2	1	GARCH-X	1381	0.417	0.252	8.054	3.165	25.394	33.093
SE2	1	GARCH-X + LSTM	1381	0.415	0.257	7.803	3.134	25.802	33.540
SE2	1	GARCH-X + Random Forest	1381	0.400	0.244	8.009	3.093	24.081	32.407
SE2	1	GARCH-X + XGB	1381	0.390	0.242	7.933	3.082	23.958	32.017
SE2	24	ARX (no GARCH)	1375	1.422	1.092	26.757	12.949	96.021	93.815
SE2	24	GARCH(1,1)	1375	1.607	1.212	29.143	13.753	81.744	94.539
SE2	24	GARCH-X	1375	1.428	1.099	27.478	13.348	110.325	94.232
SE2	24	GARCH-X + LSTM	1375	1.570	1.233	27.877	14.255	125.725	103.164
SE2	24	GARCH-X + Random Forest	1375	1.459	1.137	27.507	13.401	101.359	96.318
SE2	24	GARCH-X + XGB	1375	1.450	1.131	27.298	13.391	104.237	96.305
SE3	1	ARX (no GARCH)	1451	0.537	0.392	25.981	14.468	40.035	41.761
SE3	1	GARCH(1,1)	1451	1.633	1.015	42.006	28.863	231.985	71.343
SE3	1	GARCH-X	1451	0.424	0.245	14.638	7.421	27.043	27.113
SE3	1	GARCH-X + LSTM	1451	0.437	0.263	14.784	7.983	29.857	29.112
SE3	1	GARCH-X + Random Forest	1451	0.405	0.241	14.473	7.391	25.964	27.096
SE3	1	GARCH-X + XGB	1451	0.386	0.235	13.691	7.074	24.776	26.585
SE3	24	ARX (no GARCH)	1447	1.293	0.954	40.420	27.792	120.181	79.205
SE3	24	GARCH(1,1)	1447	1.688	1.052	43.025	29.903	242.569	73.116
SE3	24	GARCH-X	1447	1.333	0.880	37.061	24.557	138.300	68.873
SE3	24	GARCH-X + LSTM	1447	1.337	0.921	42.491	27.258	150.317	73.143
SE3	24	GARCH-X + Random Forest	1447	1.686	0.934	n/a[‡]	33.393	130.809	74.953
SE3	24	GARCH-X + XGB	1447	1.295	0.921	n/a[‡]	33.531	129.844	75.782
SE4	1	ARX (no GARCH)	1441	0.581	0.407	39.436	20.684	42.931	41.718
SE4	1	GARCH(1,1)	1441	1.821	1.092	52.876	39.605	381.142	71.583
SE4	1	GARCH-X	1441	0.498	0.266	20.656	10.673	31.644	28.006
SE4	1	GARCH-X + LSTM	1441	0.501	0.280	19.969	11.087	32.430	29.386
SE4	1	GARCH-X + Random Forest	1441	0.472	0.261	20.566	10.628	29.173	27.509
SE4	1	GARCH-X + XGB	1441	0.451	0.252	17.863	9.782	27.401	27.014
SE4	24	ARX (no GARCH)	1436	1.395	1.018	54.522	37.543	172.433	82.052
SE4	24	GARCH(1,1)	1436	1.805	1.079	50.058	38.367	385.239	70.998
SE4	24	GARCH-X	1436	1.426	0.923	44.199	31.311	185.985	69.674
SE4	24	GARCH-X + LSTM	1436	1.481	1.027	78.195	41.861	207.097	79.043
SE4	24	GARCH-X + Random Forest	1436	1.356	0.957	44.764	32.474	139.173	76.440
SE4	24	GARCH-X + XGB	1436	1.385	0.983	45.877	33.030	137.212	78.620

Note: The unusually large price-space RMSE values for SE3 at $h = 24$ under GARCH-X + Random Forest and GARCH-X + XGB are inflated by inverse-transform amplification of isolated extreme transformed-space predictions. They are not used for ranking or inference; transformed-space RMSE remains the primary comparison metric.

Table 10: Best-performing model by zone and horizon under the constrained (constant) design (RMSE in transformed price space).

Zone	h	Best model	N	$RMSE_y$	MAE_y	Price RMSE	Price MAE
SE1	1	GARCH-X + Random Forest	1434	0.378	0.218	7.721	3.016
SE1	24	ARX (no GARCH)	1427	1.351	1.003	28.032	13.381
SE2	1	GARCH-X + XGB	1381	0.390	0.242	7.933	3.082
SE2	24	ARX (no GARCH)	1375	1.422	1.092	26.757	12.949
SE3	1	GARCH-X + XGB	1451	0.386	0.235	13.691	7.074
SE3	24	ARX (no GARCH)	1447	1.293	0.954	40.420	27.792
SE4	1	GARCH-X + XGB	1441	0.451	0.252	17.863	9.782
SE4	24	GARCH-X + Random Forest	1436	1.356	0.957	44.764	32.474

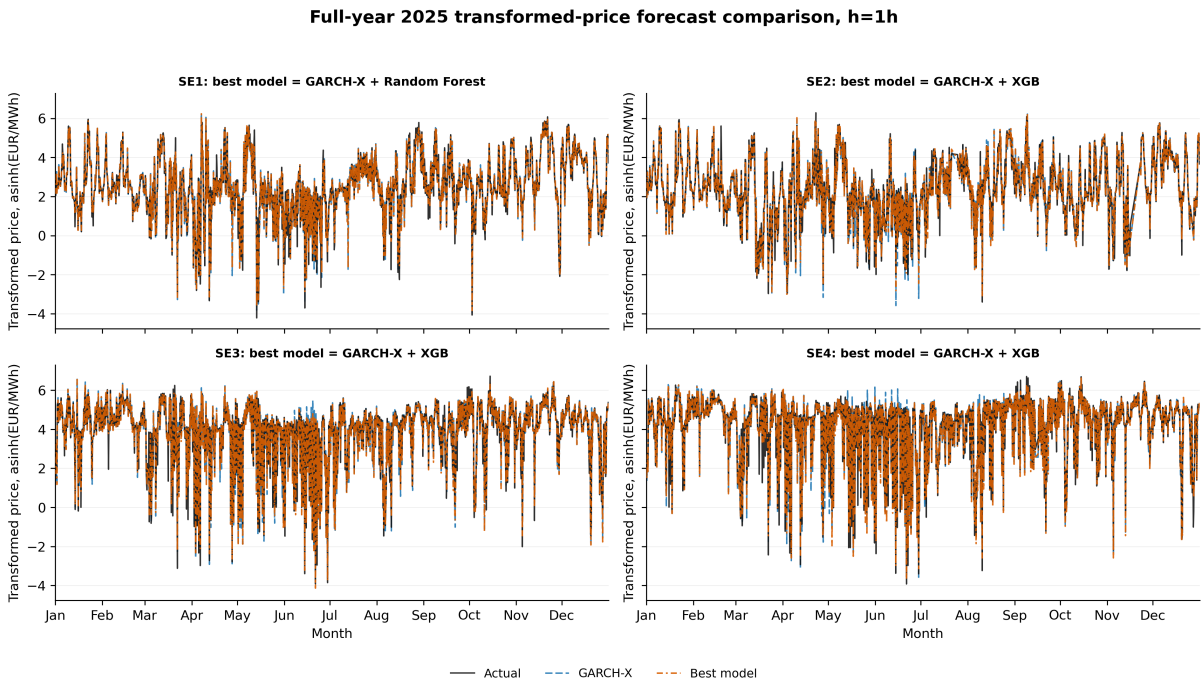


Figure 4: Out-of-sample forecast comparison at $h = 1$ under the constrained (constant) design over the full 2025 backtest period. Actual transformed prices are shown as a solid black line, the GARCH-X baseline as a blue dashed line, and the best-performing model in each zone as an orange dash-dot line.

One-hour-ahead horizon. Under the constrained (constant) design, at $h = 1$ GARCH-X still provides the dominant gain over plain GARCH(1,1), while the residual-learning layer adds a second, smaller improvement. The best residual-learning hybrid lowers transformed-space RMSE in every zone: Random Forest improves on GARCH-X by about 3.4% in SE1, while XGBoost improves on GARCH-X by about 6.3% in SE2, 8.9% in SE3, and 9.5% in SE4. All four best residual-learning comparisons are statistically significant under the Newey–West HAC Diebold–Mariano tests, with the largest short-horizon gains concentrated in SE3 and SE4 in this constrained design and sample.

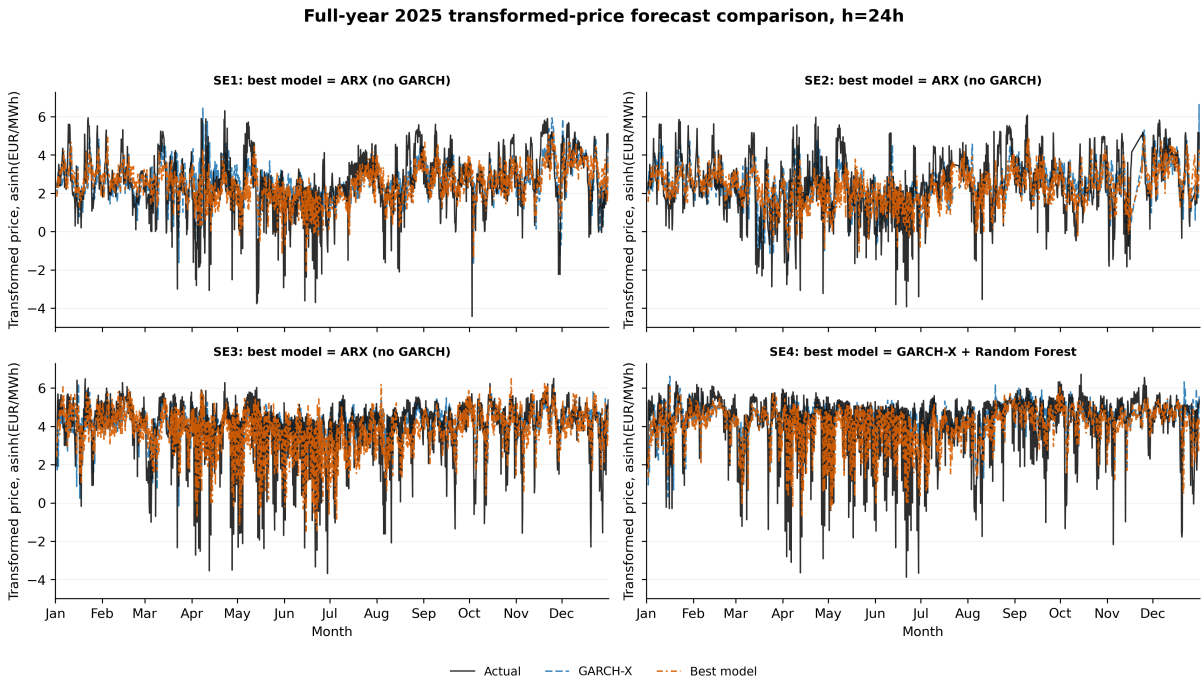


Figure 5: Out-of-sample forecast comparison at $h = 24$ under the constrained (constant) design over the full 2025 backtest period. Actual transformed prices are shown as a solid black line, the GARCH-X baseline as a blue dashed line, and the best-performing model in each zone as an orange dash-dot line. ARX is best in SE1–SE3, while Random Forest is best in SE4 in Table 10.

Twenty-four-hour-ahead horizon. At $h = 24$ under the constrained (constant) design, the ranking is more conservative but not completely uniform. Table 10 shows that ARX (no GARCH) has the lowest transformed- space RMSE in SE1 (1.351), SE2 (1.422), and SE3 (1.293), while Random Forest is best in SE4 (1.356 versus GARCH-X 1.426). This indicates that the day-ahead ranking is still driven mainly by the mean equation, with one zone-specific residual-learning exception. Under the Newey–West HAC best residual-learning comparisons in Table 12, only the SE4 Random-Forest comparison is significant at the 5% level. The broader learner-by-learner summary in Table 14 also records significant day-ahead deterioration for LSTM in SE1 and SE2, reinforcing that a residual-learning overlay is not a safe default at this horizon under the constrained design.

