How Power System Structure Shapes Short-Term Electricity Forecasting Performance: A Comparative Study of Germany and Portugal

Aaron Fischer

Dissertation written under the supervision of professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 04.01.2026

Abstract (English)

How Power System Structure Shapes Short-Term Electricity Forecasting Performance: A Comparative Study of Germany and Portugal

Aaron Fischer

Short-term forecasts of electricity consumption, production, and net imbalance are critical for power system operation, as forecast errors directly influence balancing actions, reserve activation, and system security. While most existing research focuses on improving predictive accuracy, far less attention has been given to how forecasting performance depends on the structural characteristics of the underlying power system. This thesis examines how system structure shapes short-term forecasting performance through a comparative study of Germany and Portugal.

The analysis focuses on one-hour and twenty-four-hour ahead horizons and applies a unified machine learning framework across both countries, including tree-based ensemble models and recurrent neural networks. Forecast performance is evaluated using standard accuracy metrics and a severity-based framework that distinguishes between operationally minor and high-risk deviations. Model behavior is further examined using SHAP-based interpretability.

The results reveal pronounced cross-country differences despite identical models. Germany exhibits lower errors and fewer severe deviations, while Portugal shows greater sensitivity, particularly at longer horizons. These differences are closely linked to structural factors such as system size, renewable penetration, and interconnection capacity, highlighting the importance of interpreting forecast performance within its system context.

Keywords: Short-term electricity forecasting, net imbalance, power system structure, machine learning, renewable energy integration, forecast error severity, power system operation

Abstract (Portuguese)

Como a Estrutura do Sistema Elétrico Molda o Desempenho das Previsões de Eletricidade de Curto Prazo: Um Estudo Comparativo entre a Alemanha e Portugal

Aaron Fischer

Previsões de curto prazo do consumo, da produção e do desequilíbrio líquido de eletricidade são fundamentais para a operação dos sistemas elétricos, uma vez que erros de previsão afetam diretamente o balanceamento, a ativação de reservas e a segurança do sistema. Apesar de a investigação existente se concentrar sobretudo na melhoria da precisão preditiva, pouca atenção tem sido dada à influência da estrutura do sistema elétrico no desempenho das previsões. Esta dissertação analisa esse impacto através de um estudo comparativo entre a Alemanha e Portugal.

A análise incide nos horizontes de uma hora e de vinte e quatro horas, aplicando uma estrutura unificada de aprendizagem automática, incluindo modelos ensemble baseados em árvores e redes neuronais recorrentes. O desempenho é avaliado com métricas de precisão padrão e uma abordagem baseada na severidade, complementada por análise interpretável com valores SHAP.

Os resultados mostram diferenças claras entre os dois países, apesar do uso de modelos idênticos. A Alemanha apresenta erros mais baixos e menos desvios severos, enquanto Portugal revela maior sensibilidade, sobretudo em horizontes mais longos, refletindo fatores estruturais como a dimensão do sistema, a penetração de renováveis e a capacidade de interconexão.

Palavras-chave: Previsão de eletricidade de curto prazo, desequilíbrio líquido, estrutura do sistema elétrico, aprendizagem automática, integração de energias renováveis, severidade do erro de previsão, operação do sistema elétrico

Table of Contents

1. Introduction

The growing use of renewable energy sources has fundamentally changed how modern power systems work. Wind and solar power generation cause more short-term changes and make it harder to predict what will happen. The gradual replacement of synchronous generation makes the system less stable and less flexible in terms of balancing. As a result, it has become harder to keep the real-time balance between electricity supply and demand, especially in systems where a lot of generation depends on the weather (Ejuh Che et al., 2025; Graabak & Korpås, 2016). These changes have made short-term electricity forecasting more important, not just for market efficiency but also for system security and operational reliability.

In reality, mistakes in forecasts have direct effects on both the economy and the environment. When actual consumption or production is different from what was expected, there are net imbalances that need to be fixed by activating reserves, redispatching, or trading across borders. Automatic control systems can handle small imbalances, but bigger ones can put a strain on reserves, raise operational costs, change market prices, and, in extreme cases, threaten frequency stability. Recent events in Europe show this range of outcomes: times when demand and renewable generation are misaligned have caused electricity prices to drop to zero or even negative in some markets, costing producers money (Bloomberg, 2025; Pavlík et al., 2025). On the other hand, rare but serious imbalance events have caused widespread outages and emergency interventions (REUTERS, 2025).

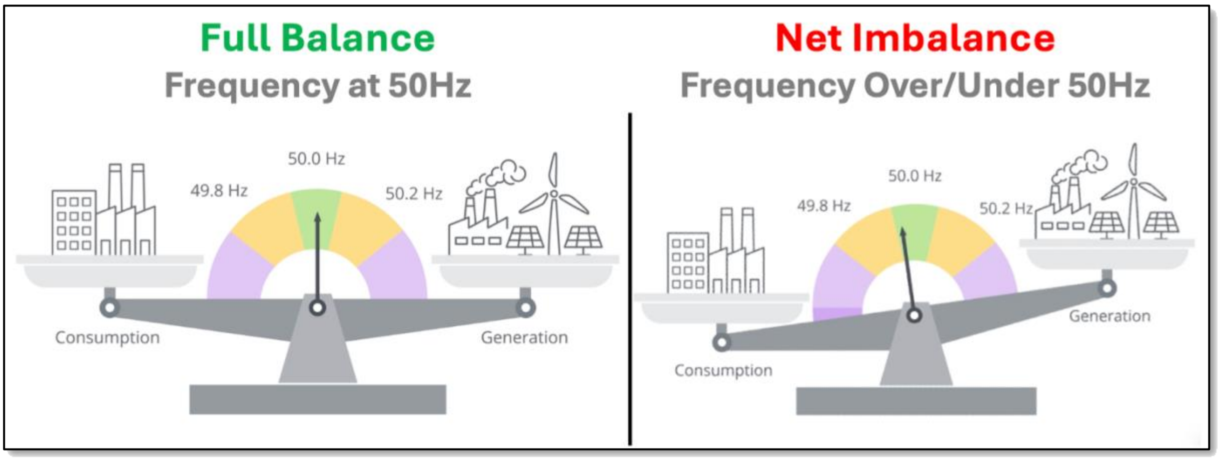


Figure 1: Overview of balancing services (adapted from regelleistung.net, "What are balancing services", accessed 14.12.2025)

The operational significance of a forecast error is contingent not only on its numerical value but also on the structural attributes of the power system in which it transpires. In a big, highly interconnected system, a deviation that is not critical may already be a big risk in a smaller or less inertial grid. This system dependence is especially clear when looking at countries with substantially different grid sizes, generation mixes, and interconnection structures. Germany has one of the largest and most interconnected transmission networks in Europe. It has a lot of reserve capacity, strong spatial smoothing effects, and many ways to balance regional imbalances. Portugal, on the other hand, is a smaller and less interconnected system. It has a higher percentage of renewables, less inertia in the system during certain operating conditions, and a strong dependence on support from Spain. These structural differences, which are clearly shown in the European transmission network topology published by ENTSO-E, mean that the same forecast errors can lead to very different operational risks in different national power systems. (ENTSOE-E Grid Map, 2025)

Despite these differences, much of the electricity forecasting literature evaluates models using aggregate error metrics such as MAE or RMSE, which are useful for comparison but say little about the operational risk of forecast errors across different power systems (Hong et al., 2020; Suganthi & Samuel, 2012).

In this context, this thesis examines short-term forecasting of electricity consumption, production, and net imbalance via a comparative analysis of Germany and Portugal. Instead of solely concentrating on the creation of a novel forecasting model, the focus is on analyzing forecast performance through a system-specific and severity-sensitive perspective. The thesis aims to investigate how structural disparities in national power systems influence not only forecast accuracy but also the severity and risk profile of forecast errors, by integrating multi-horizon forecasting with operationally based severity thresholds and model explainability techniques.

The central research question addressed is: **To what extent do structural differences in national power systems shape short-term net imbalance forecast performance and the severity of errors across time horizons, as demonstrated through a comparative analysis of Germany and Portugal?**

By answering this question, the thesis aims to contribute to a more operationally meaningful interpretation of electricity forecasts-one that treats forecasting not merely as a statistical exercise, but as a fundamental component of power system risk management.

The remainder of this thesis is structured as follows. Chapter 2 reviews the relevant literature on short-term electricity forecasting, with particular emphasis on forecasting objectives, horizons, methodologies, evaluation practices, and interpretability. Chapter 3 presents the theoretical framework, linking power system operation with time-series forecasting theory, supervised learning, and risk-oriented model evaluation. Chapter 4 describes the data sources and preprocessing steps used to construct harmonized country-level datasets for Germany and Portugal. Chapter 5 outlines the forecasting methodology, including model selection, training protocols, and evaluation metrics. Chapter 6 presents the empirical results, focusing on forecast accuracy, severity-based net imbalance analysis, and SHAP-based interpretability. Chapter 7 discusses the results in relation to the research question, operational implications, and limitations of the study. Finally, Chapter 8 concludes the thesis by summarizing the main findings, contributions, and directions for future research.

2. Literature Review

2.1 Electricity Forecasting Objectives

Traditionally, electricity forecasting has focused on predicting either system load (consumption) or generation. The primary objective operational planning, market clearing, and infrastructure investment. Load forecasting remains an important part of power system operation, as accurate demand predictions are essential for unit commitment, reserve sizing, and congestion management. At the same time, production forecasting, especially for renewable energy sources like wind and solar, has become increasingly important as these sources have become more common in European power systems (Benti et al., 2023; Hong et al., 2020; Mystakidis et al., 2024; Sweeney et al., 2020).

Recently, the focus has moved to the idea of net imbalance, which is the residual difference between actual production and consumption. Net imbalance directly shows the stress that balancing mechanisms and frequency control are experiencing, unlike forecasts of consumption or production that are looked at individually. Even though net imbalance forecasting is important for operations, it remains an underexplored topic in academic literature. Instead, it is often only touched on by combining separate load and generation forecast errors. This gap has gotten bigger in the last few years because of the increasing share of variable renewable energy sources has grown, making production more volatile and system imbalances more frequent and severe (Balázs et al., 2024; Plakas et al., 2023a).

2.2 Forecasting Horizons and Operational Context

Forecasting requirements vary significantly across different time horizons. Short-term forecasts, which usually cover a time frame of minutes to one hour, are mostly used for real-time balancing and frequency control. At these horizons, recent observations and autocorrelation are the most important factors in predictive performance. Errors can often be mitigated with fast reserves and automatic control actions.

Intra-day horizons, such as four hours ahead, help with rescheduling operations and setting aside resources. They are a transitional regime where both short-term dynamics and outside factors, like changes in the weather, become important. Day-ahead forecasts, typically set at 24-hour intervals, play a key role in market clearing and system planning (Soman et al., 2010;

Z. Zhao et al., 2021). At these longer horizons, uncertainty increases significantly due to reduced persistence and higher exposure to forecast errors in weather-driven generation.

The literature consistently indicates a decline in forecast accuracy with increasing horizons; however, there are relatively few studies that explicitly associate this loss in forecast accuracy with operational risk or the severity of consequences.

2.3 Forecasting Methodologies

For a long time, people have used classical statistical models like ARIMA and its variations to predict electricity due to their interpretability and strong theoretical foundation. These models perform well under stable conditions with strong seasonal patterns but struggle to capture non-linear relationships and complex interactions introduced by high renewable penetration and volatile weather dynamics (Chavez et al., 1999; Soman et al., 2010).

In recent years, machine learning techniques like Random Forests and gradient boosting methods like XGBoost have become increasingly popular. These models are well suited to electricity forecasting problems, as they can capture non-linear relationships, handle high-dimensional feature spaces, and integrate heterogeneous data sources typically available in tabular form. Many studies show superior performance of gradient boosting models over traditional statistical methods, particularly for short- and medium-term forecasting tasks. (Benti et al., 2023; Plakas et al., 2023b; Shahiduzzaman et al., 2021)

Deep learning architectures, including LSTM, GRU, and hybrid CNN–RNN models, are frequently seen as cutting-edge solutions, particularly for modeling temporal dependencies Some studies show that performance can improve, while others point out problems with overfitting, high computational costs, and low robustness when data availability or system conditions change. Consequently, the empirical advantage of deep learning compared to fine-tuned machine learning models remains context-dependent. (Gasparin et al., 2019; Hong et al., 2020; Mystakidis et al., 2024)

2.4 Data and Feature Engineering in the Literature

The choice of input features has a strong influence on how well a model can predict. Lagged target variables, calendar effects, and weather inputs like temperature, wind speed, and solar irradiance are some of the most common predictors. For renewable generation forecasting, weather features are often the most important part of the model (Salcedo-Sanz et al., 2018). For predicting consumption, temporal patterns and socio-economic proxies are more important (H.-X. Zhao & Magoulès, 2012).

Particularly when interpreting model behavior, leakage and overreliance on autoregressive features can become problematic. This motivates the use of reduced or structured feature sets in explainability-focused analyses, even though such choices may come at the cost of predictive accuracy.

2.5 Evaluation Practices and Interpretability

Forecast evaluation in the literature is dominated by aggregate error metrics like MAE, RMSE, and MAPE (Ledmaoui et al., 2023; Mystakidis et al., 2024). While these measures facilitate model comparison, they implicitly assume that forecast errors lead to the same consequences across magnitudes and system contexts. This assumption is increasingly questioned in operational settings, where the cost and risk associated with errors grow non-linearly with their size.

As machine learning has become more popular, model interpretability has become more important too. SHAP values, which come from cooperative game theory, are often used to link predictions to input features and to make black-box models more trustworthy. However, most studies use SHAP to explain average model behaviour and do not focus on the causes of large or high-risk forecasting errors (Mystakidis et al., 2024).

2.6 Gaps in the Literature

Even though a lot of research has been done on predicting electricity use, several gaps remain. First, net imbalance forecasting is rarely analyzed as a primary target, particularly at high temporal resolution. Second, there aren't many evaluation frameworks that take into account the severity of errors and connect them to operational risk. Third, cross-country comparisons that explicitly account for structural system differences - such as inertia, grid topology, and reserve capacity - are limited. Lastly, people don't often use explainability methods to figure out what causes large or serious forecast errors.