5.4 Residual Signal Quality for Residual-Learning Corrections

A residual-learning correction can only improve on GARCH-X if it reliably predicts the direction and rough magnitude of the baseline's errors. A correction that systematically points the wrong way, predicting a positive adjustment when the actual residual is negative, will make forecasts worse, not better. To assess this, we evaluate three complementary measures for each residual-learning hybrid: the correlation between the predicted correction and the realised GARCH-X residual, the share of sign flips (corrections pointing the wrong

way), and the share of overshoots (corrections larger in magnitude than the error they are trying to fix). These are reported in Table 11 and visualised in Figure 6.

The diagnostic logic follows the standard forecasting principle that a well-specified model should leave residuals with no exploitable autocorrelation or systematic structure, so that no further forecast-relevant information can be extracted from them (Hyndman and Athanasopoulos, 2018). In this thesis, that principle is operationalised through the three measures in Table 11: residual-correction correlation, sign flips, and overshoots.

Table 11: Residual signal quality diagnostics for residual-learning hybrids under the constrained (constant) design at the main horizons. *corr*: Pearson correlation between the predicted residual correction and the realised GARCH-X residual. *signflip_pct*: share of forecast origins where the predicted correction has the opposite sign from the actual residual (values above 50% indicate systematically misdirected corrections). *overshoot_pct*: share of origins where the correction magnitude exceeds the actual residual magnitude.

Zone	h	Model	Correlation	Overshoot (%)	Sign flip (%)	N
SE1	1	GARCH-X + LSTM	0.100	14.854	75.523	1434
SE1	24	GARCH-X + LSTM	0.034	19.762	62.579	1427
SE2	1	GARCH-X + LSTM	0.144	13.613	80.666	1381
SE2	24	GARCH-X + LSTM	0.042	28.945	58.255	1375
SE3	1	GARCH-X + LSTM	0.105	23.294	66.437	1451
SE3	24	GARCH-X + LSTM	0.138	24.395	59.986	1447
SE4	1	GARCH-X + LSTM	0.133	26.787	56.627	1441
SE4	24	GARCH-X + LSTM	0.127	31.825	61.212	1436
SE1	1	GARCH-X + Random Forest	0.289	10.460	53.347	1434
SE1	24	GARCH-X + Random Forest	0.064	18.570	56.412	1427
SE2	1	GARCH-X + Random Forest	0.317	10.065	53.584	1381
SE2	24	GARCH-X + Random Forest	0.085	17.018	58.036	1375
SE3	1	GARCH-X + Random Forest	0.304	20.538	48.312	1451
SE3	24	GARCH-X + Random Forest	0.116	28.956	58.258	1447
SE4	1	GARCH-X + Random Forest	0.342	22.415	45.177	1441
SE4	24	GARCH-X + Random Forest	0.265	27.925	60.933	1436
SE1	1	GARCH-X + XGB	0.251	17.364	45.328	1434
SE1	24	GARCH-X + XGB	0.129	22.425	48.984	1427
SE2	1	GARCH-X + XGB	0.351	19.261	45.619	1381
SE2	24	GARCH-X + XGB	0.147	21.455	48.218	1375
SE3	1	GARCH-X + XGB	0.418	24.466	43.970	1451
SE3	24	GARCH-X + XGB	0.257	32.205	51.762	1447
SE4	1	GARCH-X + XGB	0.430	27.203	41.221	1441
SE4	24	GARCH-X + XGB	0.262	32.033	53.969	1436

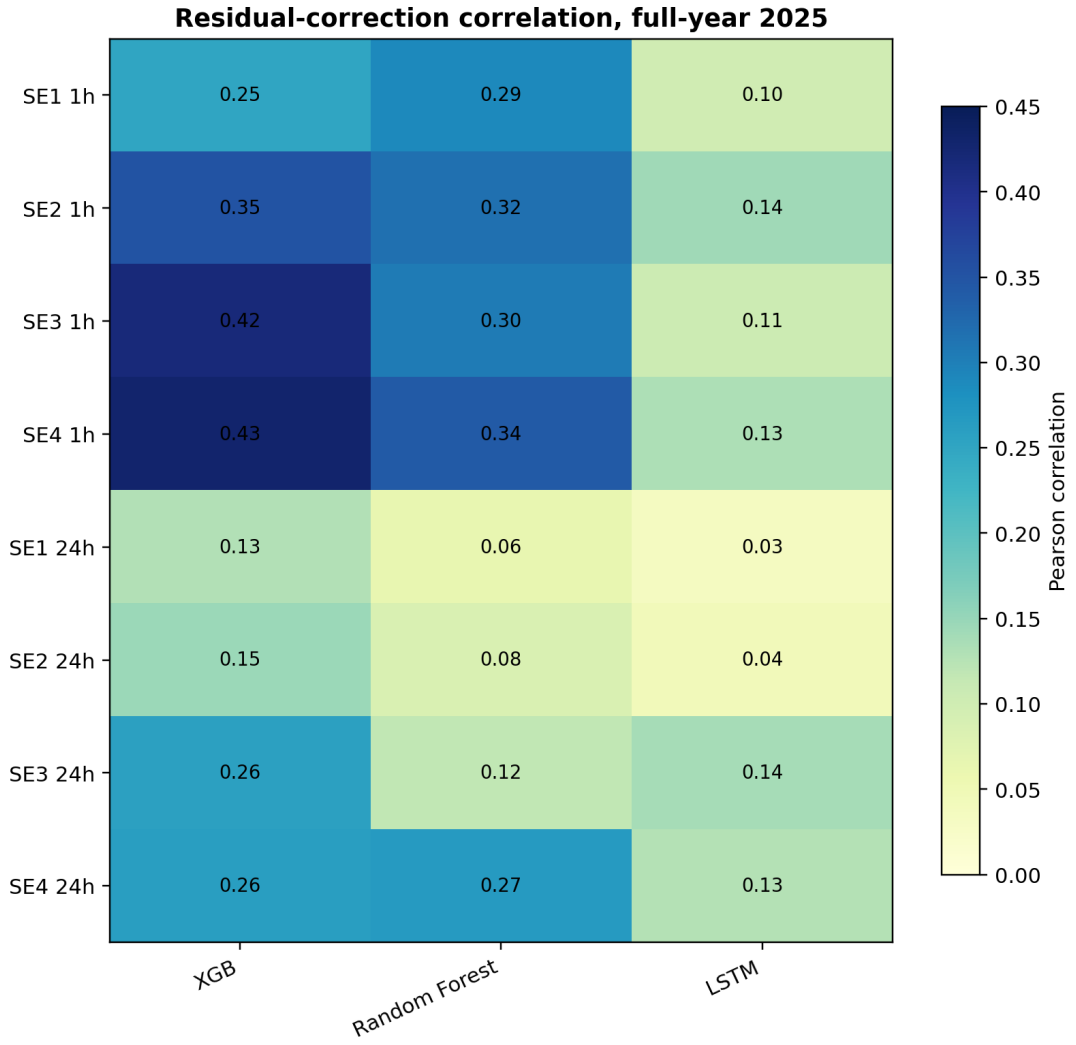


Figure 6: Residual-signal heatmap under the constrained (constant) design across zones, horizons, and residual-learning hybrids. Colour intensity represents the Pearson correlation between the predicted residual correction and the realised GARCH-X residual. Darker shading indicates stronger residual tracking; correlations near zero indicate the correction carries no exploitable signal.

Figure 6 and Appendix Figures 17–19 illustrate this visually for XGB, RF, and LSTM. At $h = 1$, meaningful residual signal appears across all four zones for the tree-based learners. XGB correlations range from about 0.25 in SE1 to 0.43 in SE4, while RF ranges from about 0.29 to 0.34. LSTM correlations are materially lower at roughly 0.10–0.14 and its sign-flip rates remain markedly higher. This hierarchy matches the forecast rankings: tree-based corrections account for every best residual-learning gain at the one-hour horizon, with the largest effects in SE3 and SE4. The Ljung–Box diagnostics in Table 7 remain useful as an in-sample diagnostic, but they should not be read as a mechanical filter for out-of-sample usefulness. They identify the clearest residual dependence in SE3 and SE4, while the full-year forecast results show that smaller residual signals in SE1 and SE2 can still be exploited by tree-based learners. At $h = 24$, residual correlations are weaker and the

forecast ranking becomes less stable: only the SE4 Random-Forest residual-learning hybrid delivers a significant best residual-learning gain, while several learner-specific comparisons remain neutral or adverse. At the main horizons, nonlinear correction value is therefore horizon- and zone-specific rather than universal.

5.5 Statistical Significance of Forecast Improvements

The previous subsection showed that residual-learning hybrids reduce RMSE relative to GARCH-X at $h = 1$ but mostly do not at $h = 24$. The natural next question is whether those RMSE differences are large enough to be real, or whether they could plausibly reflect sampling noise. A small difference in average forecast error over a single year of data could simply mean one model got lucky on the specific days in our sample. This subsection addresses that question using the Diebold–Mariano (DM) test of equal predictive accuracy (Diebold and Mariano, 1995), which compares the squared forecast errors of two models over the out-of-sample period and asks whether the difference between them is statistically distinguishable from zero. Positive DM statistics favour the residual-learning hybrid; negative ones favour GARCH-X.

Because forecast errors made close together in time are correlated rather than independent, especially at $h = 24$ where the six-hour forecast stride produces overlapping target windows, inference uses a Newey–West HAC variance estimator with a Bartlett kernel and bandwidth $h - 1$ for each horizon (Newey and West, 1987). This corrects the standard errors for the serial dependence in the loss-differential series. With forecast origins spaced six hours apart, the bandwidth is an operational approximation to the true autocorrelation structure rather than an exact finite-sample correction. More advanced joint-inference procedures, such as Model Confidence Sets or bootstrap-based comparisons, are left for future work.

Table 12 reports DM statistics and p-values for the strongest residual-learning candidate in each zone and horizon, where "strongest residual-learning hybrid" means the residual-learning model with the lowest RMSE in transformed price space, even when GARCH-X or ARX remains the best overall model in that cell. The comparison is always residual-learning hybrid against the GARCH-X baseline; this gives the residual-learning approach its best shot in every zone-horizon comparison. The main table contains eight pairwise tests; the broader summary in Table 14 extends the same logic to all 24 learner-by-horizon comparisons.

Table 12: Diebold–Mariano tests under the constrained (constant) design: best residual-learning hybrid versus GARCH-X (main horizons). Newey–West HAC variance estimation with bandwidth $h-1$ is applied to the Diebold–Mariano loss differential. Positive *dm_stat* values favour the hybrid. Reported p-values of 0.000 indicate $p < 0.001$. No multiple-testing correction is applied; with eight tests at $\alpha = 0.05$, the expected false-positive count is 0.4 under independence. The pairwise tests should therefore be read together with the economic effect sizes and residual diagnostics.

Zone	h	Best hybrid	DM stat.	p-value	N
SE1	1	GARCH-X + Random Forest	3.520	< 0.001	1434
SE2	1	GARCH-X + XGB	3.916	< 0.001	1381
SE3	1	GARCH-X + XGB	4.848	< 0.001	1451
SE4	1	GARCH-X + XGB	4.951	< 0.001	1441
SE1	24	GARCH-X + XGB	-1.137	0.256	1427
SE2	24	GARCH-X + XGB	-1.121	0.262	1375
SE3	24	GARCH-X + XGB	1.137	0.255	1447
SE4	24	GARCH-X + Random Forest	2.527	0.012	1436

The DM tests in Table 12 are based on the same out-of-sample forecast errors as Figures 9 and 10, but aggregate over the full 2025 evaluation period, whereas the figures highlight local patterns over shorter windows.

The significance pattern is sharply horizon-dependent. At $h = 1$, the residual-learning gains over GARCH-X are statistically significant in every zone, with DM statistics ranging from 3.5 in SE1 to 4.95 in SE4 (all $p < 0.001$). The strongest evidence is in SE3 and SE4, the same zones where the residual signal diagnostics in Table 11 showed the highest correlations between predicted and realised residuals. Statistical significance, effect size, and residual-structure evidence all point in the same direction at the one-hour horizon.

At $h = 24$, the result is much narrower. Only the SE4 Random-Forest comparison is significant at the 5% level. The other three zone-level comparisons cannot be distinguished from zero, with DM statistics close to zero or slightly negative. The broader summary in Table 14 also records significant *negative* day-ahead DM statistics for LSTM in SE1 and SE2, meaning that this residual-learning layer materially worsens forecast accuracy at the day-ahead horizon. The implication is not that day-ahead residual learning never works, but that it is not a reliable default across zones and model classes under the constrained design.

A caveat on multiple testing is worth stating explicitly. No formal multiple-testing correction is applied to the eight pairwise tests; with eight tests at $\alpha = 0.05$, the expected false-positive count is 0.4 under independence. The $h = 1$ results are significant by wide margins and are unlikely to reflect multiple-testing artefacts, but the sole significant $h = 24$ result (SE4 Random Forest) is more sensitive to this caveat and is correspondingly treated as a sample-specific exception throughout the discussion rather than as a general

endorsement of day-ahead residual learning. The p-values should therefore be read together with the economic effect sizes in Table 9 and the residual-signal evidence in Table 11 and Figure 6, not as stand-alone decision rules.

5.6 Model Performance under Different Volatility Regimes

Aggregate RMSE hides variation across market conditions. To examine whether residual-learning hybrids add value specifically in turbulent periods, the sample is split into low-, medium-, and high-volatility regimes using the rolling 24-hour standard deviation of y_t with the empirical 33rd and 67th percentiles as cutoffs, an approach consistent with recent energy-forecasting work comparing model accuracy across market environments (Ganczarek-Gamrot et al., 2025). Table 13 reports transformed-space RMSE by zone and regime at $h = 1$, together with the difference relative to GARCH-X within each state. Figure 7 visualises the same pattern across both horizons.

Table 13: Performance by volatility regime under the constrained (constant) design at horizon $h = 1$. Rows report bidding zone and volatility regime, while columns show transformed-space $RMSE_y$ for the GARCH-X baseline and each hybrid contender together with the difference relative to GARCH-X. Negative $\Delta RMSE_y$ indicates that the hybrid outperforms the baseline.

Zone	Regime	n	GARCH-X $RMSE_y$	XGB $RMSE_y$	RF $RMSE_y$	LSTM $RMSE_y$	ΔXGB	ΔRF	ΔLSTM
SE1	Low	499	0.322	0.306	0.311	0.320	-0.015	-0.010	-0.001
SE1	Medium	420	0.449	0.430	0.432	0.452	-0.019	-0.016	0.003
SE1	High	515	0.403	0.401	0.390	0.407	-0.002	-0.013	0.003
SE2	Low	466	0.372	0.368	0.367	0.382	-0.004	-0.004	0.010
SE2	Medium	432	0.475	0.436	0.449	0.469	-0.038	-0.025	-0.005
SE2	High	483	0.401	0.367	0.384	0.393	-0.034	-0.017	-0.008
SE3	Low	425	0.499	0.459	0.480	0.508	-0.040	-0.019	0.008
SE3	Medium	530	0.449	0.394	0.422	0.470	-0.055	-0.027	0.021
SE3	High	496	0.310	0.299	0.303	0.318	-0.011	-0.007	0.008
SE4	Low	385	0.423	0.400	0.408	0.422	-0.023	-0.015	-0.001
SE4	Medium	558	0.531	0.499	0.512	0.543	-0.032	-0.019	0.012
SE4	High	498	0.514	0.430	0.472	0.508	-0.084	-0.042	-0.006

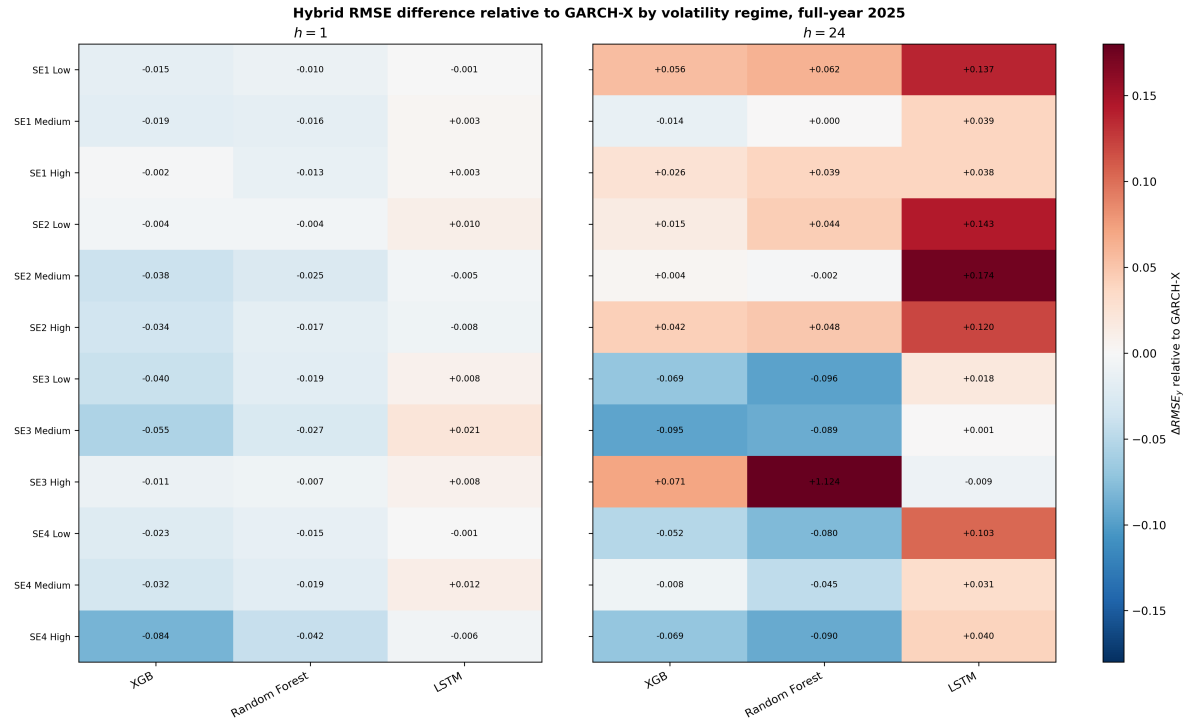


Figure 7: RMSE differences relative to the GARCH-X baseline by volatility regime (low, medium, high), bidding zone, and hybrid model at horizons $h = 1$ and $h = 24$ under the constrained (constant) information design. Negative values indicate that the hybrid model outperforms GARCH-X. To keep the full pattern visible, the colour scale is clipped at $|\Delta RMSE_y| = 0.18$; the cell annotations report the exact values.

Figure 7 and Table 13 decompose hybrid performance by volatility regime. Figure 8 condenses the corresponding full-sample percentage gains, while Figures 9–10 show how the comparison evolves through the 2025 evaluation year. Taken together, these visual summaries are consistent with the full-sample RMSE and DM test results reported in Tables 9 and 12.

The regime decomposition sharpens the full-year short-horizon pattern. At $h = 1$, tree-based hybrids improve on GARCH-X in multiple volatility states in every zone, but the largest reductions remain concentrated in SE3 and SE4. XGBoost is especially strong in SE3's medium-volatility regime and SE4's high-volatility regime: the transformed-space RMSE reductions are 0.055 in SE3-medium (about 12.2% relative to GARCH-X) and 0.084 in SE4-high (about 16.3%). RF also contributes positive gains in several states, with its largest one-hour reduction in SE4-high (0.042, about 8.2%). At $h = 24$, the picture is more fragmented: improvements are smaller, more regime-specific, and less consistent across learners, although the SE4 Random-Forest result remains visible. The regime evidence therefore supports a qualified interpretation: residual learning is useful most clearly at the one-hour-ahead horizon, with the strength of that benefit varying by zone and market state rather than following a simple binary split.

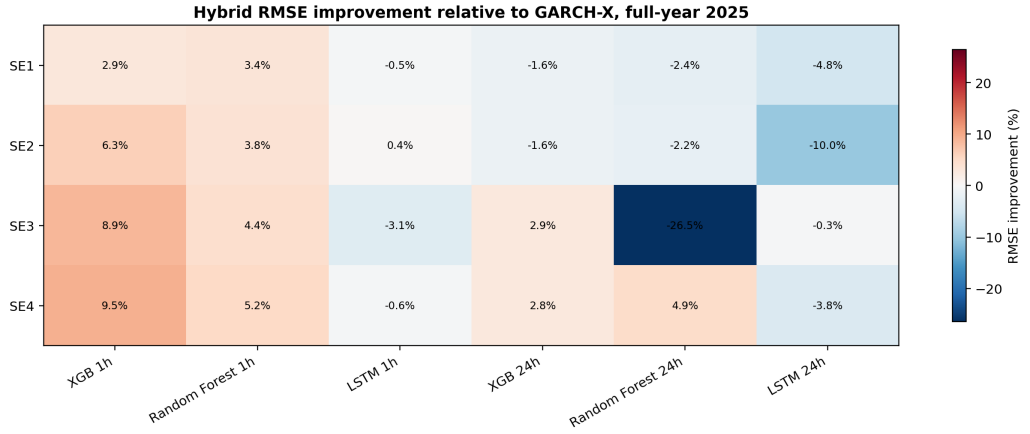


Figure 8: Relative RMSE improvement of residual-learning hybrids over GARCH-X under the constrained (constant) design across zones and horizons over the full 2025 evaluation period.

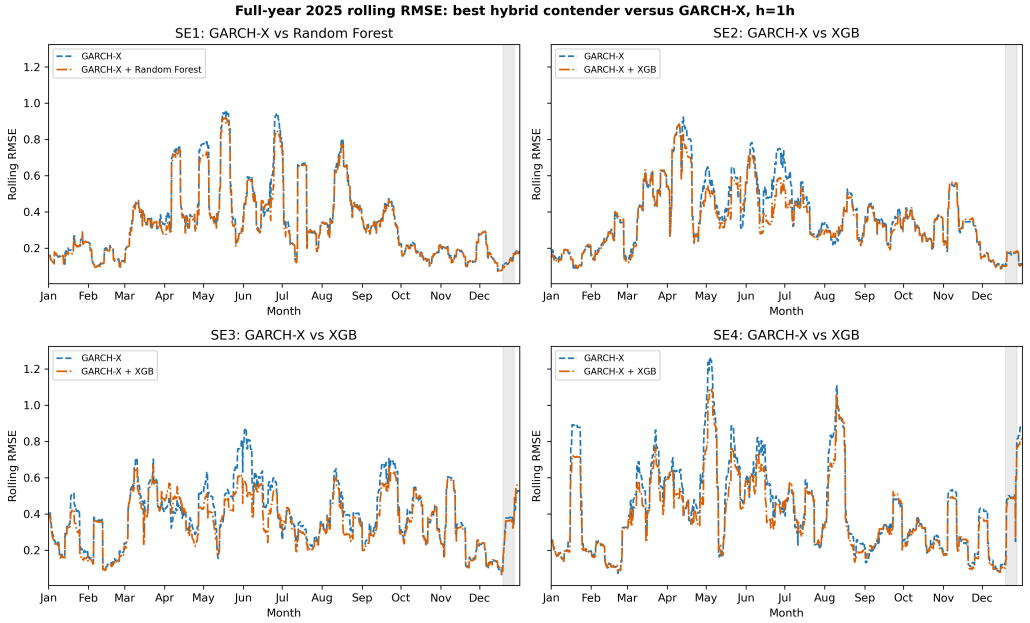


Figure 9: Rolling RMSE under the constrained (constant) design over the full 2025 evaluation period at $h = 1$: best residual-learning contender versus GARCH-X. For visual clarity, the GARCH-X baseline is shown as a blue dashed line and the best residual-learning contender as an orange dash-dot line in each zone; full time-series comparisons for GARCH-X and all hybrid contenders are reported in Appendix B.