This thesis fills in these gaps by doing a comparative study of Germany and Portugal that uses multi-horizon forecasting, country-specific severity thresholds, and SHAP-based interpretability.

3. Theoretical Framework

This chapter lays the groundwork for the forecasting methodology, evaluation strategy, and interpretability approach used in this thesis. It connects the principles of power system operation with time-series forecasting theory and supervised machine learning. This is the basis for the severity-aware interpretation that will be developed in later chapters.

3.1 Power System Operation and the Concept of Net Imbalance

Electric power systems require constant balance between electricity generation and consumption to maintain system frequency within narrow operational limits. When this balance is disturbed - resulting in net imbalance - frequency deviations occur that must be addressed through automatic and manual control actions.

From an operational point of view, net imbalance is more than just a number; it is a **risk signal.** Primary and secondary control reserves can handle small imbalances, but larger ones need tertiary reserves, redispatch, or emergency measures to keep the frequency at 50Hz. As systems get closer to stability limits like reserve exhaustion or under-frequency load shedding thresholds, the relationship between imbalance size and operational impact becomes decreasingly linear. This non-linearity necessitates the categorization of forecast errors into severity levels instead of evaluating them exclusively via aggregate metrics.

3.2 Time Series Forecasting Foundations

Electricity demand and generation exhibit strong temporal patterns, including daily, weekly, and seasonal cycles, as well as dependence on past values. Traditional time-series models are designed to capture such patterns using assumptions about stationarity and autoregressive behaviour.

In contrast, modern machine learning approaches treat forecasting as a supervised learning task, where relationships between past observations, external variables, and future outcomes are learned directly from data. While these modelling approaches differ in flexibility and interpretability, all forecasting models face increasing uncertainty as the prediction horizon extends. This growing uncertainty explains the systematic decline in forecasting performance observed at longer horizons in empirical studies.

3.3 Forecasting Horizon Theory and Operational Use

Different operational decision contexts are shown by forecasting horizons. Short-term horizons, like one hour ahead, are closely linked to real-time balancing and work best when the system's behavior stays the same over time. Longer horizons, such as day-ahead forecasts, help with market clearing and resource planning, but are more exposed to uncertainty in weather-driven generation and demand.

As horizons get longer, losing information about how the system is doing in the short term not only increases the average error, but it also makes extreme deviations more likely. From a system operation point of view, this change means that there will be more chances of severe imbalance events. This makes it even more important to look at forecasts not only for their accuracy but also for their risk implications.

3.4 Supervised Learning and Gradient Boosting

Supervised machine learning models learn predictive relationships between input features and target variables by minimizing a loss function over historical data. XGBoost and other gradient boosting methods build groups of decision trees one at a time, with each tree correcting the mistakes made by the ones before it.

Gradient boosting performs well because it can capture complex non-linear relationships while maintaining good generalization. However, the loss functions typically used treat all forecast errors equally, even though large errors are much more costly in power system operation. This motivates the use of severity-aware evaluation frameworks.

3.5 Interpretable Machine Learning and SHAP Values

Interpretability is particularly important in safety-critical and operational applications. SHAP values provide a practical framework for understanding how input features contribute to model predictions, both at an overall level and for individual cases. While SHAP does not imply causality, it helps explain how models use the available information. In this thesis, SHAP is applied to identify the factors driving large forecast errors and to relate them to system characteristics and operational risk.

3.6 Model Evaluation Theory and Risk-Oriented Assessment

Traditional model evaluation relies on average performance measures such as MAE or RMSE, assuming that forecast errors have similar consequences. In practice, power system operation exhibits a non-linear relationship between error magnitude and operational cost or risk.

Theoretically, this supports the implementation of evaluation frameworks that differentiate between error regimes instead of solely depending on aggregate statistics. Severity-aware classification connects model evaluation with operational decision-making. This means that forecasts can be understood in terms of how they might affect the system, not just how accurate the predictions are.

4. Data and Preprocessing

4.1 Data Sources

This study integrates publicly available datasets describing electricity system behavior, market prices, meteorological conditions, and temporal information for Germany and Portugal. All raw data were transformed to a common 15-minute resolution and aligned across sources before modeling. This section provides an overview of the original data sources, their formats, and the types of variables they contain.

4.1.1 Electricity System Data: Consumption, Production, and Cross-Border Flows

Electricity consumption, electricity generation by source, import and export data were obtained from the national transparency platforms in each country. All four domains originate from the **same source per country** and are already provided in a **uniform 15-minute resolution**.

Germany - SMARD (SMARD Platform, 2025)

For Germany, all operational electricity system measurements were obtained from SMARD, the transparency platform operated by the German Federal Network Agency (Bundesnetzagentur). The SMARD platform provides 15-minute total electricity consumption (grid load), 15-minute electricity generation disaggregated by technology-including wind (onshore and offshore), solar photovoltaics, biomass, hydropower, natural gas, coal, and nuclear generation (active until 2023) - as well as 15-minute physical cross-border electricity exchanges aggregated across all interconnectors.

This data represents official measurements reported by the German transmission system operators and reflect the German power system in real time.

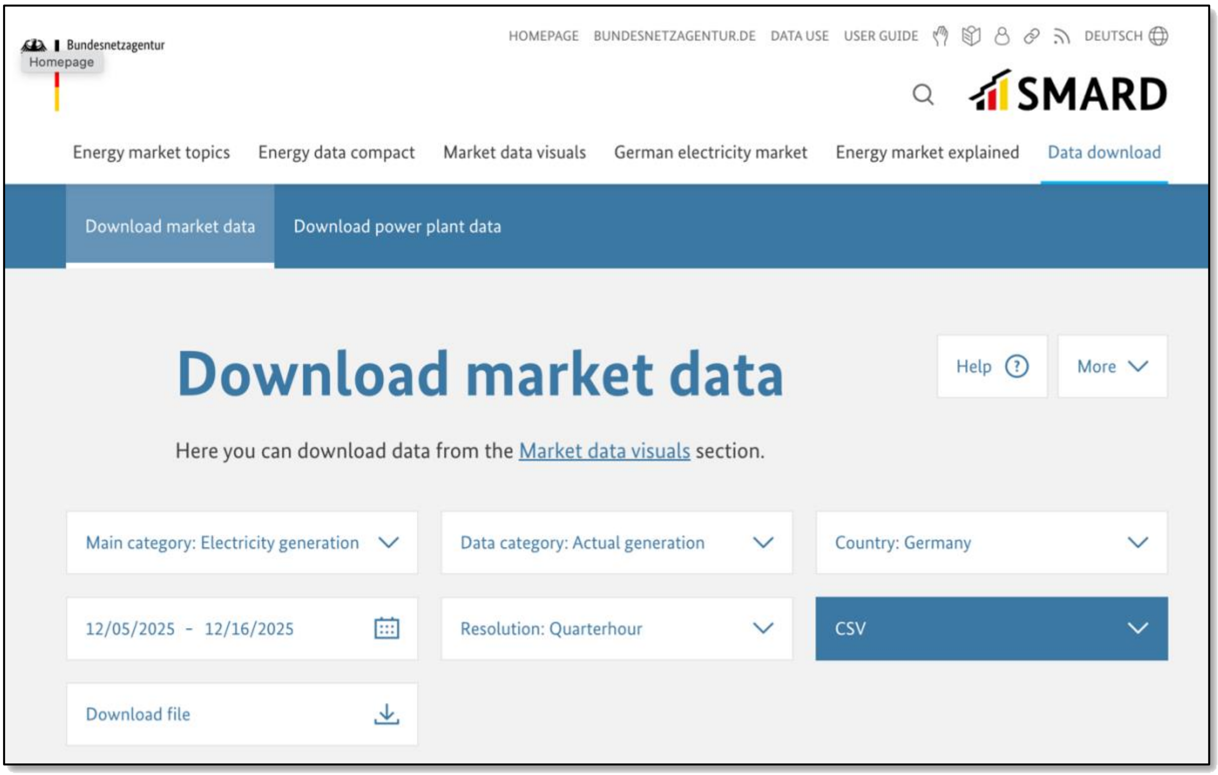


Figure 4.1: SMARD electricity market data download interface (Bundesnetzagentur, 2025)

Portugal - REN DataHub (REN DataHub, 2025)

For Portugal, equivalent operational electricity system data were collected from the REN DataHub, maintained by Redes Energéticas Nacionais (REN). The platforms API provides 15-minute national electricity consumption, 15-minute electricity generation by technology-including hydropower (run-of-river, reservoir, and pumping), wind, solar, biomass, and thermal generation-and 15-minute import and export flows between Portugal and Spain.

Because Portugal is electrically interconnected with only a single neighbouring system these cross-border flow measurements capture the full extent of Portugal's international electricity exchanges and therefore represent the complete external balancing interface of the Portuguese power system.

```
import requests
import csv
from datetime import datetime, timedelta

def daterange(start_date, end_date):
    """Yield datetime.date from start_date to end_date inclusive."""
    current = start_date
    while current <= end_date:
        yield current
        current += timedelta(days=1)

def fetch_daily_breakdown(date_obj):
    """Fetch JSON for a specific date."""
    url = "https://servicebus.ren.pt/datahubapi/electricity/ElectricityProductionBreakdownDaily"
    params = {
        "culture": "en-US",
        "date": date_obj.strftime("%Y-%m-%d")
    }
    resp = requests.get(url, params=params, timeout=30)
    resp.raise_for_status()
    return resp.json()

def write_header_once(csv_path, series_names):
    """Write CSV header once at top of file."""
    with open(csv_path, mode="w", newline="", encoding="utf-8") as f:
        writer = csv.writer(f)
        header = ["date", "time"] + series_names
        writer.writerow(header)

def append_daily_to_csv(csv_path, date_obj, json_obj):
    """Append rows for a single day to CSV."""
```

Figure 4.2: REN DataHub API Retrieval Script

4.1.2 Electricity Market Prices

Day-ahead wholesale electricity prices were collected separately, as market data does not originate from the system operation platforms.

Germany - ENTSO-E / Ember (ENTSO-E Transparency Platform, 2025)

German day-ahead electricity prices were obtained from **ENTSO-E**, accessed via the **Ember European Wholesale Price Database**, and are published in **hourly resolution** for the DE-LU bidding zone.

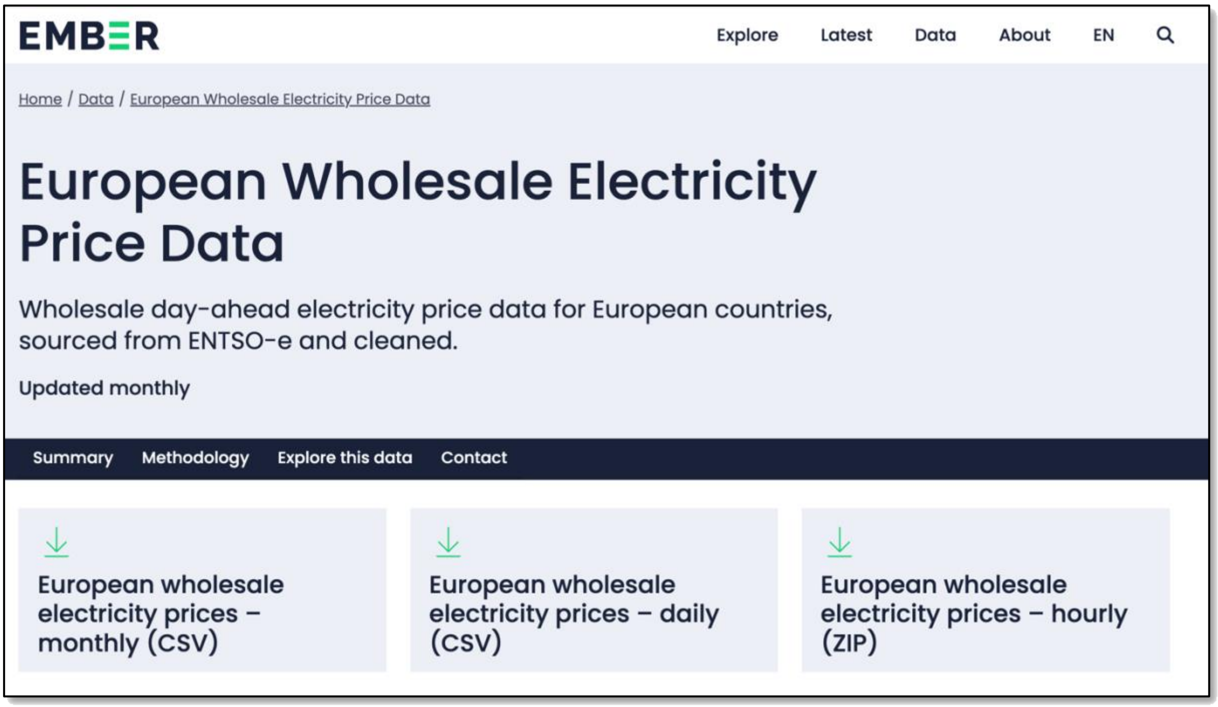


Figure 4.3: ENTSO-E / EMBER Platform Interface (ENTSO-E Transparency Platform, 2025)

Portugal - OMIE / REN DataHub (OMIE, 2025)

Portuguese day-ahead prices were retrieved from the **REN DataHub**, which integrates hourly price data published by **OMIE**, the Iberian day-ahead market operator (MIBEL area).

These market prices were later aligned with the 15-minute operational dataset using time-based interpolation, under the assumption of linear price behaviour within each hour

4.1.3 Weather and Meteorological Data

Weather data for both countries was sourced from **Open-Meteo (Open-Meteo, 2025)** at hourly resolution and includes variables relevant for electricity consumption and renewable generation, such as temperature, wind speed (100 m), cloud cover, radiation, and precipitation. As full nationwide coverage was not available, data from multiple representative locations were aggregated using weighted averages to construct country-level weather series (see Appendix A). These variables were subsequently resampled to a 15-minute resolution to match the electricity system and market data.

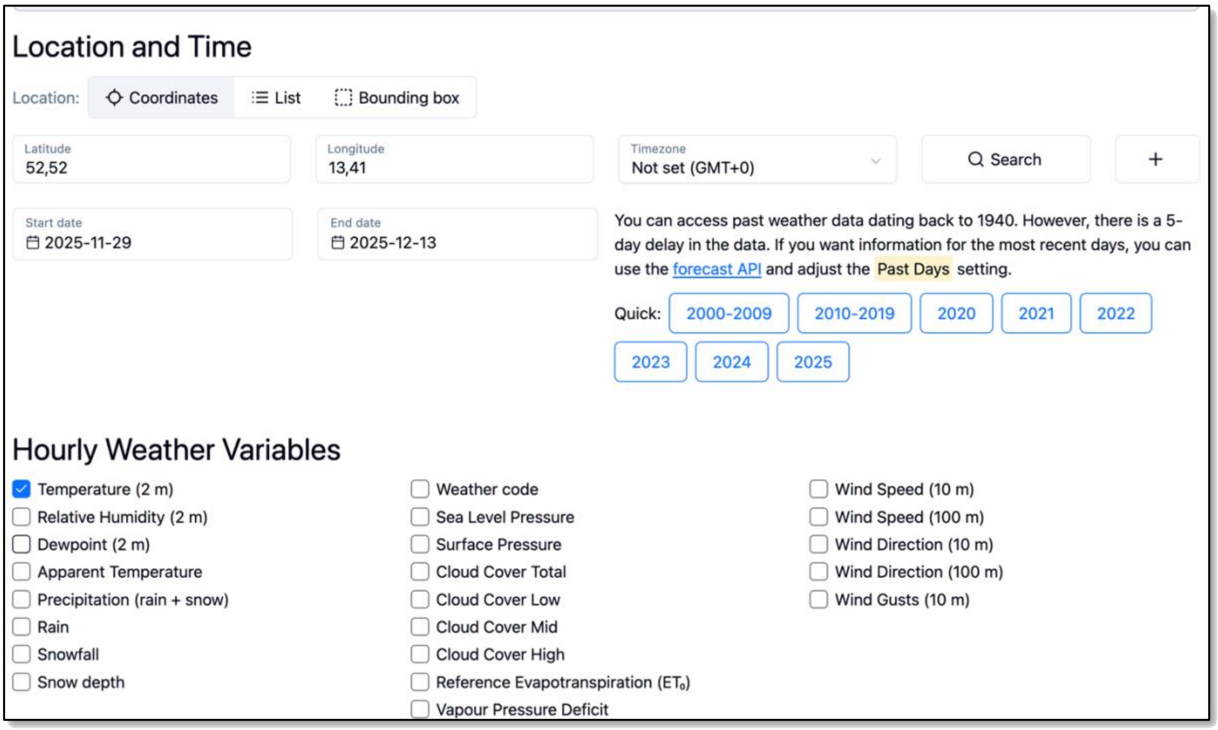


Figure 4.4: OpenMeteo Platform Interface (Open-Meteo, 2025)

4.1.4 Temporal and Calendar Features

Temporal features were constructed directly from the timestamp to complement the physical and market datasets. Cyclical encodings of the hour of day, day of week, and month of year were created using sine and cosine transformations to keep the periodic structure of these variables. In addition, a binary weekend indicator was included to distinguish weekends from weekdays. National public holidays were incorporated using the Python holidays package (Python, 2025), with country-specific calendars applied for Germany and Portugal. Together, these temporal features capture recurring daily, weekly, and seasonal patterns in electricity consumption and renewable generation.

4.1.5 Lag Features

The standardized datasets for both Germany and Portugal also include a set of engineered lag features, such as 15-minute, 1-hour, 2-hour, 24-hour, and 7-day lags for the most important variables including consumption, production, and selected weather indicators. These lagged variables are particularly important for **tree-based models such as XGBoost**, which do not process sequential input and therefore require explicit historical values to capture temporal dependencies.

All lag features are included in the standardized datasets to ensure compatibility with such classical forecasting methods. However, **they will be excluded during the LSTM modelling stage**, since recurrent neural networks already learn temporal dependencies directly from the sequences of past observations. In this context, explicit lag columns add little informational value and may introduce redundancy or noise.

4.1.6 Data Sources Overview

Table 4.1 summarises the data sources used in Sections 4.1.1-4.1.5 for Germany and Portugal.

Feature	***Portugal***	***Germany***
Net Imbalance	Calculated	Calculated
Production	Ren Data Hub	SMARD
Consumption	Ren Data Hub	SMARD
Import / Export	Ren Data Hub	SMARD
Electricity Prices	Ren Data Hub	ENTSO-E
Weather Data	Open Meteo	Open Meteo
Holidays	Python holidays	Python holidays

Lags	Calculated	Calculated

Table 4.1: Overview of data sources used for Germany and Portugal

4.2 Data Preprocessing

After collecting the raw datasets described in Section 4.1, the data was preprocessed to create harmonized and fully time-aligned 15-minute datasets for Germany and Portugal covering 2015-2025. The preprocessing workflow includes temporal harmonization, data cleaning, resolution alignment, merging of multiple data sources, and feature construction.

4.2.1 Temporal Standardization: Timezones, DST, and Resolution Harmonization

Because the raw datasets originate from multiple platforms, countries, and temporal resolutions, the first preprocessing step was to establish a unified time structure. All timestamps were converted to local time (Europe/Berlin for Germany and Europe/Lisbon for Portugal) with full daylight-saving-time (DST) awareness. This ensured that clock changes - particularly the repeated hour during the autumn DST transition - were handled consistently across all datasets.

A common 15-minute "backbone" timeline was generated for each country, spanning from 2015-01-01 00:00 to 2025-08-31 23:45. All datasets, regardless of their original resolution, were mapped to this backbone. Operational electricity system data from SMARD (Germany) and REN (Portugal) already used 15-minute resolution and therefore required minimal temporal transformation apart from localisation and alignment.

Hourly datasets - including electricity prices (Germany, Portugal) and all meteorological variables from Open-Meteo - were harmonised to the 15-minute backbone by uniformly subdividing each hourly value across the four sub-intervals. This approach assumes constant behavior within each hourly interval, given the lack of higher-resolution data. Weather time series were first aligned in UTC and then converted to local time (Europe/Berlin or Europe/Lisbon) with full DST awareness.

The outcome of this step is a unified, DST-consistent, gap-free 15-minute timeline to which all subsequent data components attach.

4.2.2 Cleaning, Reformatting, and Duplicate Handling

After temporal standardization, the datasets from each domain were cleaned and normalized to ensure structural consistency before merging. This included converting locale-specific numerical formats, such as the German comma-based decimal notation, into standard floating-point values, as well as replacing placeholder symbols (e.g., "-") with zeros in production and import/export datasets where these indicated missing or null entries.

Only exact duplicate rows were removed, as duplicate timestamps can legitimately occur in the data due to system reporting practices or daylight-saving-time transitions. Finally, all datasets were ordered chronologically after transformation.

Together, these cleaning steps ensured structural consistency across sources while keeping the original temporal characteristics of each dataset.

4.2.3 Construction of Country-Level Master Datasets

Once cleaned and temporally aligned, the different data domains were merged into consolidated master datasets for Germany and Portugal. For Germany, consumption, production by technology, physical imports and exports, electricity prices, and weather variables were merged on the common 15-minute time grid using left joins, ensuring that no timestamps were removed. Because the German data sources already followed consistent structures, this step required only minimal adjustment.

For Portugal, the raw data showed greater heterogeneity, particularly in historical production records. Production datasets were therefore standardized prior to concatenation, and missing periods were appended after temporal alignment. Hourly electricity prices and weather variables were then mapped to the 15-minute resolution. As in the German case, only exact duplicate rows were removed, and all unique timestamps were preserved.

The result is a pair of complete, country-level datasets with consistent temporal alignment and full coverage across the study horizon.

4.3 Forecast Targets

The harmonised 15-minute datasets constructed in the previous sections form the basis for defining the forecasting targets used in the modelling stage. The analysis focuses on short-term system behaviour that is directly relevant for operational planning. For each country, three target variables are defined: electricity consumption (MW), electricity production (MW), and net imbalance, calculated as the difference between production and consumption:

$$\text{Net Imbalance}_t = \text{Production}_t - \text{Consumption}_t$$

Forecasts are produced for three operationally relevant horizons commonly used in electricity system analysis: one hour ahead (four 15-minute steps), four hours ahead (sixteen steps), and twenty-four hours ahead (ninety-six steps). For each horizon, the target series is constructed by shifting the original time series forward by the corresponding number of 15-minute intervals.

4.4 Train/Test Splitting

Because electricity system variables exhibit strong temporal dependence and evolving structural patterns, the dataset is split chronologically rather than randomly. The training period spans from 2015 to 2023, while the years 2024-2025 are held back as a strict out-of-sample test set.

This splitting strategy ensures that future information cannot leak into the training process and that model performance is evaluated on unseen data. It also allows structural changes-such as increasing renewable shares, shifting temperature patterns, and different market conditions-to be reflected in test performance. In addition, both the training and test periods are sufficiently long to capture seasonal effects, multi-year variability, and rare events, consequently supporting robust model estimation and meaningful evaluation.

All feature scaling, lag construction, rolling-window calculations, and sequence preparation are performed exclusively using training-period data in order to keep proper out-of-sample evaluation.

5 Methodology

5.1 Forecasting Horizons and Experimental Design

This thesis examines short-term electricity forecasting across two operationally relevant horizons-1 hour and 24 hours ahead-for both Germany and Portugal and for all target variables (consumption, production, and net imbalance). The emphasis on the 1-hour and 24-hour horizons is consistent with the way European electricity markets and system operations are organised. While additional operational relevance can be found in intermediate horizons such as the 4-hour horizon, this thesis includes those results only in the appendix A and B to provide completeness and methodological transparency. The main analysis therefore concentrates on the 1-hour and 24-hour horizons, which are of greatest relevance for business decisions, financial outcomes, and system security.

In the European context, the **24-hour horizon corresponds to the day-ahead market**, a single European auction that determines prices, cross-border exchanges, and generation schedules for the following day. Day-ahead results function as the reference point for the entire operational cycle: they inform unit commitment decisions, determine which thermal generators must be online, define interconnector usage, and establish the baseline against which all intraday deviations are evaluated. Because many generation units have long start-up times, the day-ahead horizon is structurally the most influential for both cost formation and resource adequacy (ENTSO-E, 2023). Consequently, forecasting performance at the 24-hour horizon directly shapes the accuracy of system-wide planning decisions and the economic efficiency of the following operational cycle.

While the 24-hour horizon is for structural planning, the **1-hour horizon** represents the final meaningful opportunity for system operators and market participants to intervene ahead of real-time. In practice, this horizon aligns closely with intraday balancing, redispatch decisions, and last-minute trading. All of these are critical for maintaining frequency stability and preventing imbalance penalties. European intraday markets allow trading up to minutes before delivery, making short-term forecasting essential for market participants who must refine their positions continuously (ACER, 2025). System operators rely on forecasts within this near real-time window to activate balancing reserves, adjust cross-border schedules, and respond to unforeseen deviations in renewable output or load patterns. Review studies on short-term load forecasting consistently mention that forecasting horizons of one hour or less are vital for real-time operational security, network congestion management, and frequency control (Mystakidis et al., 2024; Ullah et al., 2024).

Taken together, the 1-hour and 24-hour horizons capture the two most meaningful and important horizons of system operation-real-time flexibility and day-ahead structural planning-and all models introduced in Section 5.2 are therefore evaluated across both horizons to ensure comparability and operational relevance.

5.2 Model Portfolio and Training Protocol

The forecasting framework in this thesis uses a diverse portfolio of statistical, machine learning, and deep learning models. Each model was selected for its ability to capture different temporal and non-linear patterns in electricity consumption, production, and net imbalance. All models are trained under a unified experimental design to ensure comparability across Germany and Portugal, across all target variables, and the forecasting horizons. The following paragraphs describe the models used and the respective training procedure applied to each class.

Classical statistical models provide a baseline for comparison with more advanced methods. An ARIMA model is included as a statistical benchmark, reflecting the traditional forecasting framework based on linear autoregressive and seasonal structures. Its performance is generally inferior in modern short-term forecasting, as electricity demand exhibits non-linear and non-stationary behavior that such models cannot capture adequately. (Z. Zhao et al., 2021).

The next tier consists of machine learning regressors. XGBoost is used as the primary model in this category because of its strong empirical performance in energy forecasting and its ability to model nonlinear interactions between weather, time-of-day factors, and lagged consumption or production values. Random Forest acts as a non-boosted comparison, while LASSO regression provides a linear method to benchmark against more flexible models. These algorithms operate on engineered feature matrices containing lagged target variables, calendar encodings, weather indicators, rolling statistics, and market-related features. All machine learning models are trained using a consistent horizon-specific setup: separate models are fitted for the 1h and 24h ahead predictions, with the forecasting target constructed by shifting the time series according to the number of 15-minute steps corresponding to each horizon. Additionally, machine learning methods have often shown that they can match or even surpass deep learning models in various energy forecasting applications. This reinforces their relevance as competitive alternatives (Benti et al., 2023; Plakas et al., 2023b; Shahiduzzaman et al., 2021).

Deep learning models form the third component of the portfolio and are used to directly model temporal sequences rather than engineered lag vectors. LSTMs and GRUs are included because of their strong performance and increasing use in energy forecasting, where higher shares of renewable generation have introduced non-linear and volatile patterns that can be effectively modeled with sequence-based approaches. Their rising prominence in the literature is documented in recent review work (Hong et al., 2020; Mystakidis et al., 2024).

Multiple LSTM variants are trained, including multi-output configurations, dual-pipeline architectures for consumption and production, and models with lag-augmented features. GRU networks are used as computationally lighter alternatives. To ensure comparability across models, the input sequences for all recurrent networks are standardised following the preprocessing in Section 4, and the same horizon-specific label generation used in the machine learning models is applied to the sequence targets.