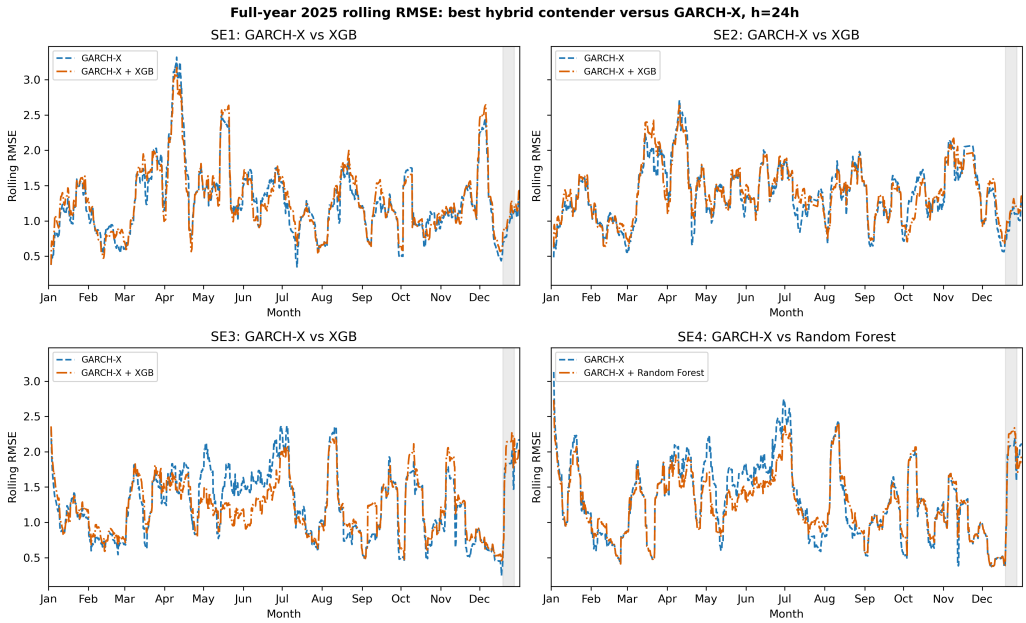


Figure 10: Rolling RMSE under the constrained (constant) design over the full 2025 evaluation period at $h = 24$: best residual-learning contender versus GARCH-X. For visual clarity, the GARCH-X baseline is shown as a blue dashed line and the best residual-learning contender as an orange dash-dot line in each zone; full time-series comparisons for GARCH-X and all hybrid contenders are reported in Appendix B.

5.7 Event Window: Late December 2025

The late-December 2025 episode is used as an illustrative stress window for the residual-correction failure mode. The full-period evidence in Table 9, Table 12, and Figures 4–5 remains the basis for the average performance results, while the event window shows how the models behave during an abrupt regime change.

Within this window, GARCH-X remains imperfect but is generally more stable than the residual-learning corrections. The largest residual-learning degradations arise when the residual learner adds a correction with the wrong sign or with excessive magnitude relative to the baseline. In those cases, the corrected forecast overshoots the realised price move rather than dampening the baseline error. This pattern is shown most clearly in Appendix Figure 12.

At $h = 1$, the instability appears as occasional but economically meaningful overshooting during the negative-price spell. At $h = 24$, the same mechanism interacts with multi-step uncertainty in future fundamentals, which is consistent with the weaker case for residual-learning correction at the day-ahead horizon. The point is not that all large errors occur inside this window, but that this episode reveals the residual-correction failure mode in a compact and interpretable setting.

For RQ3, this window matters because it speaks to robustness and interpretability rather

than to average accuracy. A transparent baseline that misses the level by a moderate amount is easier to diagnose than a hybrid correction that amplifies the error through an opaque residual adjustment. The event window therefore illustrates why hybrid complexity is justified only when residual diagnostics indicate stable, learnable structure beyond the econometric baseline under the maintained information set.

5.8 Scope of Alternative Exogenous-Information Designs

Section 4.7 documents two alternative future-path designs for non-deterministic exogenous variables: a seasonal-naive path and an oracle path. These designs clarify the information-set problem but are not used for the thesis's quantitative model rankings. The seasonal-naive design is operationally feasible, while the oracle design is infeasible at the forecast origin and serves only as a conceptual upper bound. During implementation, the alternative-design exports were useful as stress tests, but their price-space diagnostics were numerically unstable after rolling standardisation and inverse arcsinh transformation. For that reason, no model ranking or headline conclusion is drawn from them. All reported forecast comparisons, statistical tests, and practical recommendations should therefore be interpreted as conditional on the constrained carry-forward information design.

6 Discussion

Table 14: Formal hypothesis-test summary for the main horizons (1h and 24h).

Hypothesis	Zone	h	Model A	Model B	$RMSE_A$	$RMSE_B$	DM stat.	p-value	N
H2	SE1	1	GARCH-X	GARCH-X + XGB	0.392	0.380	1.499	0.134	1434
H2	SE1	24	GARCH-X	GARCH-X + XGB	1.370	1.392	-1.137	0.256	1427
H2	SE2	1	GARCH-X	GARCH-X + XGB	0.417	0.390	3.916	< 0.001	1381
H2	SE2	24	GARCH-X	GARCH-X + XGB	1.428	1.450	-1.121	0.262	1375
H2	SE3	1	GARCH-X	GARCH-X + XGB	0.424	0.386	4.848	< 0.001	1451
H2	SE3	24	GARCH-X	GARCH-X + XGB	1.333	1.295	1.137	0.255	1447
H2	SE4	1	GARCH-X	GARCH-X + XGB	0.498	0.451	4.951	< 0.001	1441
H2	SE4	24	GARCH-X	GARCH-X + XGB	1.426	1.385	1.037	0.300	1436
H3	SE1	1	GARCH-X	GARCH-X + Random Forest	0.392	0.378	3.520	< 0.001	1434
H3	SE1	24	GARCH-X	GARCH-X + Random Forest	1.370	1.403	-1.826	0.068	1427
H3	SE2	1	GARCH-X	GARCH-X + Random Forest	0.417	0.400	3.896	< 0.001	1381
H3	SE2	24	GARCH-X	GARCH-X + Random Forest	1.428	1.459	-1.565	0.117	1375
H3	SE3	1	GARCH-X	GARCH-X + Random Forest	0.424	0.405	3.931	< 0.001	1451
H3	SE3	24	GARCH-X	GARCH-X + Random Forest	1.333	1.686	-0.872	0.383	1447
H3	SE4	1	GARCH-X	GARCH-X + Random Forest	0.498	0.472	4.886	< 0.001	1441
H3	SE4	24	GARCH-X	GARCH-X + Random Forest	1.426	1.356	2.527	0.012	1436
H4	SE1	1	GARCH-X	GARCH-X + LSTM	0.392	0.394	-0.480	0.631	1434
H4	SE1	24	GARCH-X	GARCH-X + LSTM	1.370	1.436	-2.417	0.016	1427
H4	SE2	1	GARCH-X	GARCH-X + LSTM	0.417	0.415	0.340	0.734	1381
H4	SE2	24	GARCH-X	GARCH-X + LSTM	1.428	1.570	-3.168	0.002	1375
H4	SE3	1	GARCH-X	GARCH-X + LSTM	0.424	0.437	-2.391	0.017	1451
H4	SE3	24	GARCH-X	GARCH-X + LSTM	1.333	1.337	-0.200	0.841	1447
H4	SE4	1	GARCH-X	GARCH-X + LSTM	0.498	0.501	-0.443	0.658	1441
H4	SE4	24	GARCH-X	GARCH-X + LSTM	1.426	1.481	-1.592	0.111	1436

Unless otherwise noted, the discussion refers to the full-year 2025 main results at $h = 1$ and $h = 24$ under the constrained (constant) design. Appendix B provides supporting event-window and rolling-RMSE figures, and Appendix C documents alternative information-set assumptions.

6.1 RQ1: Can Machine Learning Methods Complement GARCH-X?

The answer is yes, but only in horizon- and zone-specific ways. At $h = 1$, the best residual-learning hybrid improves on GARCH-X in every zone in the full-year backtest: Random Forest in SE1 and XGBoost in SE2–SE4. The largest reductions are in SE3 and SE4, where the residual diagnostics also point to the strongest remaining structure after the GARCH-X filter. At $h = 24$, the conclusion is different in this constrained (constant) exogenous-information design. ARX is best in SE1–SE3, Random Forest is best in SE4, and only the SE4 day-ahead residual-learning gain is statistically significant in the best residual-learning Diebold–Mariano comparison in this sample. Residual-learning gains therefore do not support a general day-ahead residual-learning default under the constrained design. No formal multiple-testing correction is applied to these eight tests, so the p-values are interpreted jointly with economic effect sizes and residual diagnostics rather than as stand-alone decision rules.

6.2 RQ2: Cross-Zone Differences in Price Dynamics

The main cross-zone difference is not that the volatility model works in fundamentally different ways from north to south. The more important difference is how much structure remains in the residuals after GARCH-X has filtered the series. One identification caveat is important: because generation-mix shares are measured nationally and therefore enter identically for SE1–SE4, the cross-zone mechanism evidence is identified primarily from zone-specific load, net-flow, and residual-structure variation rather than from directly estimated zonal generation-mix differences. Tables 7, 11, and 13 show a graded pattern: SE3 and SE4 retain the strongest residual structure and deliver the largest one-hour residual-learning gains, while SE1 and SE2 show smaller but still statistically significant short-horizon improvements. The selected lag structures point in the same direction: all zones depend on recent hours, but the southern zones require denser short-run and intraday timing structure.

As in the literature discussion, evidence on generation-mix mechanisms is used here as external market context rather than as a direct estimate from the thesis's explanatory matrix, because the generation-share variables enter at the national rather than the zone-specific level.

The diagnostic pattern and the forecast pattern are mutually consistent rather than independent findings. Table 7 remains an in-sample standardised-residual diagnostic, not a mechanical rule for excluding hybrids. It nevertheless identifies SE3 and SE4 as the zones with the clearest remaining autocorrelation after GARCH-X filtering, and the same zones are also the ones in which the GJR diagnostic in Section 5.1 indicates economically meaningful asymmetric leverage. The full-year forecast results then refine this picture: residual-learning value is not confined exclusively to the southern zones, but it is strongest where the diagnostics and effect sizes align.

One interpretation is that part of the short-horizon residual-learning gain in SE3 and SE4 reflects nonlinear or asymmetric price responses that the symmetric GARCH-X baseline does not model directly. This is why a GJR-X or EGARCH-X baseline inside the same residual-learning pipeline is a natural robustness extension, rather than a separate research question.

Figure 11 sheds further light on the cross-zone mechanism by decomposing XGBoost feature importance into economic channels.

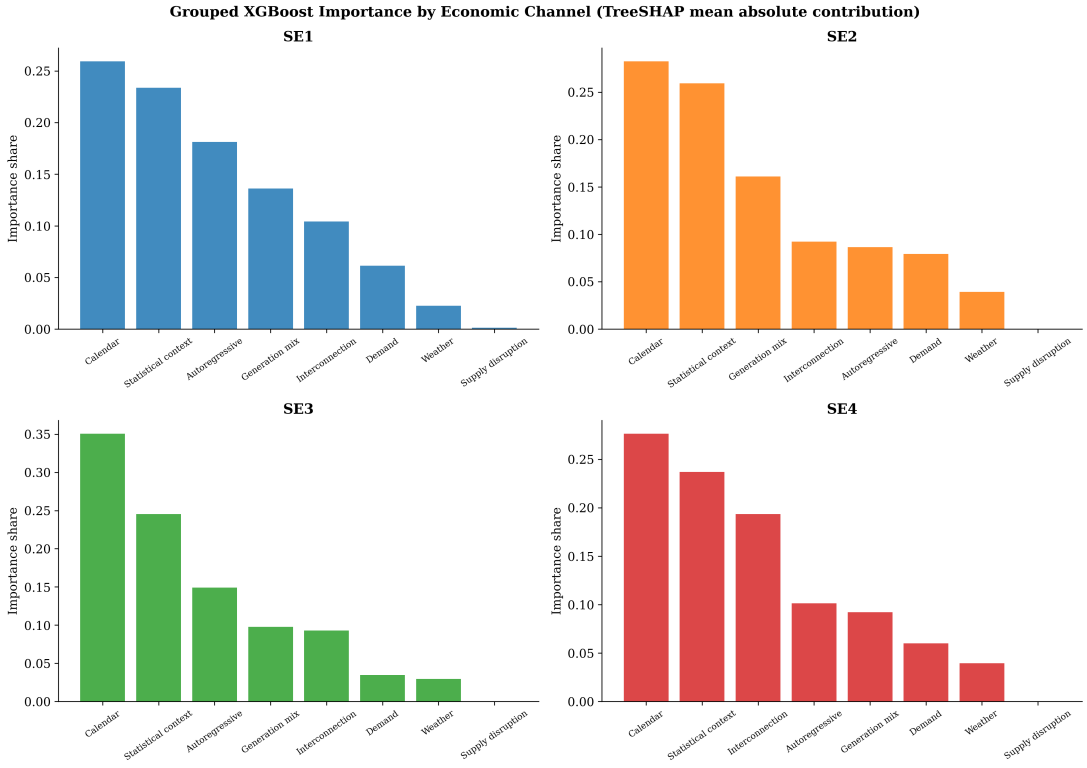


Figure 11: XGBoost residual-correction feature importance by economic channel and bidding zone ($h = 1$). Importance is measured using native TreeSHAP mean absolute contributions from the XGBoost residual corrector and then aggregated from individual features into economic channels: Calendar, Generation mix, Interconnection, Autoregressive, Statistical context, Demand, Weather, and Supply disruption where present. Shares sum to one within each zone.