Finally, the portfolio includes hybrid convolutional-recurrent models, specifically 1D-CNN × LSTM and 1D-CNN × GRU architectures. These models first use convolutional layers to extract short-term temporal features-such as ramps in consumption, sudden renewable fluctuations, or morning/evening load peaks-before applying recurrent layers to capture longer-range temporal dependencies. Empirical evidence in the literature shows that CNN-LSTM hybrids frequently outperform both classical and standalone deep learning models in short-term electricity forecasting, particularly under volatile and weather-driven conditions (Al-Ali et al., 2023; Ullah et al., 2024).

Reviews of renewable forecasting techniques similarly identify hybrid deep learning methods as a state-of-the-art approach for managing nonlinear and chaotic energy patterns.
These findings justify their inclusion for both consumption/production forecasting and for net imbalance forecasting, where accurate representation of rapid renewable changes is essential.

All models are trained using a consistent evaluation methodology based on time-series cross-validation. For the machine learning models, a five-fold expanding-window structure is used, ensuring that validation folds always occur strictly after training folds. For deep learning models, training and validation splits follow the same chronology, and training is conducted with early stopping based on validation loss to avoid overfitting. Hyperparameter tuning for XGBoost is performed through random search over a predefined parameter space, using cross-validated mean absolute error as the selection criterion.

Random Forest and LASSO were also tuned using standard parameter settings. Deep learning models are trained on GPU acceleration with fixed hyperparameters selected through exploratory experiments documented in the model notebooks.

The final performance evaluation of all models is conducted on the 2024-2025 out-of-sample test set. Net imbalance forecasts are generated by subtracting separately trained consumption and production forecasts. This design mirrors operational practice, where load and generation are forecast independently using distinct models and information sets, and imbalance emerges as a residual rather than a primary forecast target. While this approach introduces uncertainty from both components, it preserves the structural separation required for the following severity analysis and interpretability. Because all models are trained and evaluated in the same way, any differences in results come from the models themselves - not from differences in how they were trained or tested.

5.3 Evaluation Metrics

A reliable comparison between Germany and Portugal requires consideration of the German electrical system's size, which is approximately ten times larger than Portugal's. Consequently, absolute errors like MAE or RMSE are inappropriate for cross-country performance ranking, as they would inflate the forecasting challenge for the larger system. We therefore choose the Mean Absolute Percentage Error (MAPE) as the primary evaluation and model selection metric for the individual Consumption and Production forecasts. MAPE is a scale-independent measure, allowing us to compare relative forecast difficulty consistently between the two countries. Furthermore, a symmetric measure is appropriate since forecast deviations in either direction for these targets have comparable operational consequences.

MAPE is not suitable for evaluating net position forecasts because this variable can take values close to zero, which makes percentage errors unstable and difficult to interpret. For this reason, the evaluation of net position forecasts focuses on absolute error measures and their operational implications. This is done using two complementary approaches.
First, absolute forecast errors measured in megawatts (MW) are assessed using metrics such as MAE and RMSE, which provide an indication of the volume of balancing reserves required to correct forecast deviations. Second, a custom severity-based evaluation framework is applied, as described in Section 6.2. This framework classifies absolute forecast errors into operationally meaningful categories - minor, significant, severe, and critical - based on their potential impact on system operation. By scaling these categories relative to each country's physical characteristics, such as reserve capacity and system inertia, the analysis focuses on forecast errors that create genuine operational risks rather than simply minimizing average statistical error.

5.4 SHAP-Based Explainability

While forecasting accuracy is the primary objective of the modelling framework, understanding why models behave the way they do is also essential, particularly in the context of electricity system operation where model outputs can influence financially and operationally important decisions. Modern machine learning and deep learning models often function as non-linear black boxes, which makes a systematic interpretability approach necessary. For this purpose, the thesis employs SHAP (SHapley Additive exPlanations) values to analyse the contribution of individual features to model predictions.

SHAP provides both global and local interpretability by quantifying each feature's marginal contribution relative to a baseline prediction. This allows examination of which variables shape the forecasts on average, as well as how specific inputs influence individual predictions. Such insights are valuable for diagnosing model behaviour during normal conditions and during periods of unusual system behaviour, such as rapid renewable fluctuations or extreme weather events.

In line with the evaluation strategy of selecting models exclusively based on MAPE, SHAP analysis is conducted **only for the best-performing model** for each forecasting horizon and target variable. Since this is the model that would be used in practice and that forms the basis for all subsequent conclusions, its internal decision structure is the only one that requires detailed interpretability assessment. For tree-based models, exact SHAP values are computed; for deep learning models, gradient-based approximations are used to obtain consistent feature attributions.

The SHAP analysis also supports cross-country comparison. Because the same modelling framework and feature set are applied to both Germany and Portugal, differences in SHAP value distributions can be interpreted as genuine differences in system behaviour rather than artefacts of model design-for example, a stronger influence of wind speed in Germany or a more pronounced contribution of hydropower-related variables in Portugal. Examining how feature relevance changes between the 1-hour and 24-hour horizons further clarifies the shift from short-term dynamics to structural daily patterns.

In summary, SHAP-based explainability provides a transparent view into the strongest-performing model's behaviour and supports a deeper understanding of the physical and operational drivers behind forecast accuracy across countries, targets, and horizons. To ensure interpretability and comparability, SHAP analyses are applied to fixed-parameter XGBoost models with a reduced feature set that excludes direct autoregressive production and consumption terms, thereby avoiding dominance by lagged variables and allowing weather, calendar, market, and generation-related drivers to be assessed consistently across systems.

6. Results

6.1 Forecast Accuracy

6.1.1 General Observations Across All Models

Across the full set of experiments, several consistent patterns emerge. First, the ARIMA baseline performs significantly worse than all other models. This result is unsurprising, as ARIMA is a linear univariate method that does not incorporate critical exogenous information such as weather variables, holiday effects, market behavior, or nonlinear interactions. Its role is therefore limited to a simple benchmark.

Second, tree-based machine learning models outperform regression-based and most deep learning methods. XGBoost consistently delivers the lowest MAPEs, followed by Random Forest, while LASSO regression performs notably worse due to its linear structure.

Third, single-output deep learning architectures clearly outperform their multi-output versions, and lag-augmented or CNN-hybrid variants do not improve accuracy. This suggests that the temporal structure of these national-level series is efficiently captured by simple recurrent layers without requiring added convolutional complexity.

Across all horizons and models, two broader empirical regularities emerge. Consumption is consistently easier to forecast than production, while Portugal shows higher MAPEs than Germany. These differences likely reflect structural characteristics of the respective electricity systems. Germany's larger and more integrated grid results in smoother consumption and production trajectories, offering more stable temporal patterns for forecasting. In contrast, Portugal's system is smaller and more strongly influenced by weather-dependent renewable generation, making production patterns more volatile and inherently harder to predict. Appendix B provides the complete set of performance tables and detailed plots for all models evaluated across countries, targets, and forecasting horizons.

6.1.2 1-Hour Horizon Specific Analysis

The 1-hour horizon represents the most immediate operational forecasting situation. As shown in Figure 6.1, XGBoost performs best across both targets and countries, achieving MAPEs between **0.97%** and **7.55%**. Random Forest performs slightly worse but remains competitive, while the best deep learning models (LSTM-Single and GRU-Single) exhibit somewhat higher errors at this short horizon. These results are consistent with the findings of Song et al. (2025), who report low single-digit MAPE values for very short-term electricity load forecasting using an XGBoost-based approach, suggesting that the performance observed in this study is reasonable.

A notable characteristic of the 1-hour results is that **consumption forecasts are substantially more accurate than production forecasts**, across all models and in both countries. The discrepancy is particularly clear in Portugal. This pattern reflects the comparatively smooth and behavior-driven nature of consumption versus the much more volatile generation dynamics, which are strongly influenced by short-term weather conditions, renewable availability, and system balancing needs.

Similarly, **Germany achieves lower errors than Portugal across all models**, a pattern that is already visible at this short horizon. This suggests that recent information has higher predictive stability in Germany, whereas Portugal's system experiences more abrupt fluctuations. These 1-hour findings illustrate that engineered feature sets (such as lags, calendar variables, and recent system states) provide substantial predictive value at short horizons-value that tree-based models are particularly effective at exploiting.

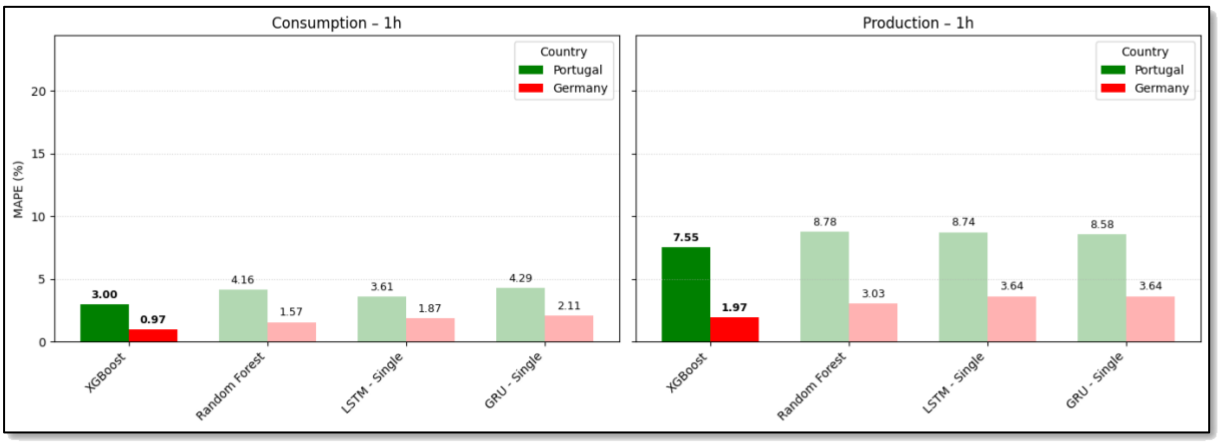


Figure 6.1: One-hour-ahead forecasting performance for electricity consumption and production in Germany and Portugal

6.1.3 24-Hour Horizon Specific Analysis

At the 24-hour horizon, accuracy decreases noticeably for all models, confirming that **longer forecast horizons lead to worse performance**. This decline reflects the reduced usefulness of recent observations and the growing influence of factors that are not directly captured by the historical time series.

Despite this decrease, **XGBoost remains the best-performing model**, again outperforming Random Forest and all deep learning variants. For Portugal, consumption errors increase to around 6-7%, while production reaches 17-21%. Germany once again achieves lower errors, with XGBoost producing 3.52% consumption MAPE and 11.25% production MAPE.

The same two systematic patterns remain strongly at the 24-hour horizon: production continues to be harder to forecast than consumption, especially in Portugal. Additionally, Portugal remains more challenging to predict than Germany, even at longer horizons.

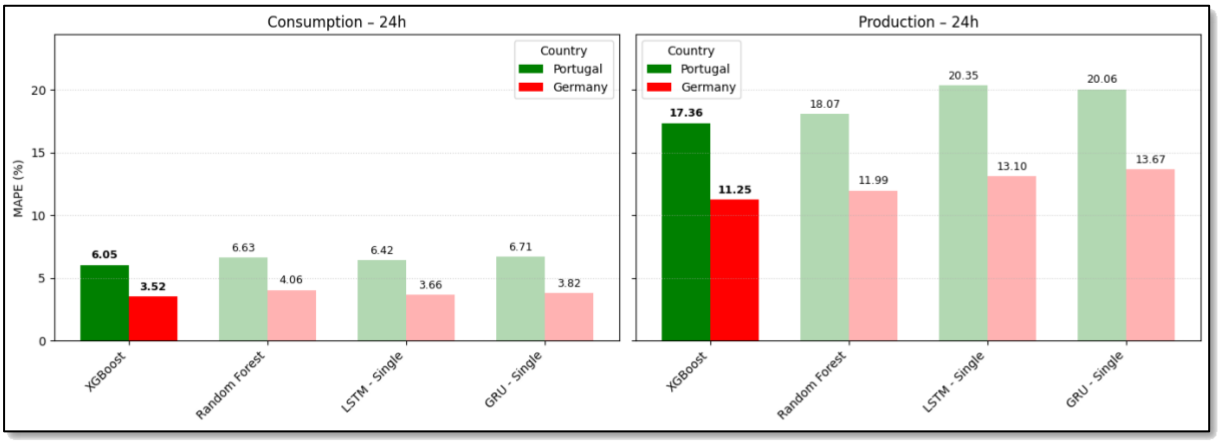


Figure 6.2: Twenty-four-hour-ahead forecasting performance (MAPE) for electricity consumption and production in Germany and Portugal

6.1.4 Deep Learning Performance and Horizon Sensitivity

Deep learning models generally perform worse than the tree-based machine learning models across all horizons, and this performance gap widens as the forecast horizon increases. One explanation is that recurrent models depend on short-term sequential patterns that lose relevance at longer horizons, whereas tree-based models rely on lagged and structured features that continue to provide useful information.

Although the dataset has high temporal resolution, it does not offer the large training volume typically needed for deep learning models to fully exploit their capacity. This increases the risk of overfitting and can limit the ability of these models to learn stable temporal patterns, particularly for multi-output and CNN-based architectures, which consistently perform worse than simpler single-output recurrent models.

Taken together, these results highlight the need to examine the underlying drivers of forecasting performance more closely, motivating the deeper analyses presented in the remainder of Section 6

6.2 Net Imbalance Forecasting and Severity Analysis

The accuracy metrics presented above provide a useful summary of predictive performance, but they do not capture the operational implications of forecast errors. Errors of similar statistical magnitude can have very different consequences depending on system size, operating conditions, and available balancing resources. For this reason, the following analysis complements standard accuracy measures with a severity-based classification that links forecast errors to their potential operational relevance.

6.2.1 Operational Motivation for Severity Thresholds

Forecast errors in net imbalance forecasting need to be evaluated in terms of their operational consequences, which increase non-linearly as the power system moves away from normal operating conditions. In European transmission system operation, imbalance severity is therefore assessed relative to system load and current operating conditions rather than using fixed absolute thresholds. This approach is reflected in ENTSO-E operational guidelines and is supported by both system stress tests and observed system events.

The ENTSO-E **Incident Classification Scale (ICS)** (ENTSOE-E, 2019) establishes a clear percentage-of-load logic for disturbance classification, defining deviations below approximately 1% of system load as minor events within normal operational tolerance, while larger deviations are classified as reportable incidents due to their potential impact on system security. This distinction reflects the transition from routine automatic control to conditions in which active balancing actions become necessary.

In addition to this classification framework, the ENTSO-E **System Defence Plan** (ENTSOE-E, 2022) evaluates system behaviour under increasingly larger relative imbalances. These studies show that deviations in the range of approximately 1-5% of system load already place the system in an alert state. Under such conditions, extensive reserve activation is required, frequency volatility increases, and operational costs rise. Beyond this range, system stability increasingly relies on emergency defence mechanisms rather than standard balancing actions. Imbalances around 10% of system load are considered extreme scenarios, in which escalating failures and loss-of-load events become realistic if corrective actions are delayed or constrained.

Observed system events closely mirror this escalation logic. In Germany, the June 2019 wind forecast error events demonstrated that even before defence mechanisms are in use, large relative imbalances rapidly exhaust balancing reserves and necessitate coordinated international support, highlighting the fragility that appears well below blackout conditions (50Hertz et al., 2019). In smaller and lower-inertia systems such as Portugal, this transition occurs at significantly lower absolute levels. ENTSO-E analyses of Iberian disturbances show that relative deviations of only a few percent of national load can already produce steep frequency excursions and require coordinated cross-border balancing, particularly under high renewable penetration and low-inertia conditions (ENTSO-E, 2022; IEA, 2022).

Together, ENTSO-E classification frameworks, system defence stress tests, and historical incidents point to **distinct operational states** defined by relative imbalance magnitude: a normal state below 1%, a stressed but controllable state between approximately 1-5%, a defence-level state between 5-10%, and a critical state beyond 10%, where blackout becomes a realistic risk. These states motivate the relative severity categories adopted in this thesis and provide an operationally grounded basis for interpreting forecast errors across systems with different physical characteristics.

6.2.2 Classification of Severity Thresholds and Operational Impact of Mispredictions

Based on the operational considerations outlined in Section 6.2.1, net imbalance forecast errors are classified using a **relative severity framework expressed as a percentage of**

contemporaneous system load. This formulation ensures that forecast errors are evaluated in proportion to prevailing system conditions and allows consistent interpretation across countries with markedly different system sizes, inertia characteristics, and reserve structures.

Minor deviations, defined as forecast errors below **1%** of system load, remain within the normal operating state of the power system. Errors in this range are absorbed through routine automatic control actions, primarily via frequency containment and restoration reserves, without significantly affecting system security or market outcomes.

Significant deviations, defined as forecast errors between approximately **1%** and **5%** of system load, go beyond routine tolerance and require active balancing actions, including increased reserve activation and redispatch. Although overall system stability is usually maintained, these conditions are associated with higher operational costs, greater frequency variability, and potential market effects such as price volatility or increased redispatch.

Severe deviations, covering forecast errors in the range of roughly **5%** to **10%** of system load, place the system under considerable stress and approach the limits of normal operational controllability. Under such conditions, stability increasingly depends on defence-level mechanisms rather than standard balancing processes, and operational margins become thin. The system becomes more sensitive to delays, forecast errors, or simultaneous disturbances, increasing the likelihood of escalating effects.

Critical deviations, defined as forecast errors exceeding approximately **10%** of system load, fall outside the range of normal and defence-level operation. Under these conditions, outcomes such as blackout, system separation, or widespread loss of load become plausible rather than unlikely, especially when emergency actions are delayed, limited, or occur during stressed system conditions.

These severity categories represent escalating states of operational risk rather than deterministic outcomes. A given forecast error does not imply a fixed frequency response or guaranteed system failure; instead, the classification reflects the likelihood that increasingly disruptive control actions will be required to maintain system stability. By using this relative severity framework, the analysis directly links statistical forecast errors to their potential operational consequences, allowing a risk-aware evaluation of net imbalance forecasting performance.

6.2.3 Net Imbalance Forecast Model (XGBoost)

The XGBoost model is used to forecast **net imbalance**, defined as the residual between independently predicted electricity consumption and electricity production. Net imbalance is not a directly observed physical quantity but an emergent system-level outcome that is created from the interaction of demand and generation forecasts. Consequently, its forecast quality depends on the accuracy of both components, and uncertainty carries over from both sides into the imbalance estimate. Errors in consumption and production can either offset or reinforce each other, depending on their direction and timing. This introduces a structural source of uncertainty that arises naturally in real-world forecasting pipelines.

At an aggregate level, the model captures short-term system dynamics with reasonable accuracy in both countries. For short lead times, forecast errors remain relatively concentrated, and the majority of deviations fall within the minor or significant severity regimes defined in Section 6.2.2. Germany shows a comparatively thin error distribution, consistent with its large system size, high inertia, and more stable aggregate demand and generation patterns. Portugal a substantially wider distribution of net imbalance errors. This reflects stronger sensitivity to weather-driven renewable generation, greater local variability, and more limited internal balancing capability. These structural characteristics increase the volatility of the net imbalance residual and reduce forecastability, even when identical model architectures and feature sets are applied.

Forecast uncertainty increases systematically with horizons in both systems, but the effect is stronger in Portugal. Day-ahead forecasts rely more heavily on weather predictions, behavioural uncertainty, and renewable output variability, which widens the distribution of forecast errors. While this decline in average accuracy is expected, it does not fully capture the operational relevance of forecast errors. As discussed in Sections 6.2.1 and 6.2.2, errors of similar statistical magnitude can have fundamentally different consequences depending on whether they remain within routine operational tolerance or push the system toward defence-level or critical operating regimes.

For this reason, the evaluation of net imbalance forecasts in this thesis goes beyond aggregate accuracy metrics and explicitly considers the **severity distribution of forecast errors**. The following section therefore examines the joint relationship between realised net imbalance and forecast error using scatter plots overlaid with relative severity thresholds. This approach allows us a more nuanced assessment of how forecast deviations evolve as the system approaches increasingly sensitive operating conditions and provides direct insight into the operational risk implications of net imbalance forecasting performance across countries and horizons.

6.2.4 Forecast Error Scatter Plots and Severity Bands

To assess how forecasting performance varies across operating conditions, net-imbalance forecast errors are plotted against the true net imbalance for both horizons and both countries. Severity bands are defined relative to system load (Section 6.2.1), reflecting the ENTSO-E percentage-of-load framework rather than fixed absolute thresholds. This allows forecast deviations to be interpreted in terms of relative system stress and potential operational relevance under different system states.

These severity bands are not intended as operational alarms, but as an analytical tool to compare how identical forecast errors result in system stress under different structural conditions.

1-Hour Horizon

At the 1-hour horizon, Germany exhibits a tightly concentrated error distribution centered around zero across the full range of net imbalance values. Approximately **34 %** of forecast steps fall into the minor category (<1 % of load), while **61 %** are classified as significant (1–5 %). Severe and critical deviations are rare, accounting for **4.9 %** and **0.2 %** of observations. Even during periods of relatively large net imbalance, forecast errors remain proportionally small relative to system load, and the scatter plot shows no systematic skew or increase of errors.

This behaviour reflects the structural characteristics of the German power system: large scale, high synchronous inertia, diversified generation, and substantial internal balancing capability. Short-term fluctuations are smoothed through aggregation, and the model benefits from strong temporal persistence at this horizon.

Portugal shows a different pattern at the same horizon. Only **13 %** of forecast errors fall within the minor band, while **44 %** are significant, **28 %** are severe, and **15%** are already classified as critical relative to contemporaneous load. The scatter plot shows a visibly wider vertical range, even at moderate net imbalance levels, indicating that relatively small absolute forecast errors can lead to significant relative system stress in a smaller, lower-inertia system.

Importantly, this contrast appears even though an identical modelling pipeline, feature set, and evaluation framework are used. The higher occurrence of severe and critical classifications in Portugal therefore reflects system characteristics rather than model deficiencies: higher renewable shares, reduced inertia, stronger dependence on weather-driven generation, and limited internal balancing options all increase sensitivity to forecasting inaccuracies.

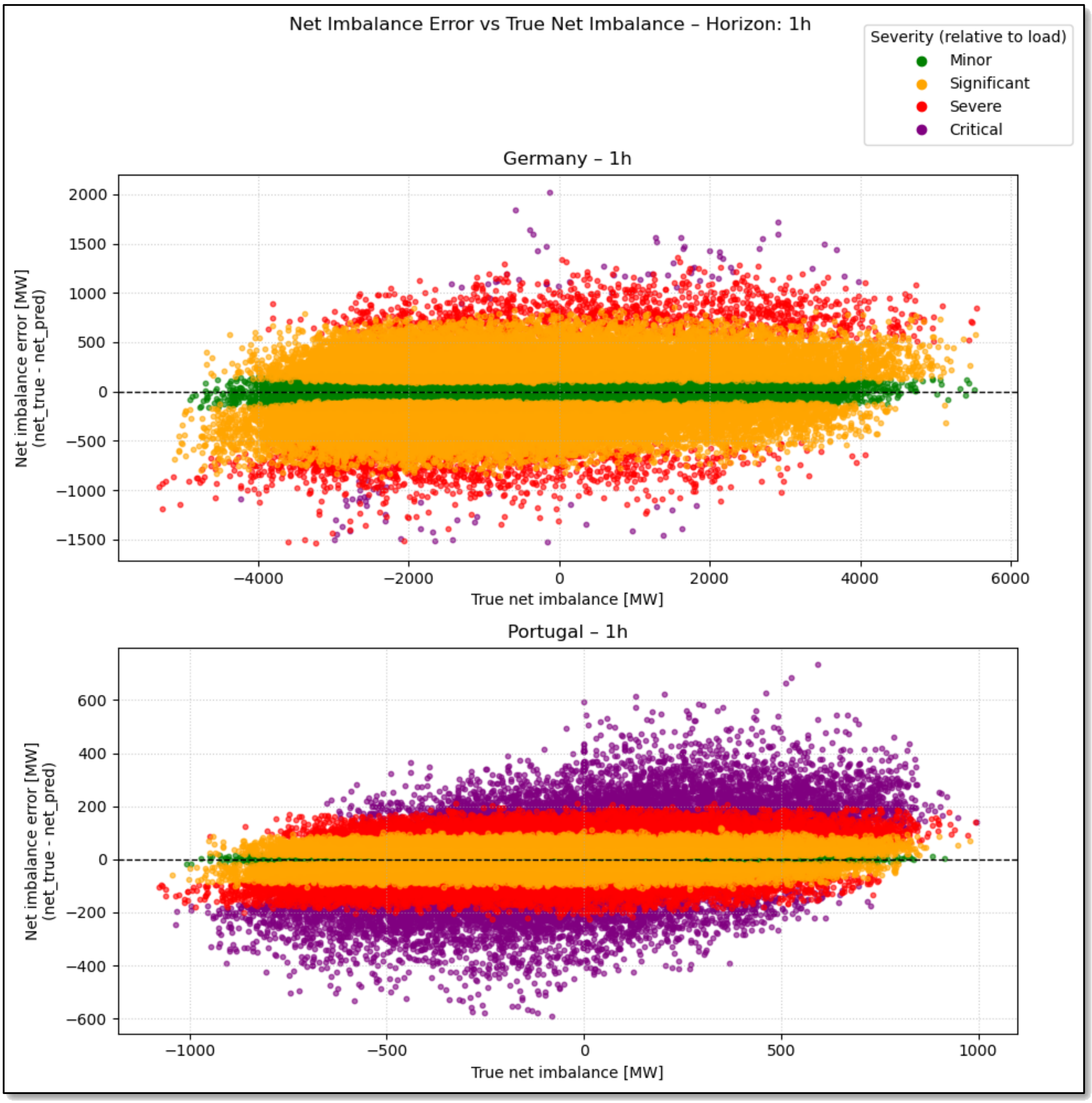


Figure 6.3: Relationship between net imbalance forecast error and true net imbalance at the one-hour horizon, with severity classification for Germany and Portugal

24-Hour Horizon

At the 24-hour horizon, forecast uncertainty increases substantially in both systems, as expected for day-ahead prediction. In Germany, the distribution shifts toward higher severity levels: minor errors fall to **6 %**, significant errors account for **25 %**, severe errors rise to **27 %**, and critical deviations reach **42 %**. The scatter plot displays a funnel shape, with forecast error magnitude increasing alongside absolute imbalance, reflecting scale-dependent uncertainty and reduced informational value of recent system states.

Despite this widening distribution, the error cloud remains generally symmetric around zero, indicating no systematic directional bias in day-ahead forecasts. In particular, the high share of critical classifications at the 24-hour horizon reflects the scaling of relative error under a static, out-of-sample evaluation and does not imply that the German system routinely operates near emergency conditions.

Portugal's 24-hour horizon reveals a more extreme escalation. Minor errors fall below **5 %**, while severe and critical deviations together account for over **75 %** of all forecast steps (**22 % severe, 54 % critical**). The scatter plot shows a strong linear expansion of relative error with increasing imbalance magnitude, particularly under positive imbalance conditions. This pattern highlights the difficulty of anticipating day-ahead residual imbalances in a smaller, renewable-dominated system with limited buffering capacity.

Comparing the two systems, the asymmetry is strong. Germany experiences a decrease in relative forecast quality, yet still benefits from structural resilience that reduces operational consequences. Portugal, under the same modelling assumptions, shows a much higher share of high-severity classifications, indicating greater exposure to forecasting uncertainty as the horizon increases.