To interpret what drives the XGBoost residual corrections, we use TreeSHAP, which estimates how much each input variable contributed to each individual prediction the model made and then averages those contributions across the sample. This gives a more reliable picture of which variables actually matter than the simpler "gain-based" importance measure, which tends to overstate the role of features with many possible values.

Figure 11 groups the resulting feature contributions into economically interpretable channels. Calendar features (hour-of-day, day-of-week) account for the largest share in every zone at roughly 26–35%, which is consistent with the strong intraday and weekly seasonality of electricity demand. Statistical context, meaning the fitted GARCH-X mean and conditional volatility passed to the residual learner as features, contributes another 23–26% across zones. This confirms that the machine-learning correction is building on the econometric baseline rather than operating independently of it. Generation mix contributes about 9–16%, and interconnection features are most visible in SE4, where cross-border exposure is structurally most relevant.

6.3 RQ3: Robustness and Interpretability

Under the thesis's main information set, the most robust and interpretable choices depend on the horizon. Tables 9, 12, and 13 show that residual-learning overlays have a clear role at $h = 1$, especially in SE3 and SE4, but that transparent mean-equation models remain the safer default at $h = 24$ except for SE4. The event-window evidence in Appendix B illustrates that residual-learning corrections can become unstable when the residual mapping changes abruptly. Appendix C also underscores that model performance is jointly determined by the information set used for future exogenous variables, not by model class alone. Robustness in electricity price forecasting is therefore a joint property of model structure, residual signal, and operational information availability. A more complex model is credible only when the extra layer improves forecast accuracy in a stable and interpretable way under the maintained information set.

Taken over the full 2025 evaluation year under the constrained design, the cross-zone pattern is therefore graded rather than binary. At $h = 1$, the largest and most robust residual-learning gains remain in SE3 and SE4, while SE1 and SE2 also show smaller but statistically significant improvements under the constrained design. At $h = 24$, gains are limited overall and mainly concentrated in SE4, where the Random Forest residual-learning hybrid provides a statistically significant improvement in this sample. The evidence supports a stronger short-horizon residual-learning case in the southern zones, but not a strict claim that residual-learning overlays help only there.

6.4 Practical Implications

The results have different implications depending on which forecasting decision is at stake.

For producers, retailers, and market participants monitoring short-horizon movements in the hourly day-ahead price series, a residual-learning overlay is justified most clearly at h=1 in this constrained design. The full-year results show statistically significant best residual-learning improvements in all four zones, with the strongest gains in SE3 and SE4. The operational lesson is not that every Swedish zone requires the same model stack, but that the case for a residual-correction layer is strongest where diagnostics and effect sizes align. This matters because short-horizon errors in the expected day-ahead price profile can affect position management, balancing expectations, and risk monitoring, even though the thesis does not directly model intraday transaction prices.

For day-ahead planning, the results favour transparent mean-equation models over a general residual-learning overlay. ARX is best in SE1–SE3, and only SE4 shows a statistically significant Random-Forest gain over GARCH-X at $h = 24$ in this constrained design and sample. A simpler model is also easier to monitor, audit, and explain to counterparties or management when a forecast turns out to be wrong. This matters because day-ahead bidding decisions involve contract obligations and financial exposure; a model that degrades under stress, as the late-December 2025 event window illustrates, carries operational risk that can outweigh a small average-case improvement.

For system operators and policymakers, the north–south asymmetry is worth attention but should be read as a gradient rather than a binary split. The largest one-hour residual-learning gains occur in SE3 and SE4, consistent with stronger cross-border exposure and less predictable residual patterns in the south, but smaller gains also appear in SE1 and SE2. That distinction matters for how modelling resources and risk-management tools are allocated across the Swedish transmission system.

To make the horizon- and zone-specific recommendations operational, Table 15 summarises the preferred models under the constrained (constant) design for the main horizons in each bidding zone.

6.5 Limitations and External Validity

Three design choices bound the external validity of the results. First, the main evaluation covers the 2025 calendar year. This is broader than a single winter quarter, but it is still one market year, so cross-year stability remains untested. The results could differ in calmer years or during stress episodes comparable to the 2021–2022 energy crisis. Second, the main specification uses a constrained exogenous-information design and a common 90-day rolling window; different information assumptions or adaptive windows could change

Table 15: Practical model guidance by zone and horizon under the constrained (constant) design.

Zone–horizon	Recommended model	Comment
SE1, $h = 1$	GARCH-X + Random Forest	Small but significant gain over GARCH-X; baseline already strong
SE2, $h = 1$	GARCH-X + XGBoost	Modest, significant hybrid improvement; limited residual structure
SE3, $h = 1$	GARCH-X + XGBoost	Largest short-horizon gains; strongest residual signal
SE4, $h = 1$	GARCH-X + XGBoost	Large and robust gains in medium/high-volatility regimes
SE1–SE3, $h = 24$	ARX (no GARCH)	No robust hybrid advantage; gains driven by mean specification
SE4, $h = 24$	GARCH-X + Random Forest	Useful day-ahead exception; residual-learning gain significant in this sample

rankings, especially in the more volatile zones. Third, transformed-space RMSE is the primary ranking criterion, while price-space diagnostics remain supplementary because the inverse transformation can amplify rare extreme errors. Because all main accuracy metrics are reported in arcsinh-transformed space, the practical magnitude of a stated percentage improvement depends on the price level at which it is evaluated: the inverse transformation $p_t = \sinh(y_t)$ has gradient $\cosh(y_t)$, so a 0.02 unit reduction in transformed-space RMSE corresponds to roughly 1–2 EUR/MWh near a representative price of 60 EUR/MWh and to materially more during spike events where $|y_t|$ is large. Reported percentage gains should therefore be read as transformed-space quantities, not as direct EUR/MWh accuracy improvements. These choices limit how far the results travel, and the main conclusions should therefore be read as conditional on the constrained information design, but they do not weaken the within-design comparisons.

External validity is also limited by the single-market setting: all evidence comes from Swedish bidding zones, so the north–south pattern documented here should not be generalised to markets with different interconnection structures, generation mixes, or market designs without further testing. A natural extension, beyond the scope of this thesis, would be to repeat the analysis in at least one additional bidding zone, for example NO4 in the Norwegian system or a German bidding zone, to test whether the SE3/SE4 residual-learning advantage at $h = 1$ generalises to markets with different generation mixes and interconnection structures. The cross-market benchmarking principles emphasised by Lago et al. (2021) would support such an extension, but the present thesis confines itself to Swedish data so that the within-market comparison across SE1–SE4 remains the unit of analysis.

The thesis is not an exhaustive model-selection exercise. It deliberately holds the inten-

tionally narrow model set of GARCH-X plus three residual learners and a conservative tuning strategy fixed to keep the full-year comparisons interpretable across zones and horizons and to focus on an architecture test rather than optimisation of any single learner. Broader searches over model classes, hyperparameters, and window lengths are natural robustness exercises, but they would answer a different question from the operational comparison studied here.

The ARX (no GARCH) ablation separates the structural mean equation from the conditional-variance component, but it does not isolate pure autoregressive price persistence because it retains the same exogenous fundamentals as GARCH-X. A pure AR-only benchmark would be a useful further decomposition. It was left outside the final model set to keep the comparison focused on operationally plausible structural forecasts rather than a full ablation tree.

A further limitation specific to the LSTM is that its network weights are held fixed between weekly retraining points, while the GARCH-X baseline and the tree-based corrections update at every forecast origin. This means the LSTM operates with staler parameter estimates at short horizons and the comparison is not perfectly symmetric. The consequence is likely a downward bias in reported LSTM performance at $h = 1$, where frequent re-estimation is likely most valuable.

The explanatory data also impose limits on the interpretation of cross-zone mechanisms. Load and net-flow variables vary by bidding zone, but generation shares are measured at the national level and therefore enter identically for SE1–SE4. As a result, statements about hydro dominance in the north or greater renewables exposure in the south should be read as institutional context rather than as variation directly identified from the generation-share regressors. The TreeSHAP channel decomposition reduces the bias of gain-based feature importance, but it remains a model-explanation device rather than causal evidence about the structural drivers of Swedish electricity prices.

Finally, the formal inference remains pairwise and conditional on the forecast design. The Newey–West HAC adjustment addresses serial dependence in the loss-differential series, including the overlapping forecast windows induced by the 6-hour stride, but it remains a pairwise comparison. No formal multiple-testing correction is applied to these eight tests, so the p-values are interpreted jointly with economic effect sizes and residual diagnostics rather than as stand-alone decision rules. A stationary block bootstrap, Romano–Wolf step-down procedure, or formal Model Confidence Set would be the cleaner way to assess joint model superiority in a larger extension.

7 Conclusion

The central question of the thesis is whether machine-learning residual corrections add forecast value on top of a carefully specified GARCH-X baseline in Swedish bidding zones under a constrained (constant) exogenous-information design. The answer is that they sometimes do, but the conditions are more specific than a general argument for hybrid modelling would suggest. Moving from a plain GARCH(1,1) benchmark to GARCH-X, with autoregressive lags and exogenous fundamentals in the mean equation, is the largest and most consistent improvement. At $h = 24$ in this constrained (constant) exogenous-information design, the ARX ablation then shows that much of the day-ahead forecast value comes from the mean equation itself: ARX is best in SE1–SE3, while only SE4 supports a statistically significant Random-Forest residual correction. At $h = 1$, by contrast, residual learning has a clearer role, with the best residual-learning hybrid improving on GARCH-X in every zone and the largest gains in SE3 and SE4.

For RQ1, the implication is that residual-learning corrections are a selective complement to GARCH-X in this constrained design, not a universal improvement. For RQ2, the strongest residual-learning gains in SE3 and SE4 are consistent with structural differences between zones, especially stronger cross-border exposure and residual patterns that are harder to capture with a linear model, but the full-year evidence shows that the zone split is graded rather than absolute. Claims about generation-mix differences remain contextual rather than directly identified from the regressors, because the generation-share variables enter at the national rather than the zone-specific level. For RQ3, the preferred operational choice depends on the horizon: residual learning is most defensible for one-hour-ahead forecasts of the hourly day-ahead price series under the constrained information set. While transparent mean-equation models remain the day-ahead default across zones in this design, with SE4 as a useful but sample-specific exception.

These findings are tied to a specific empirical design: a common 90-day rolling estimation window, a full-year 2025 out-of-sample evaluation period for the main horizons, and a constrained treatment of future exogenous variables. Within that design, the case for machine-learning residual correction is strongest where the baseline still leaves forecastable residual patterns and where the forecast does not rely on unrealistically rich knowledge of future fundamentals. No formal multiple-testing correction is applied to these eight tests, so the p-values are interpreted jointly with economic effect sizes and residual diagnostics rather than as stand-alone decision rules.

The practical implication is straightforward: transparent mean-equation models should be the default, and residual-learning layers should be reserved for settings where diagnostics show stable, learnable residual structure under the maintained information set. In this thesis, that case is strongest for one-hour forecasts in SE3 and SE4, with smaller but

statistically significant one-hour gains also present in SE1 and SE2. Day-ahead forecasting under the constrained design does not support a general residual-learning default, despite the SE4 Random-Forest result. Table 15 in Section 6.4 summarises these horizon- and zone-specific recommendations under the constrained design and is intended as a concise reference for practitioners. The most natural extensions are multi-year out-of-sample tests, evaluation in other Nordic or European markets, and experiments with shorter or adaptive estimation windows, particularly for the more volatile zones. The thesis therefore supports selective residual-learning modelling, not residual-learning modelling as a general rule.

8 References

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY. ACM.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–265.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

ENTSO-E (2026). Electricity market transparency. Accessed 22 April 2026.

European Commission (2013). Commission regulation (eu) no 543/2013 of 14 june 2013 on submission and publication of data in electricity markets and amending annex i to regulation (ec) no 714/2009 of the european parliament and of the council. Official Journal of the European Union, accessed 22 April 2026.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Ganczarek-Gamrot, A., Gorczyca-Goraj, A., Pilot, K., and Kania, K. (2025). Electricity price volatility and the performance of machine learning forecasting models in European energy markets. *Energies*, 18(24):6535.

Han, H. and Kristensen, D. (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. *Journal of Business & Economic Statistics*, 32(3):416–429.

Henje Scott, O. and Mellander, W. (2025). Scenario-based long-term electricity price forecasting in NO4 (Norway) using a hybrid machine learning model. Student thesis, KTH Royal Institute of Technology.