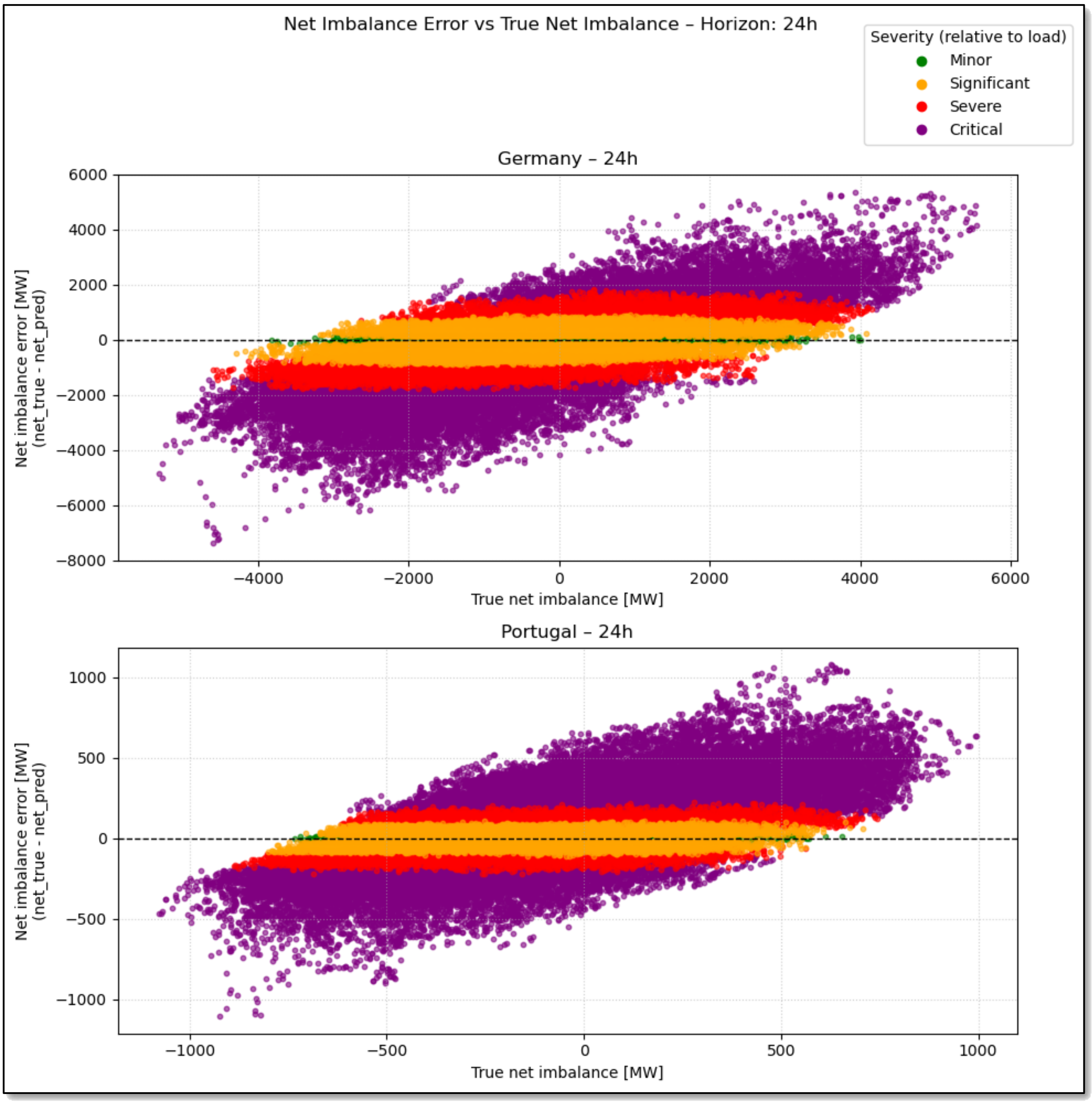


Figure 6.4: Relationship between net imbalance forecast error and true net imbalance at the twenty-four-hour horizon, with severity classification for Germany and Portugal

General Remark

It is important to emphasize that the frequency of severe and critical classifications in this analysis **does not imply that either system routinely operates near blackout conditions**. Severity categories are defined relative to current load and are used here as a comparative stress metric under a fixed forecasting pipeline.

In operational practice, TSOs continuously refine forecasts, deploy probabilistic tools, update schedules intraday, and rely on layered reserve mechanisms. Under such conditions, the absolute number of high-severity outcomes would be substantially lower than those implied by a static out-of-sample evaluation.

However, because Germany and Portugal are evaluated using **identical models, features, horizons, and evaluation procedures**, the relative differences in severity distributions are robust and highly informative. Even under improved forecasting accuracy, the same qualitative pattern would persist: Germany remains structurally easier to forecast and produces fewer high-severity deviations, while Portugal's smaller size, higher renewable share, and lower inertia make it inherently more sensitive to forecast errors-especially at longer horizons. This robustness is further supported by the upper- and lower-bound sensitivity analysis presented in Appendix C, which confirms that the relative severity patterns and cross-country contrasts remain qualitatively unchanged under alternative threshold assumptions.

These findings motivate the severity-conditioned interpretability analysis in Section 6.3, where the drivers of high-risk forecast deviations are examined in detail.

6.2.5 SHAP-Based Interpretation of Forecast Drivers

To understand the mechanisms driving the net-imbalance forecasts and identify which factors most strongly influence model behaviour, SHAP values were computed for the consumption and production models of both countries at both horizons. The underlying model remains the same XGBoost configuration used throughout this chapter; only the feature set was simplified to improve interpretability. Autoregressive and aggregated variables were removed so that the explanations reflect direct physical, temporal, and market drivers rather than internal model summaries. The resulting SHAP attributions therefore highlight how meteorological variability, calendar structure, and market signals shape the forecast without obscuring effects from highly correlated lag features.

A consistent pattern emerges across all model configurations. Consumption is structurally easier to forecast than production because it exhibits strong daily, weekly, and seasonal regularities that are well captured by time encodings. Production, in contrast, is shaped by volatile renewable generation, which introduces substantially higher variance. This asymmetry becomes particularly visible in Portugal, where the share of wind and solar generation is large and the system is small enough that even local fluctuations influence the aggregate production forecast.

Germany 1-Hour Horizon

The German 1-hour SHAP values reveal a highly regular short-term structure. Consumption forecasts are dominated by temporal features such as time of day and day of week, complemented by modest contributions from temperature and shortwave radiation. Production forecasts also show strong temporal patterns, together with stable contributions from dispatchable technologies such as natural gas and coal and a moderate role for wind generation. Renewable variability does affect the model, but its influence remains comparatively small due to Germany's large and diversified system, which smooths local meteorological fluctuations.

These SHAP patterns align with the narrow error distribution observed earlier: short-term imbalance forecasting in Germany is structurally stable, and large deviations occur mainly under unusual operating conditions. No single feature displays disproportionate volatility, and the overall explanatory structure stays consistent across both components of the forecast.

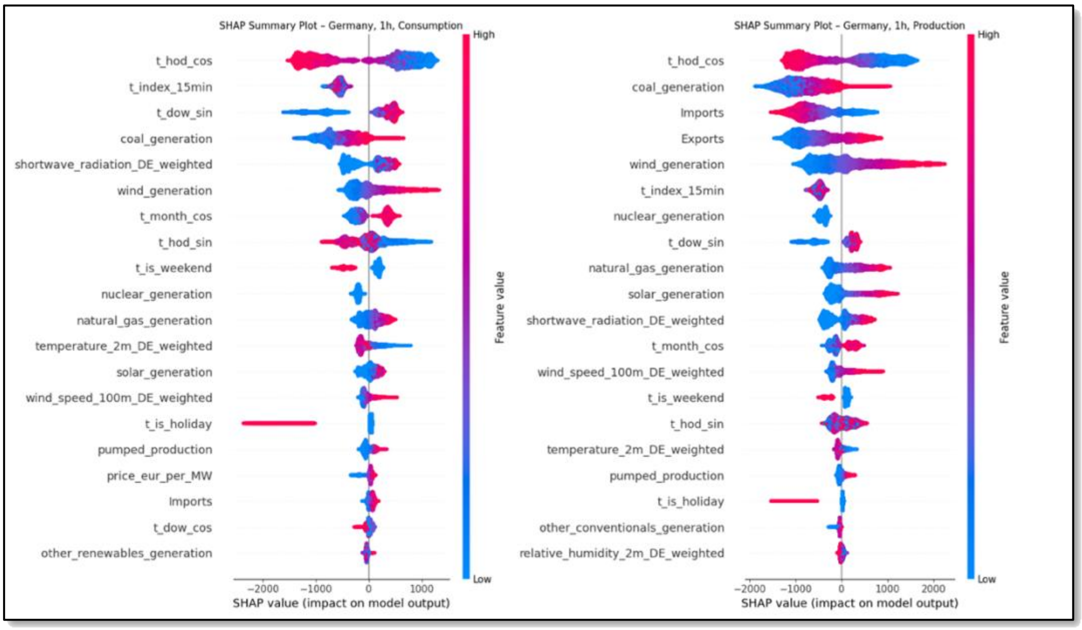


Figure 6.5: SHAP summary plots for the one-hour-ahead consumption and production forecasting models in Germany

Portugal 1-Hour Horizon

Portugal exhibits a much more meteorologically sensitive SHAP structure even at the 1-hour horizon. While daily timing patterns remain relevant, renewable-related variables such as wind generation, solar generation, hydropower inflows, and shortwave radiation have a substantially stronger influence on the model. The SHAP distributions for these features are broad and variable, indicating that the forecast is highly responsive to short-term renewable fluctuations. The production model relies more heavily on weather-driven inputs than the German model.

These patterns correspond directly to the wider scatter and higher variance described in Section 6.2.4. They show that Portugal's net-imbalance forecast is significantly more exposed to short-horizon volatility, not because of model limitations but because meteorological conditions materially affect both consumption and production in a system of this size.

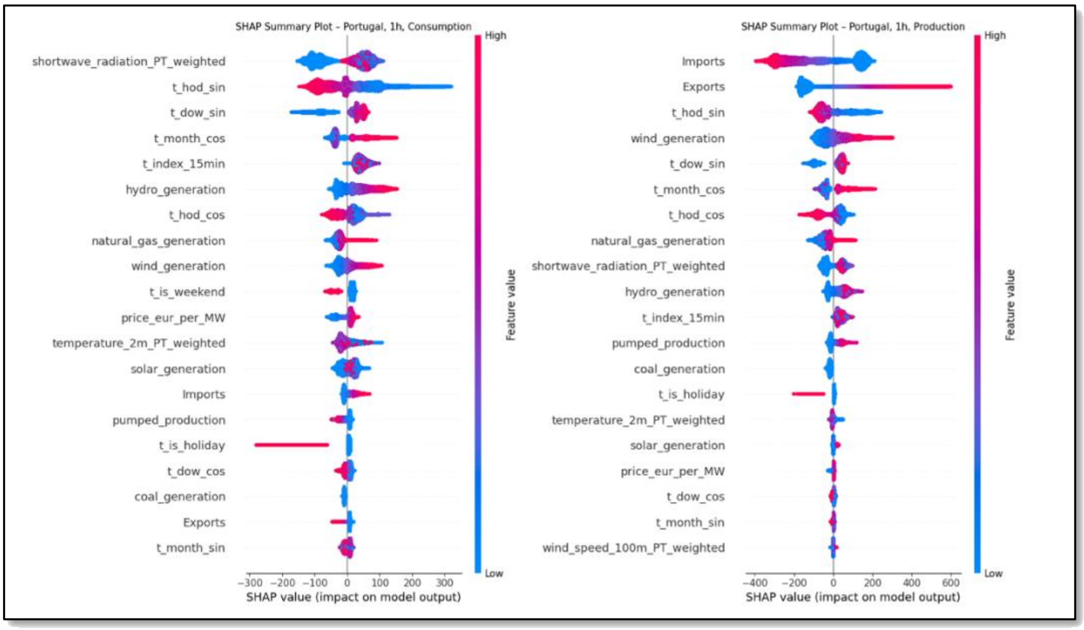


Figure 6.6: SHAP summary plots for the one-hour-ahead consumption and production forecasting models in Portugal

Comparison - Germany vs Portugal (1h)

The contrast between the two countries at the short horizon follows logically from their structural characteristics. Germany's larger and more diversified system reduces the impact of renewable variability and maintains a stable explanatory structure dominated by temporal and thermal variables. Portugal's system, in contrast, reacts strongly even to moderate changes in renewable output, which is reflected in the magnitude and spread of its SHAP values. This means that the same forecasting model, applied to otherwise identical feature sets, encounters fundamentally different predictability environments due to differences in system size, renewable penetration, and internal balancing capability.

Germany 24-Hour Horizon

At the 24-hour horizon, Germany's SHAP distributions widen, as expected from the increased uncertainty of day-ahead forecasting. Temporal encodings continue to structure the forecast, but renewable-related features such as wind generation, solar generation, and radiation gain greater importance. Temperature becomes more influential for the consumption model, reflecting the sensitivity of day-ahead load to meteorological conditions. Although renewable variables introduce more spread into the SHAP values, their influence remains balanced against contributions from dispatchable generation, which stabilises overall system behaviour. The explanatory patterns remain consistent, and the model does not display notable shifts in feature dominance.

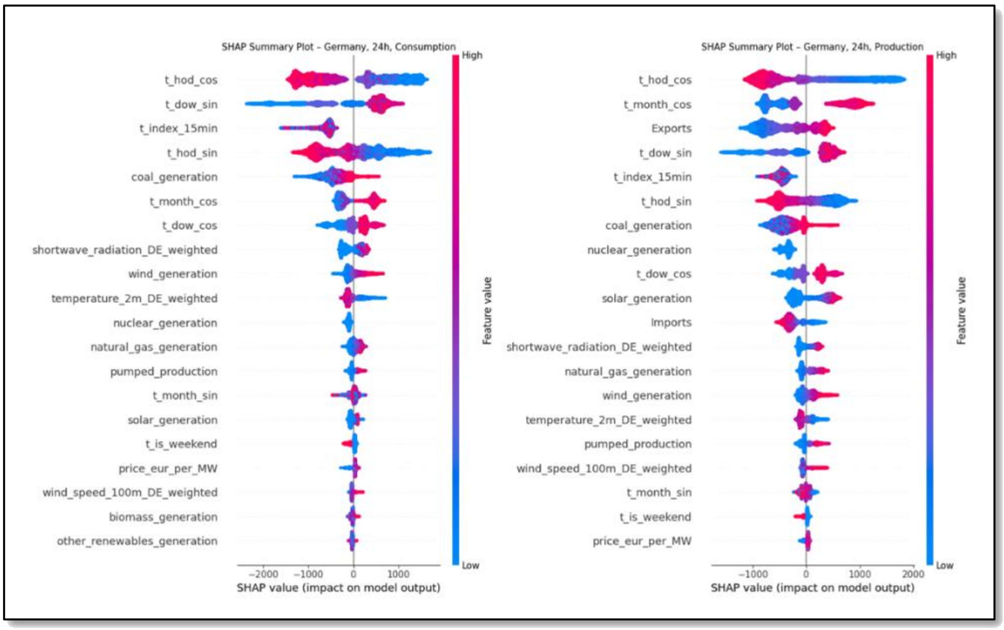


Figure 6.7: SHAP summary plots for the twenty-four-hour-ahead consumption and production forecasting models in Germany

Portugal 24-Hour Horizon

Portugal's 24-hour SHAP structure is dominated by meteorological inputs to an even greater extent than at the 1-hour horizon. Wind and solar generation, radiation, temperature, and hydro-related variables all show large and highly variable SHAP magnitudes, indicating that day-ahead forecast uncertainty is almost entirely shaped by renewable-driven fluctuations. Imports and exports become more prominent than in Germany, which is consistent with Portugal's need to rely on cross-border balancing when renewable forecasts error substantially.