Hickey, E., Loomis, D. G., and Mohammadi, H. (2012). Forecasting hourly electricity prices using ARMAX–GARCH models: An application to MISO hubs. *Energy Economics*, 34(1):307–315.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, 2nd edition.

Koopman, S. J., Ooms, M., and Carnero, M. A. (2007). Periodic seasonal Reg-ARFIMA–GARCH models for daily electricity spot prices. *Journal of the American Statistical Association*, 102(477):16–27.

Lago, J., Marcjasz, G., De Schutter, B., and Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1–2):499–526.

Marcjasz, G., Serafin, T., and Weron, R. (2018). Selection of calibration windows for day-ahead electricity price forecasting. *Energies*, 11(9):2364.

Misiorek, A., Trück, S., and Weron, R. (2006). Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models. *Studies in Nonlinear Dynamics & Econometrics*, 10(3):1–36.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Nowotarski, J. and Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81(1):1548–1568.

SMHI (2026). Smhi open data. Accessed 22 April 2026.

Statistics Sweden (2026). Population statistics. Accessed 22 April 2026.

Uniejewski, B., Weron, R., and Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, 33(2):2219–2229.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081.

Weron, R. and Misiorek, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, 24(4):744–763.

Ziel, F. and Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70:396–420.

9 Appendix

Appendix A: Selected Lag Structure by Bidding Zone

A.1 Lag Selection Algorithm

The purpose of the lag-selection procedure is to obtain compact but informative zone-specific lag sets for both the autoregressive price terms and the exogenous regressors. The objective is not to maximise in-sample fit mechanically, but to retain a sparse specification that captures short-run and intraday dependence without overfitting the rolling estimation windows.

The procedure is applied separately to each bidding zone on the curated modelling dataset. First, candidate autoregressive lags for the transformed price series are screened using ACF and PACF evidence, with recent and seasonal anchor lags retained when they show statistically meaningful dependence. Second, candidate lags for each exogenous variable are screened using prewhitened cross-correlation analysis so that apparent lead-lag relations are not driven by shared persistence alone. Third, a lagged design matrix is constructed from the screened endogenous and exogenous candidates, while deterministic calendar controls are always retained. Fourth, sparse stepwise selection is run under three information criteria, AIC, BIC, and HQIC. Finally, a lag is included in the recommended specification if it is selected by at least two of the three criteria.

In pseudo-code, the algorithm is as follows: screen plausible target lags; screen plausible exogenous lags variable by variable; build a joint sparse design; estimate competing stepwise specifications under AIC, BIC, and HQIC; and retain only those lag terms that survive majority voting across the three criteria. The resulting recommendation is therefore data-driven, zone-specific, and intentionally conservative.

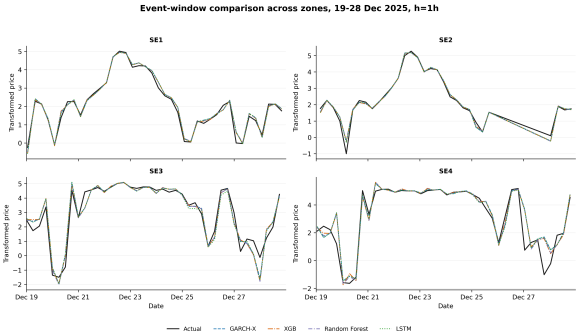
Table 16: Selected lag structure by bidding zone (full specification export).

Zone	Variable	Type	Lags
SE1	y	endog	1,3,16,23,27
SE1	domestic_flow_mw	exog	1,2
SE1	load_mw	exog	1,144
SE1	outage_planned_mw	exog	39,41
SE1	share_hydro	exog	1,5
SE1	share_wind	exog	1
SE2	y	endog	1,2,3
SE2	load_mw	exog	1,2,71
SE2	share_hydro	exog	5
SE2	share_nuclear	exog	1,5
SE2	share_wind	exog	1,3
SE3	y	endog	1,2,3,5,16,23,26
SE3	domestic_flow_mw	exog	1
SE3	foreign_flow_mw	exog	96
SE3	load_mw	exog	167,168
SE3	outage_planned_mw	exog	136
SE3	share_wind	exog	1
SE4	y	endog	1,2,3,8,13,23,26
SE4	domestic_flow_mw	exog	1,2
SE4	foreign_flow_mw	exog	134
SE4	load_mw	exog	47,65,168
SE4	outage_planned_mw	exog	64
SE4	share_wind	exog	1
SE4	temp_celsius	exog	95

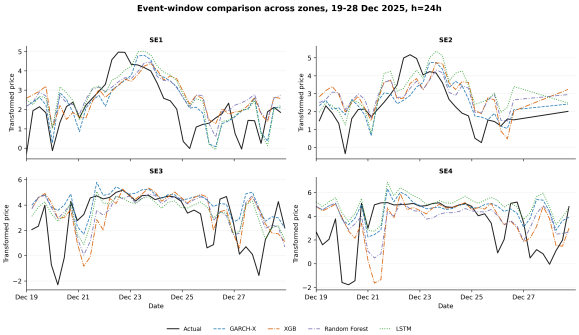
Appendix B: Supporting Diagnostics under the Constrained (Constant) Design

All tables and figures in this appendix use the constrained (constant) design that underlies Section 5. The figures correspond to the full-year 2025 main evaluation at $h = 1$ and $h = 24$. The seasonal-naive design and the oracle benchmark are documented separately in Appendix C and are not mixed into the artifacts below.

B.1 Event-Window Support Figures



(a) Event-window zoom at $h = 1$



(b) Event-window zoom at $h = 24$

Figure 12: Supporting event-window forecast comparison under the constrained (constant) design at $h = 1$ and $h = 24$. The highlighted interval, 19–28 December 2025, is the most pronounced multi-zone negative-price spell in the full-year backtest. Line types distinguish actual prices, GARCH-X, XGBoost, Random Forest, and LSTM within each panel. Periods where the hybrid pushes the forecast further from realised prices illustrate the regime-shift failure mode.

B.2 Full-Period Rolling RMSE for GARCH-X and Hybrid Contenders

The figures below report the full time-series rolling RMSE comparisons behind the compressed main-text plots. They use the same full-year 2025 out-of-sample forecast errors as Tables 9 and 12, but show all hybrid contenders and the GARCH-X baseline simultaneously in each zone.

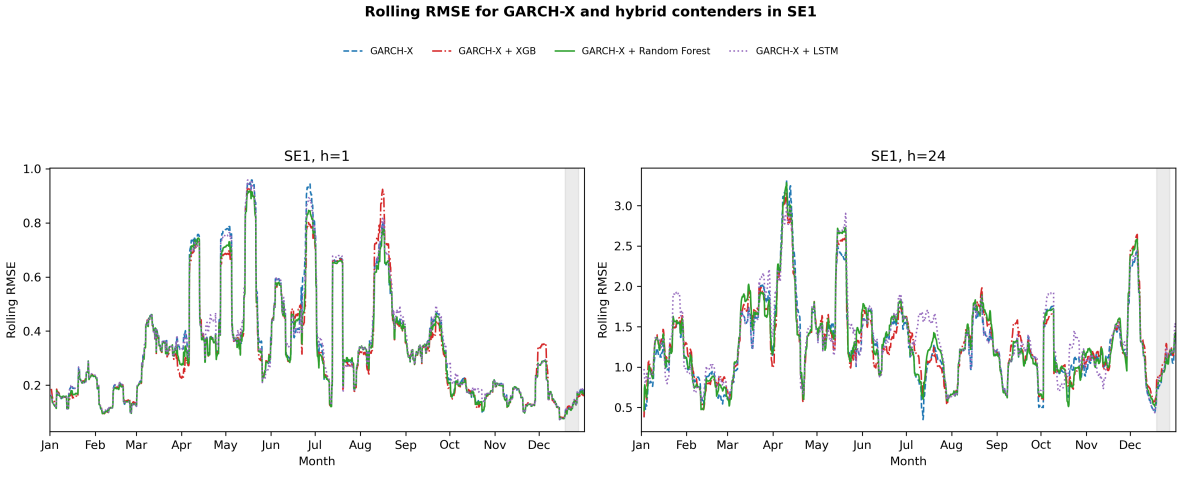


Figure 13: Rolling RMSE for GARCH-X and hybrid contenders in zone SE1 at horizons $h = 1$ and $h = 24$ under the constrained information design. The figure corresponds to the full-year 2025 out-of-sample evaluation window and uses distinct line styles for GARCH-X, XGBoost, Random Forest, and LSTM.

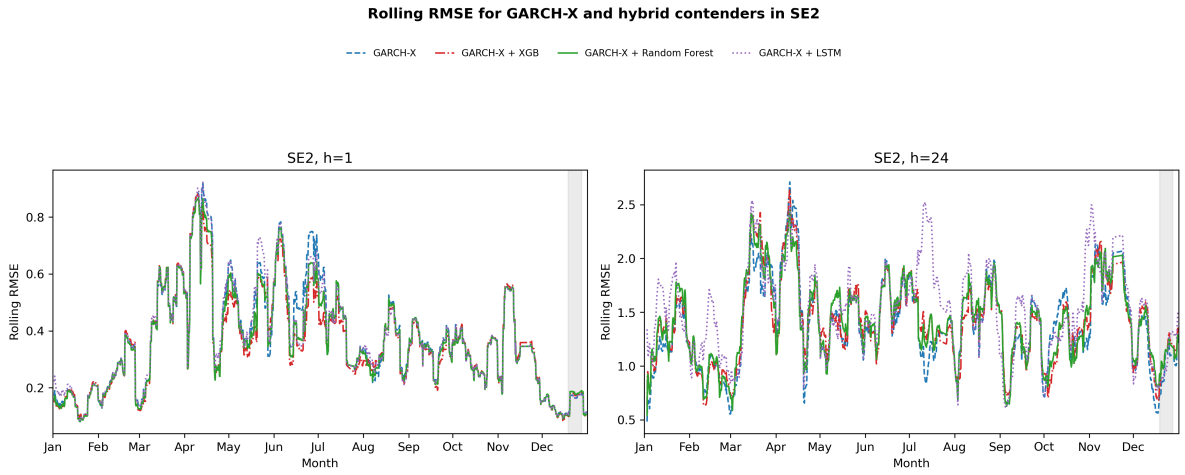


Figure 14: Rolling RMSE for GARCH-X and hybrid contenders in zone SE2 at horizons $h = 1$ and $h = 24$ under the constrained information design. The figure corresponds to the full-year 2025 out-of-sample evaluation window and uses distinct line styles for GARCH-X, XGBoost, Random Forest, and LSTM.

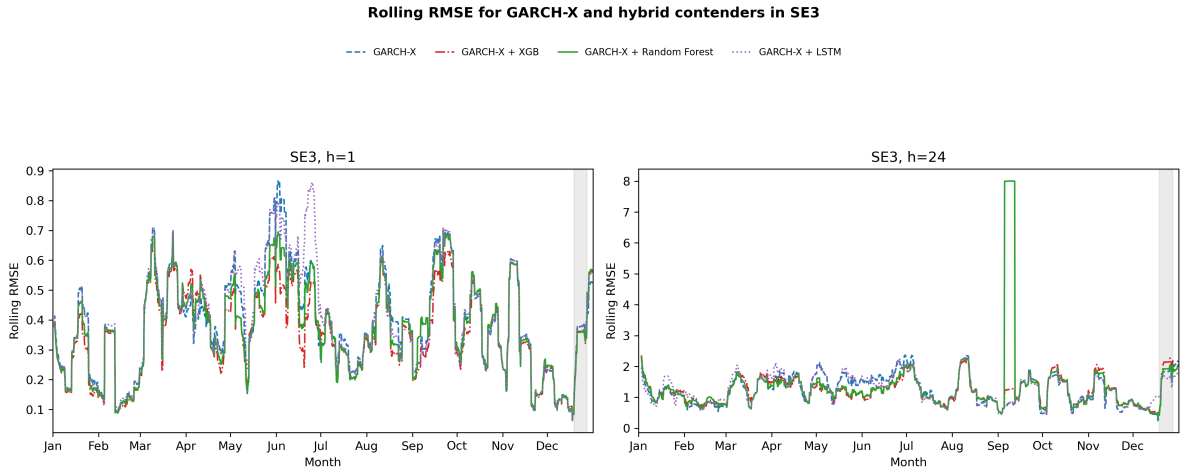


Figure 15: Rolling RMSE for GARCH-X and hybrid contenders in zone SE3 at horizons $h = 1$ and $h = 24$ under the constrained information design. The figure corresponds to the full-year 2025 out-of-sample evaluation window and uses distinct line styles for GARCH-X, XGBoost, Random Forest, and LSTM.