The intensity of the SHAP values is considerably higher than in Germany, and the renewable variables often dominate the model output. These patterns mirror the markedly wider scatter and heavier-tailed error distribution described in Section 6.2.4 and confirm that renewable variability is the primary driver of Portuguese forecast uncertainty at the day-ahead horizon.

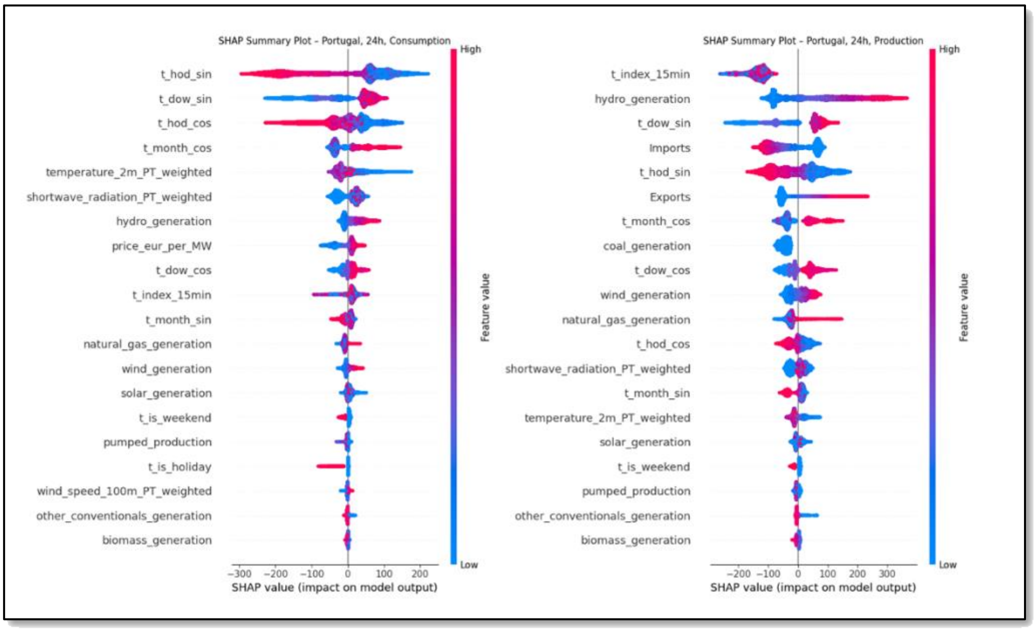


Figure 6.8: SHAP summary plots for the twenty-four-hour-ahead consumption and production forecasting models in Portugal

Comparison - Germany vs Portugal (24h)

At the day-ahead horizon, the differences observed at 1 hour become clearer. In Germany, explanatory importance is spread across temporal, thermal, and renewable features, indicating a system that can absorb forecast uncertainty across multiple dimensions. Portugal's SHAP structure, by contrast, becomes almost entirely driven by renewable variability. The relative and absolute influence of wind and solar generation increases, and the model's output becomes far more sensitive to meteorological uncertainty. These results indicate that Portugal's higher error variance reflects system characteristics rather than differences in modelling.

Synthesis

Taken together, the SHAP analyses show that the mechanisms behind net-imbalance forecasting differ significantly between the two countries. Germany's large and diversified system maintains stable predictive relationships even under day-ahead uncertainty, while Portugal's strongly renewable-driven system introduces meteorological variability into both production and consumption forecasts. The difference in predictability between consumption and production reflects basic structural characteristics of the system: consumption follows regular temporal patterns, while production depends more strongly on variable renewable output. As shown in Appendix D, focusing the analysis on severe and critical forecast errors makes these patterns even clearer. In Portugal, the most extreme deviations are largely linked to large renewable forecast errors, while in Germany they are associated with more isolated and unusual system conditions. This interpretation is further reinforced when conditioning the SHAP analysis exclusively on severe and critical forecast errors (Appendix D), where the same country-specific and production-driven mechanisms not only remain but become more clearer.

This confirms that the severity patterns documented earlier are not random but driven by the physical characteristics of the two systems. The SHAP analysis therefore provides a clear explanation for why Portugal exhibits a disproportionately large share of high-severity forecast deviations and why renewable-driven systems show significantly greater forecasting challenges.

6.2.6 Cross-Country Interpretation of Forecast Challenges

The analysis shows clear differences in forecast error behavior between Germany and Portugal, both in the size of errors and their severity. These differences remain even when the same model and feature set are used, indicating that the observed difference is mainly driven by structural characteristics of the power systems rather than by modelling choices.

A central finding of the SHAP-based feature attribution is the notable role of renewable production variables, particularly wind and solar generation, in shaping the forecast errors in Portugal. Their relative contribution is consistently higher than in Germany across both short-term (1h) and day-ahead (24h) horizons. This aligns with Portugal's heavier reliance on weather-driven renewables, which accounted for over 60% of electricity generation in recent years (IEA, 2022). Such sources introduce greater variability and uncertainty into net imbalance forecasting, especially during volatile weather conditions. In contrast, Germany's more diversified generation mix and broader geographical scope help reduce these fluctuations to some extent.

Additionally, the SHAP analysis shows that in Portugal, production-side variables play a larger role in driving large forecast errors than consumption-related variables. In Germany, this balance is more even. This difference highlights the greater difficulty of forecasting renewable generation compared to electricity demand, particularly in smaller systems with limited spatial and temporal averaging. These dynamics become even clearer when focusing only on severe errors. In this subset, renewable production explains a larger share of model-attributed variance in Portugal than in the full sample, supporting the link between forecast uncertainty and renewable variability

However, not all explanatory factors are captured in the feature set used in this analysis. Some structural aspects like grid inertia, reserve allocation protocols, and real-time balancing market efficiency are only indirectly reflected in the data. Portugal's lower grid inertia, resulting from reduced conventional generation and a greater share of renewable energy sources, makes the system more responsive to net imbalance deviations. Its relatively limited cross-border interconnection and smaller reserve base further reduce the ability to absorb surprises. Germany, by contrast, benefits from large-scale reserve mechanisms, tighter integration with neighboring markets, and a more robust intra-day balancing infrastructure.

These omitted variables likely contribute to the residual forecast challenges that remain even after feature optimization. While the model captures key physical and behavioral drivers, a full understanding of grid-level forecast performance must also consider the systemic and operational context of the model. This highlights the necessity of country-specific interpretation: a model that performs adequately in Germany may systematically underestimate volatility or overpredict control capability in Portugal.

A key structural difference lies in the grid infrastructure. Germany operates a highly interconnected transmission network, offering flexibility in balancing regional imbalances. Portugal, in contrast, is located at the edge of the Iberian Peninsula and has limited interconnections, both internally and with neighboring systems. Figure 6.9 illustrates this difference in the grid layout, highlighting the denser German transmission grid compared to the simplet Portuguese one.

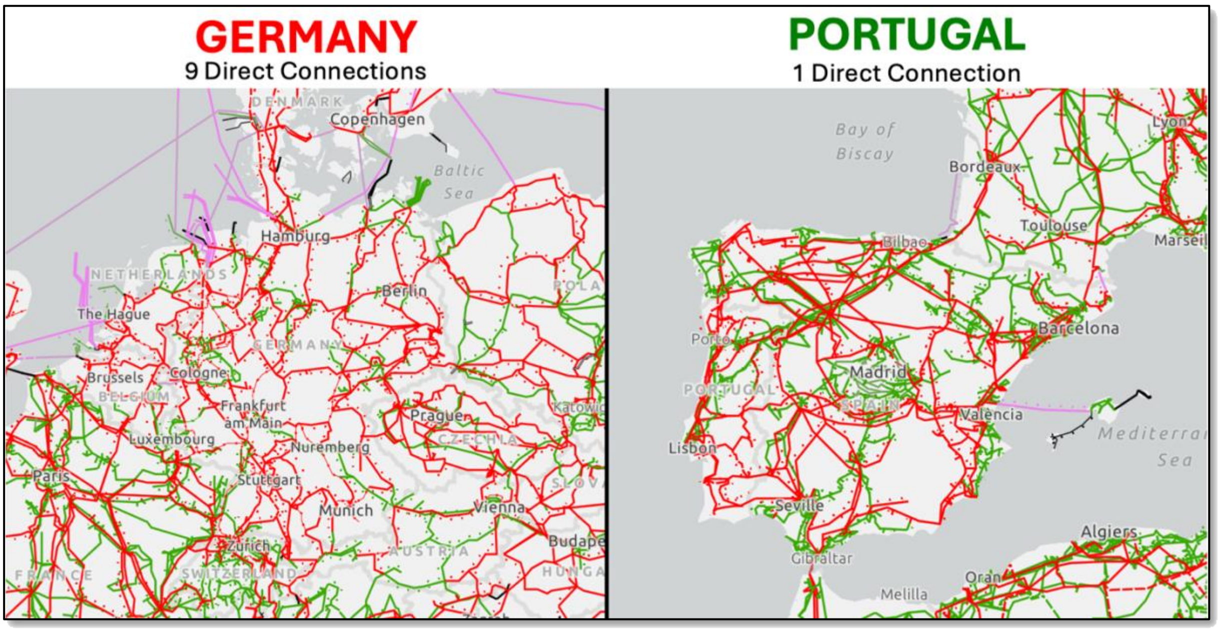


Figure 6.9: Transmission grid differences between Germany and Portugal (adapted from ENTSO-E Grid Map, 2025).

In summary, the cross-country analysis shows that net imbalance forecast performance is closely linked to the physical, economic, and institutional characteristics of each power system. Portugal's greater reliance on variable renewable generation, together with lower inertia and more limited balancing capacity, leads to a higher frequency of large forecast errors, both in absolute and relative terms. These challenges are not only technical but also have clear operational consequences, as discussed throughout this section. Interpreting model results therefore requires both data-driven evaluation and an understanding of system-specific conditions.

7. Discussion

7.1 Summary of Main Empirical Findings

This thesis investigated short-term electricity forecasting for consumption, production, and net imbalance in Germany and Portugal across multiple horizons using a unified modelling framework. Several consistent empirical findings are observed in this analysis.

First, electricity consumption is systematically easier to forecast than electricity production across all models, horizons, and countries. Production forecasts show higher errors due to their stronger dependence on volatile, weather-driven renewable generation. This difference is stronger in Portugal, where the share of renewable energy is higher and spatial averaging effects are weaker.

Second, forecast accuracy decreases as the forecasting horizon increases from 1 hour to 24 hours. While short-term forecasts benefit from recent system states and stable short-term patterns, longer horizons rely more on exogenous drivers such as weather forecasts and broader demand patterns, which introduce additional uncertainty.

Third, net imbalance forecast errors are not evenly distributed across severity levels. Although most errors fall into minor or significant categories, Portugal shows a significantly higher share of severe and critical deviations relative to system size. This finding remains despite the use of identical model architectures and feature sets, indicating that structural system characteristics play a central role.

Finally, SHAP-based interpretability reveals clear differences in the drivers of forecast errors. Renewable generation variables contribute disproportionately to large errors in Portugal, while Germany shows a more balanced influence between consumption-side and production-side features. These insights provide a physical explanation for the observed cross-country performance gap.

7.2 Horizon-Specific Challenges and Reliability

Forecast reliability declines systematically with increasing horizons. At the 1-hour horizon, recent observations, lagged values, and short-term weather patterns provide strong predictive signals, resulting in relatively low errors and a dominance of minor imbalances. At the 24-hour horizon, however, the informational value of recent system states decreases, and forecast performance becomes increasingly sensitive to weather uncertainty, behavioural shifts, and market dynamics.

This horizon effect has direct operational relevance. Errors at the day-ahead horizon shape generation schedules, reserve procurement, and cross-border capacity allocation. As the severity distribution shifts toward larger deviations at longer horizons, the risk of costly redispatch actions or reserve shortfalls increases even when average forecast accuracy remains acceptable.

Climate change may further intensify these challenges by increasing exposure to extreme weather and variability in renewable generation. Although this thesis does not establish a causal relationship, the results indicate that systems with a high reliance on renewables - such as Portugal - are more sensitive to forecast uncertainty at longer horizons. This highlights the importance of effective uncertainty management in day-ahead planning.

7.3 Cross-Country Generalizability and System Dependence

One of the central contributions of this thesis is the demonstration that identical forecasting models do not imply identical operational risk. Even when forecast accuracy is evaluated using scale-independent metrics such as MAPE, the operational implications of forecast errors differ significantly between Germany and Portugal.

Country-specific severity thresholds are therefore essential. A deviation that is manageable in Germany may already place significant strain on reserves or approach stability limits in

Portugal. Differences in grid inertia, reserve capacity, interconnection strength, and overall system size strongly influence how forecast errors lead to operational risk.

The SHAP analysis shows that Portugal's forecast errors are mainly driven by renewable variability, while Germany's more interconnected and diversified grid helps reduce these effects. The grid figures highlight the additional flexibility provided by Germany's dense transmission network

Taken together, these results underline that forecast performance must be interpreted within the physical and institutional context of the system in which the model operates. Generalizing forecasting results across countries without accounting for system characteristics risks systematically underestimating operational vulnerability in smaller or more renewable-intensive grids.