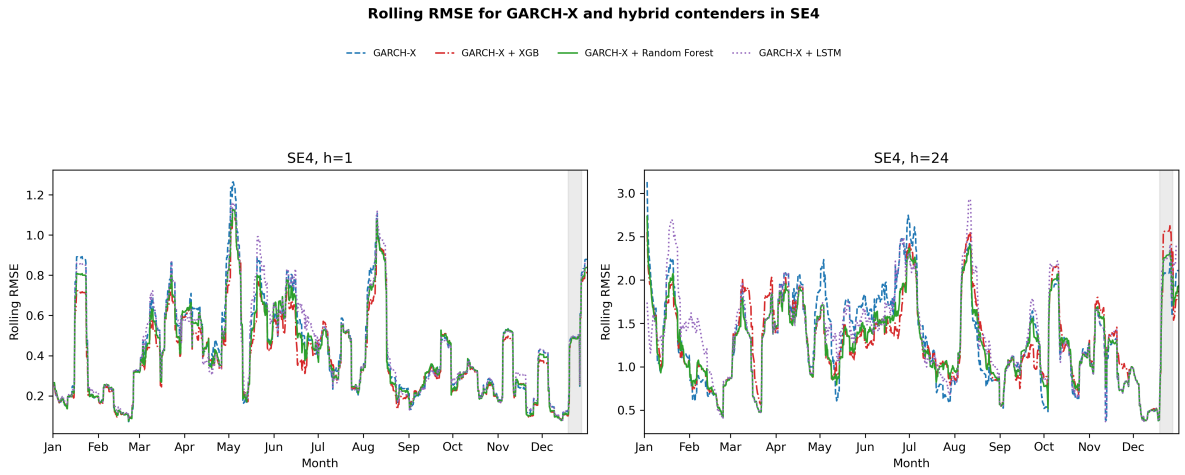


Figure 16: Rolling RMSE for GARCH-X and hybrid contenders in zone SE4 at horizons $h = 1$ and $h = 24$ under the constrained information design. The figure corresponds to the full-year 2025 out-of-sample evaluation window and uses distinct line styles for GARCH-X, XGBoost, Random Forest, and LSTM.

Appendix C: Documented Alternative Exogenous-Information Designs

This appendix records two alternative future-path designs considered during the information-set design stage. They are not used for quantitative model ranking in the thesis because the exported price-space diagnostics were numerically unstable under the rolling standardisation and inverse arcsinh transformation. The constrained design in Section 5 is therefore the only information-set design used for reported forecast comparisons.

The seasonal-naive design repeats the last 24 observed hours of each non-deterministic exogenous regressor cyclically. The design is operationally feasible, but it imposes a stronger persistence structure than the main carry-forward rule.

The oracle design replaces the extrapolated path with realised future exogenous values. It is useful as a conceptual upper-bound design, but it is infeasible at the forecast origin and is therefore unsuitable for the thesis's operational model comparison.

Appendix D: Feature Engineering Details

D.1 Cyclical Time Encoding

Hour-of-day and day-of-week are encoded using sine and cosine transformations rather than integer indicators, preserving the cyclic continuity of these features (i.e. hour 23 and

hour 0 are adjacent):

$$\texttt{sin\_hour}_t = \sin\left(\frac{2\pi \cdot h_t}{24}\right), \qquad \texttt{cos\_hour}_t = \cos\left(\frac{2\pi \cdot h_t}{24}\right), \tag{12}$$

$$\texttt{sin\_dow}_t = \sin\left(\frac{2\pi \cdot d_t}{7}\right), \qquad \texttt{cos\_dow}_t = \cos\left(\frac{2\pi \cdot d_t}{7}\right), \tag{13}$$

where $h_t \in \{0, \ldots, 23\}$ is the hour of day and $d_t \in \{0, \ldots, 6\}$ is the day of week. These four features replace indicator variables for hours and days, reducing the number of calendar parameters from 30 to 4 while providing a smooth representation of periodic demand and trading patterns.

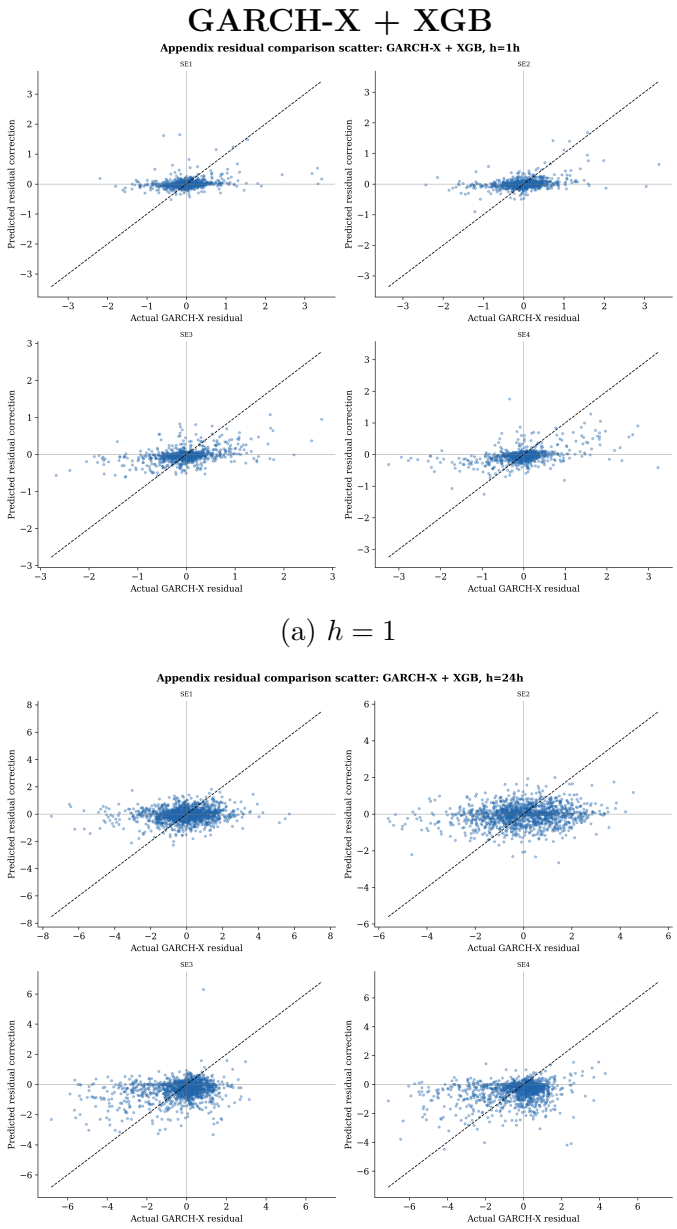
D.2 Standardisation within Rolling Windows

All non-binary exogenous regressors are standardised within each 90-day training window before estimation:

$$\tilde{x}_t^{(j)} = \frac{x_t^{(j)} - \bar{x}_{\text{train}}^{(j)}}{s_{\text{train}}^{(j)}}, \tag{14}$$

where $\bar{x}_{\text{train}}^{(j)}$ and $s_{\text{train}}^{(j)}$ are the mean and standard deviation of feature j over the training window. The same scaling parameters are applied to the corresponding test observations to prevent data leakage. This procedure is repeated at each forecast origin, ensuring that standardisation uses only information available at that point in time.

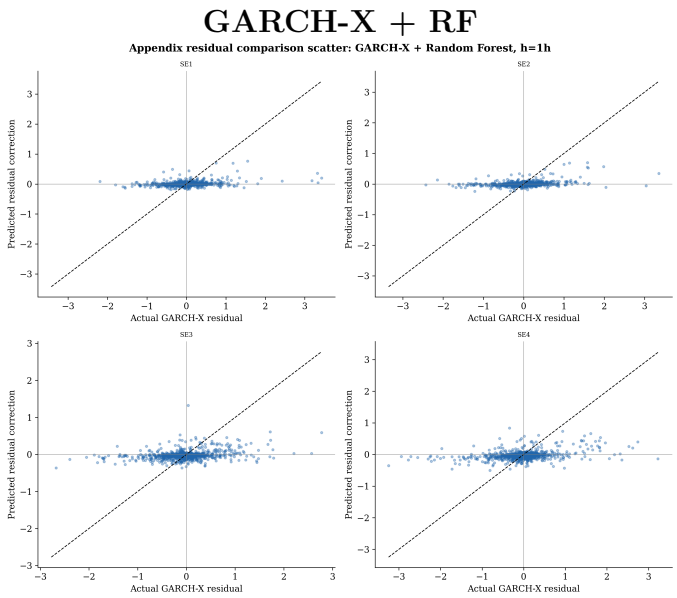
D.3 Residual-Correction Scatter Diagnostics



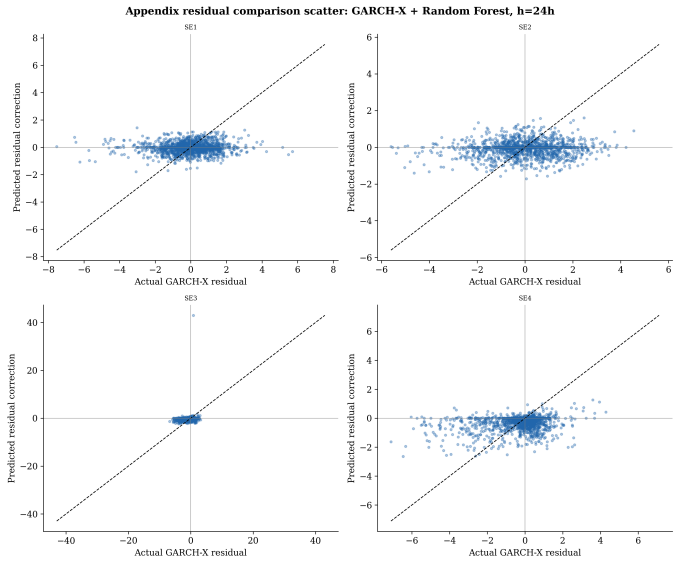
(a) $h = 1$

(b) $h = 24$

Figure 17: Residual-correction scatter under the constrained (constant) design for GARCH-X + XGB at $h = 1$ and $h = 24$. Horizontal axis: realised GARCH-X residual. Vertical axis: predicted XGB correction. Dashed diagonal represents perfect residual prediction; points above overshoot, points below undershoot.

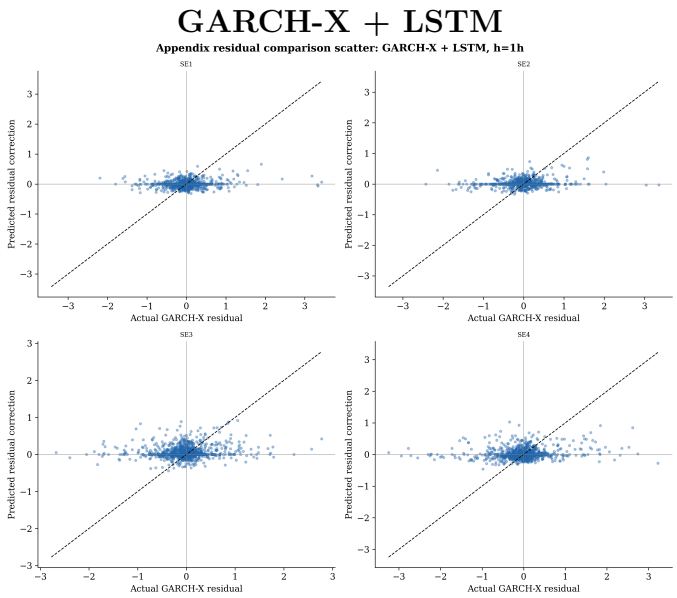


(a) $h = 1$

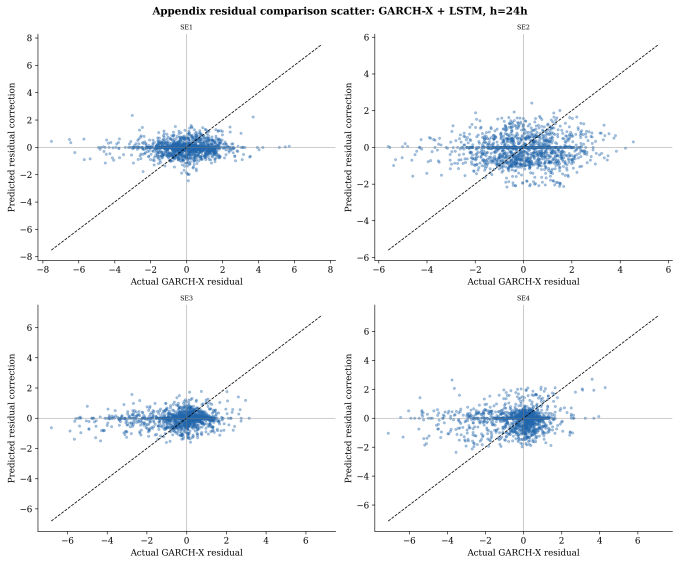


(b) $h = 24$

Figure 18: Residual-correction scatter under the constrained (constant) design for GARCH-X + Random Forest at $h = 1$ and $h = 24$. Axes and diagonal reference line as in Figure 17.



(a) $h = 1$



(b) $h = 24$

Figure 19: Residual-correction scatter under the constrained (constant) design for GARCH-X + LSTM at $h = 1$ and $h = 24$. Axes and diagonal reference line as in Figure 17.

Appendix E: Empirical Workflow Overview

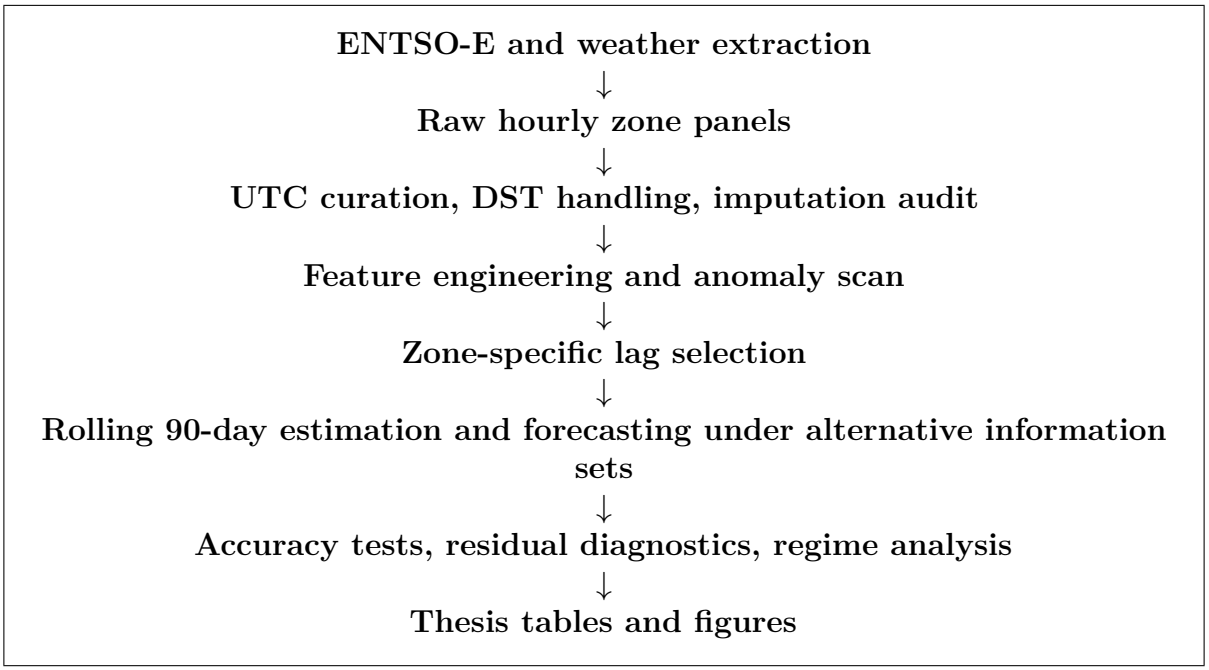


Figure 20: High-level empirical workflow used in the thesis.

The empirical pipeline begins with hourly ENTSO-E API extraction restricted to the series that enter the model-ready panel: prices, load, domestic and foreign net-flow aggregates, retained generation shares, and planned and unplanned outages. Zone-level hourly skeletons are then built and merged with SMHI weather data aggregated to bidding-zone level using SCB population weights. For transparency, the data pipeline retains a raw panel in local time, but it contains the same modelling variables before cleaning rather than additional unused market series. A canonical panel in UTC is also constructed so that daylight-saving duplicates and missing local-clock hours can be handled explicitly. Missing load, flow, and temperature observations are filled using local historical patterns and edge-aware weekly imputations, outage gaps are set to zero because outages are reported by exception, and invalid retained generation-share observations are discarded and re-imputed. Missing electricity prices are never imputed; instead, those rows are removed from the final model-ready panel and recorded in the curation audit and flag tables. On the curated panel, the transformed dependent variable $y_t = \operatorname{asinh}(price_t)$, deterministic calendar terms, autoregressive lags, rolling moments, and outage standardisations are created centrally before modelling. Lag selection is then run separately by bidding zone, yielding compact endogenous and exogenous lag sets. Forecast evaluation uses a rolling 90-day estimation window and repeated forecast origins over the full-year 2025 out-of-sample period for GARCH(1,1), ARX (no GARCH), GARCH-X, and the residual-learning hybrids at the main horizons. Multi-step forecasts are generated under explicit exogenous-information designs, with the constrained design used in the main text and the

seasonal-naive and oracle designs documented as alternative information-set concepts. The backtest exports prediction objects from which transformed-space and price-space RMSE and MAE are computed. Residual ACF, Ljung–Box, ARCH–LM, Diebold–Mariano tests with Newey–West HAC variance estimation, residual-signal diagnostics, and volatility-regime decompositions are then computed on the same forecast objects. Finally, the thesis export routines map these outputs into the tables and figures used in Sections 5 and 6 and in the appendix.

Appendix F: LSTM Hyperparameter Configuration

The LSTM residual corrector described in Section 4.5.3 is implemented in PyTorch as a recurrent network whose target is the GARCH-X residual-target series e_t; the same exogenous feature set as the tree-based contenders is re-shaped into rolling input sequences. The principal hyperparameters used in the rolling out-of-sample evaluation are listed in Table 17 and are consistent with the training configuration summarised in the Section 4.5.3 footnote. Unlike the GARCH-X baseline and the tree-based residual learners, which are re-estimated at every forecast origin, the LSTM is retrained at most once per week within each rolling backtest stream; this is the LSTM staleness limitation acknowledged in the limitations subsection of the discussion. Hyperparameters were not tuned individually per zone in order to keep the comparison across the four bidding zones symmetric and to avoid zone-specific overfitting within the 2025 evaluation window. A broader hyperparameter search is a natural robustness extension but lies outside the operational comparison studied in this thesis.

Table 17: Principal hyperparameters used for the LSTM residual corrector in the rolling out-of-sample evaluation. Training-related values match the Section 4.5.3 footnote.

Hyperparameter	Value
Framework	PyTorch (`torch.nn.LSTM`)
Input sequence length	48 hours
Hidden units	32
Recurrent layers	1
Target	GARCH-X residual series
Loss function	Mean squared error
Optimiser	Adam
Learning rate	1×10^{-3}
Batch size	64
Training epochs (per fit)	up to 50
Early-stopping patience	4
Dropout rate	0.10
Weight decay	1×10^{-5}
Gradient clipping	1.0
Re-fitting cadence	Weekly within rolling backtest

Appendix G: In-Sample Volatility Diagnostics for the GARCH Specification

The conditional-variance specification used throughout the thesis is a GARCH(1,1) process with Student-t innovations. While the main text uses GARCH-X point forecasts for ranking, this appendix reports two standard in-sample variance diagnostics applied to the GARCH(1,1) variance component on a representative 90-day in-sample window (matching the rolling estimation window length). The purpose is to establish that the variance specification is itself empirically supported, independent of the mean-equation forecast comparison reported in Section 5.

QLIKE diagnostic. The QLIKE loss is computed as

$$\overline{\text{QLIKE}} = \frac{1}{T}\sum_{t=1}^{T}\left[\ln(\hat{\sigma}_t^2) + \frac{r_t^2}{\hat{\sigma}_t^2}\right],$$

where r_t is the residual from the GARCH(1,1) mean equation, r_t^2 is the realised variance proxy, and $\hat{\sigma}_t^2$ is the fitted GARCH(1,1) conditional variance. Smaller values indicate a closer match between the fitted variance and the realised proxy on this scale. The in-sample QLIKE values on the 90-day window are: SE1 = 0.820, SE2 = 1.145, SE3 = 0.071, and SE4 = -0.125. The QLIKE level is not directly comparable across zones because price scales and residual variances differ; the values are reported here as a within-zone fit diagnostic rather than as a cross-zone ranking.

Mincer–Zarnowitz regression. As a calibration check, squared raw residuals are regressed on the fitted GARCH(1,1) conditional variance:

$$r_t^2 = \alpha + \beta\,\hat{\sigma}_t^2 + u_t,$$

with HAC standard errors. A well-specified variance forecast produces an intercept close to zero and a slope coefficient close to one. Estimated coefficients by zone are reported in Table 18: slopes lie between 0.92 and 0.96 and R^2 values lie between 0.85 and 0.92 across all four zones, indicating that the GARCH(1,1) variance specification is well-calibrated against the in-sample squared-residual proxy. Small positive intercepts (0.08–0.18) suggest mild under-prediction of variance in extreme states, consistent with the heavy-tailed Student-t innovation specification absorbing only part of the tail behaviour.

Table 18: Mincer–Zarnowitz regression coefficients for the in-sample GARCH(1,1) variance estimates by bidding zone, computed on the representative 90-day window. The regression is $r_t^2 = \alpha + \beta\hat{\sigma}_t^2 + u_t$ with HAC standard errors. A well-calibrated variance specification has $\alpha \approx 0$ and $\beta \approx 1$.

Zone	$\hat{\alpha}$	p_α	$\hat{\beta}$	p_β	R^2
SE1	0.183	0.021	0.923	< 0.001	0.851
SE2	0.094	0.001	0.962	< 0.001	0.924
SE3	0.081	0.003	0.953	< 0.001	0.906
SE4	0.096	0.004	0.952	< 0.001	0.907

Taken together, the QLIKE and Mincer–Zarnowitz results provide in-sample empirical support for the GARCH variance specification: fitted variances are approximately unbiased against the squared-residual proxy and explain 85–92% of its variation. These are in-sample diagnostics on the GARCH(1,1) variance component; out-of-sample variance forecast evaluation of GARCH-X over the full-year 2025 backtest period is a natural extension and lies outside the mean-forecast comparison that defines the main thesis.

Appendix H: AI Disclosure

H.1 Tools Used and How

Generative AI tools were used during the research and writing process for this thesis. The tools used were ChatGPT/CODEX (OpenAI) and Claude (Anthropic). Their use was confined to the following tasks:

- **Drafting and editing assistance:** rewording paragraphs, tightening prose, checking grammar, and improving clarity in successive drafts. All substantive content, arguments, and interpretations are our own.
- **Code review and debugging:** explaining error messages, suggesting refactorings, and reviewing Python implementations of the GARCH-X, XGBoost, Random Forest, and LSTM pipelines. All code was executed, inspected, and validated by us.
- **Sense-checking:** discussing methodological choices (e.g., HAC bandwidth, transformation handling, lag selection logic) to test our own reasoning before committing to a design.
- **LaTeX formatting:** assistance with table layouts, figure placement, and bibliography formatting.

H.2 Contribution to Thesis Quality

AI tools contributed in two main ways. First, they accelerated routine writing tasks such as editing, reformatting, and drafting boilerplate, which freed time for the substantive empirical work. Second, they functioned as a sounding board for methodological decisions, allowing us to articulate and stress-test our reasoning before implementing a given design. Neither of these uses replaced our own judgment; they supported it.

H.3 Risks Identified and Mitigations Applied

The main risks identified were inaccurate references, incorrect methodological suggestions, and over-reliance on AI-generated prose. These risks were mitigated by verifying all cited sources against the original publications and checking methodological suggestions against the relevant econometric literature.

H.4 Insights Gained

Using AI tools in a structured way reinforced two lessons. First, the tools are most useful when the user already understands the underlying material; they help articulate and refine, but they do not substitute for domain knowledge. Second, verification is non-negotiable: every AI-generated claim, reference, or code suggestion that entered the thesis was independently checked. The practical workflow that emerged, drafting independently, using AI for editing and stress-testing, and verifying everything, is one we expect to apply in future research and professional work.