7.4 Interpretability versus Accuracy

Interpretability is an important complement to forecast accuracy in operational settings. SHAP-based analysis helps explain how the best-performing models behave and identifies which variables influence forecasts under different conditions.

In this thesis, SHAP analysis is used for three main purposes: to confirm that models rely on physically meaningful inputs, to help explain the sources of forecast errors, and to highlight system-specific risk factors such as renewable variability. SHAP does not imply causality and is not intended to guide control actions. Instead, it supports transparency, diagnostic analysis, and informed interpretation of model behaviour.

To support these objectives, direct autoregressive terms and highly aggregated variables were excluded from the SHAP analysis. Without this step, feature importance would be dominated by uninformative, aggregated variables, making it difficult to assess the role of external drivers or to compare results across countries. This methodological choice ensures that the SHAP results remain clear and meaningful.

7.5 Net Imbalance as a Residual: Methodological Implications

In this thesis, net imbalance forecasts are obtained by combining separate forecasts for electricity consumption and production, rather than by modeling imbalance directly. This approach reflects real-world system operation, where load and generation are forecast independently using different models, data sources, and update schedules, and imbalance arises as a result rather than as a primary planning target. As a consequence, uncertainty from both components carries over into the imbalance estimate, especially when consumption and production errors move in different directions.

From a purely statistical standpoint, modeling net imbalance directly could reduce forecast variability. However, doing so would make it harder to understand where imbalance risk originates and would limit interpretability. By keeping consumption and production forecasts separate, the framework allows high-severity imbalance events to be traced back to their underlying causes, which is essential for the severity-based evaluation and SHAP analysis used in this thesis. The higher imbalance severity observed at longer horizons and in smaller, renewable-heavy systems should therefore be understood as reflecting real operational risk, rather than as a side effect of the modeling approach.

7.6 Operational and Business Relevance

Forecast errors directly lead to operational actions and economic consequences. Minor and significant deviations are typically absorbed through routine reserve activation, but severe errors can require redispatch, emergency imports, or output reductions in renewable generation, leading to substantial system costs.

Market effects are particularly visible during periods of overproduction or misaligned forecasts. Recent events, such as episodes of zero or negative electricity prices driven by excess renewable generation, illustrate how forecast inaccuracies can distort market outcomes even in the absence of system instability. These effects are felt by generators, system operators, and ultimately consumers.

The severity-based framework developed in this thesis helps distinguish between forecast errors that matter mainly in a statistical sense and those that have real operational importance. It shows that reducing the occurrence of large errors can be more important than small improvements in average accuracy, particularly in systems with limited balancing capacity..

7.7 Limitations and Omitted Variables

This study has several limitations that should be acknowledged.

First, the forecasting models do not explicitly include a number of important system and operational factors, such as grid inertia, reserve activation strategies, real-time market mechanisms, network congestion, or unexpected outages. These factors affect how forecast errors influence system stability, but they are difficult to represent using publicly available, country-level data. As a result, their effects are only captured indirectly through historical system behaviour.

Second, the forecasting framework is based on deterministic point forecasts and symmetric error measures. This allows for clear comparison across countries and horizons, but it does not fully reflect the non-linear and asymmetric costs associated with large imbalance errors. In practice, forecast deviations become more costly as the system moves closer to stressed operating conditions, which cannot be fully represented within a point-forecast approach.

Third, computational constraints limited the extent of hyperparameter tuning and the exploration of more complex model architectures, especially deeper or longer-sequence deep learning models. While initial experiments did not suggest that such models would significantly change the main results, a more extensive search could lead to improvements. In addition, direct benchmarking of absolute forecasting performance against existing studies is challenging, as comparable benchmark values are scarce due to differences in data resolution, aggregation level, forecast horizon, feature availability, and evaluation protocols across the literature. In the context of this study, this limitation is of secondary importance, as the primary focus lies on relative performance differences between Germany and Portugal under an identical modelling framework rather than on absolute accuracy levels.

Finally, the analysis focuses on two countries and a limited historical period. Germany and Portugal were chosen to represent structurally different power systems, but the findings should not be interpreted as directly transferable to all electricity systems. Instead, the results show how the same forecasting methods can perform very differently depending on system size, renewable penetration, and balancing capability.

Overall, these limitations do not weaken the main conclusions of the thesis. Rather, they highlight the importance of interpreting forecasting results within the operational and structural context in which they are applied.

7.8 Synthesis

Overall, the results show that net imbalance forecasting is not just a statistical prediction task but also a system-specific risk assessment problem. Forecast performance, error severity, and operational impact depend both on model quality and on the underlying characteristics of the power system. Effective forecasting therefore requires not only advanced modelling methods but also evaluation approaches that reflect how forecasts are used in real system operation.

8. Conclusion and Future Work

8.1 Answering the Research Question

This thesis set out to investigate whether modern machine learning-based forecasting models can reliably predict electricity system imbalances across different countries and operational horizons, and how forecast errors should be interpreted from an operational risk perspective. The results demonstrate that identical forecasting approaches do not translate into identical operational risk across power systems.

While forecast accuracy for consumption and production is generally high at short horizons in both countries, Portugal exhibits a markedly higher share of severe and critical net imbalance errors under the same modeling setup. This finding confirms that aggregate accuracy metrics alone are insufficient to assess forecast quality in operational contexts. Severity-based evaluation reveals risks that remain hidden when errors are averaged over time or treated symmetrically.

The SHAP-based analysis further shows that renewable generation variables - particularly wind and solar - play a substantially larger role in driving forecast errors in Portugal than in Germany, across both short-term and day-ahead horizons. This highlights renewable-driven volatility as a key source of imbalance risk in smaller, lower-inertia systems. Together, these findings directly answer the research question: imbalance forecasting performance must be interpreted in a system-specific and severity-aware manner, as the same numerical error can have fundamentally different operational implications depending on grid structure and generation mix.

International institutions have explicitly identified Portugal as increasingly exposed to electricity security challenges due to high renewable shares, reduced conventional generation, and growing net-load variability (IEA, 2022). Recent national responses further underscore this exposure, with Portuguese authorities signaling the need to retain dispatchable capacity with autonomous start capability as a stability backstop during extreme conditions (Observador, 2024). Similar developments in Germany, where transmission system operators emphasize hydrogen-ready power plants as future stabilizing resources, illustrate that imbalance risk is not eliminated by scale alone but mitigated through infrastructure, flexibility, and system design (AMPRION, 2024). These real-world developments reinforce the relevance of the severity framework and interpretability results presented in this thesis.

8.2 Contributions to Theory and Practice

Methodologically, this thesis introduces a severity-aware framework for evaluating net imbalance forecasts that accounts for differences in system size, inertia, and reserve capacity. By defining country-specific severity thresholds based on operational practice, the framework shows how forecast errors can be interpreted in terms of operational risk rather than as abstract numerical values.

From a practical perspective, the results highlight the value of interpretable machine learning for system operators and planners. The combination of accurate forecasting with SHAP-based diagnostics enables early identification of risk drivers, supports trust in model outputs, and provides insights into why and when forecasts are likely to fail. Rather than replacing operational judgment, such tools can complement existing decision-making processes by highlighting periods of elevated imbalance risk.

8.3 Future Research Directions

This work points to several directions for future research. These directions reflect a broader view of forecasting, where the goal is not only to produce accurate point predictions, but also to manage operational risk. From this perspective, the focus shifts from minimizing average forecast error to identifying and reducing large errors that can threaten system stability.

On the modelling side, probabilistic forecasting methods could provide uncertainty estimates that better support reserve planning and operational decisions, especially in systems with high shares of renewable generation. By explicitly representing forecast uncertainty, these approaches would allow operators to assess the likelihood of severe imbalance events rather than relying only on single-point forecasts. In addition, online learning or state-aware models could improve robustness under changing conditions, such as extreme weather events or periods of low system inertia.

An important methodological extension would be the explicit disaggregation of production forecasts by generation type. The SHAP analysis indicates that different renewable sources contribute unevenly to forecast uncertainty, suggesting that a single, unified production model may covers source-specific dynamics. Future work could therefore explore tailored forecasting pipelines for wind, solar, hydro, and conventional generation, with model architectures, feature sets, and loss functions optimized for the physical characteristics of each production type. Such an approach may improve both forecast accuracy and interpretability, particularly in systems where renewable variability dominates imbalance risk.

From an operational perspective, future studies could investigate loss functions that penalize large forecast errors more strongly than small deviations, reflecting the non-linear costs associated with severe imbalances. While this may reduce average error performance, it could improve system reliability by explicitly prioritizing the avoidance of high-impact events. This shift further reinforces the interpretation of forecasting models as tools for managing operational risk rather than purely statistical error minimization.

Finally, infrastructure-related responses-such as increased cross-border interconnection, energy storage deployment, hydrogen-based flexibility, or subsea transmission links-represent important mechanisms for mitigating imbalance risk. Evaluating these measures quantitatively lies beyond the scope of this thesis; however, the severity-aware framework and interpretability results developed here provide a foundation for future assessments that integrate forecasting performance with system planning, resilience analysis, and cost-benefit evaluation.

Bibliography

50Hertz, AMPRION, Tennet, & Transnet BW. (2019). *Investigation on system imbalances in Germany in June 2019*.

ACER. (2025). *Electricity Market Integration Report 2025*. https://acer.europa.eu/

Al-Ali, E. M., Hajji, Y., Said, Y., Hleili, M., Alanzi, A. M., Laatar, A. H., & Atri, M. (2023). Solar Energy Production Forecasting Based on a Hybrid CNN-LSTM-Transformer Model. *Mathematics*, *11*(3), 676. https://doi.org/10.3390/math11030676

AMPRION. (2024, March 24). *TRANSMISSION SYSTEM OPERATORS PUBLISH FIRST DRAFT OF GRID DEVELOPMENT PLAN FOR 2037/2045*. https://www.amprion.net/Press/Press-Detail-Page_50944.html

Balázs, I., Fodor, A., & Magyar, A. (2024). Short-term system imbalance forecast using linear and non-linear methods. *Energy Systems*. https://doi.org/10.1007/s12667-024-00667-7

Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects. *Sustainability*, *15*(9), 7087. https://doi.org/10.3390/su15097087

Bloomberg. (2025, August 12). *French Power Prices Tumble to Zero as Warm Weather Dents Demand*. https://www.bloomberg.com/news/articles/2025-12-08/french-power-prices-tumble-to-zero-as-warm-weather-dents-demand

Chavez, S. G., Bernat, J. X., & Coalla, H. L. (1999). *Forecasting of energy production and consumption in Asturias (northern Spain)*.

Ejuh Che, E., Roland Abeng, K., Iweh, C. D., Tsekouras, G. J., & Fopah-Lele, A. (2025). The Impact of Integrating Variable Renewable Energy Sources into Grid-Connected Power Systems: Challenges, Mitigation Strategies, and Prospects. *Energies*, *18*(3), 689. https://doi.org/10.3390/en18030689

ENTSO-E. (2022). *ICS Investigation Expert Panel*.

ENTSO-E. (2023). *Annual Report*. https://www.entsoe.eu/

ENTSO-E Transparency Platform. (2025). *ENTSO-E Transparency Platform* [Dataset]. https://transparency.entsoe.eu

ENTSOE-E. (2019). *Incident Classification Scale Methodology*.

ENTSOE-E. (2022). *SYSTEM DEFENCE PLAN*.

ENTSOE-E Grid Map. (2025). *ENTSOE-E Grid Map* [Dataset]. https://www.entsoe.eu/data/map/

Gasparin, A., Lukovic, S., & Alippi, C. (2019). *Deep Learning for Time Series Forecasting: The Electric Load Case* (No. arXiv:1907.09207). arXiv. https://doi.org/10.48550/arXiv.1907.09207

Graabak, I., & Korpås, M. (2016). Variability Characteristics of European Wind and Solar Power Resources—A Review. *Energies*, *9*(6), 449. https://doi.org/10.3390/en9060449

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., & Zareipour, H. (2020). Energy Forecasting: A Review and Outlook. *IEEE Open Access Journal of Power and Energy*, *7*, 376–388. https://doi.org/10.1109/OAJPE.2020.3029979

IEA. (2022). *Portugal Electricity Security Policy*. https://www.iea.org/articles/portugal-electricity-security-policy#

Ledmaoui, Y., El Maghraoui, A., El Aroussi, M., Saadane, R., Chebak, A., & Chehri, A. (2023). Forecasting solar energy production: A comparative study of machine learning algorithms. *Energy Reports*, *10*, 1004–1012. https://doi.org/10.1016/j.egyr.2023.07.042

Li, M., & Wang, Y. (2022). Power load forecasting and interpretable models based on GS_XGBoost and SHAP. *Journal of Physics: Conference Series*, *2195*(1), 012028. https://doi.org/10.1088/1742-6596/2195/1/012028

Mystakidis, A., Koukaras, P., Tsalikidis, N., Ioannidis, D., & Tjortjis, C. (2024). Energy Forecasting: A Comprehensive Review of Techniques and Technologies. *Energies*, *17*(7), 1662. https://doi.org/10.3390/en17071662

Observador. (2024, April 29). *Apagão. Governo quer centrais do Baixo Sabor e Alqueva com arranque autónomo*. https://observador.pt/2025/04/29/apagao-governo-quer-centrais-do-baixo-sabor-e-alqueva-com-arranque-autonomo/

OMIE. (2025). *REN - Redes Energéticas Nacionais—Prices* [Dataset]. https://www.datahub.ren.pt

Open-Meteo. (2025). *Open-Meteo Weather API* [Dataset]. https://open-meteo.com

Pavlík, M., Kurimský, F., & Ševc, K. (2025). Renewable Energy and Price Stability: An Analysis of Volatility and Market Shifts in the European Electricity Sector (2015–2025). *Applied Sciences*, *15*(12), 6397. https://doi.org/10.3390/app15126397

Plakas, K., Andriopoulos, N., Birbas, A., Moraitis, I., & Papalexopoulos, A. (2023a). A Forecasting Model for the Prediction of System Imbalance in the Greek Power System. *ITISE 2023*, 18. https://doi.org/10.3390/engproc2023039018

Plakas, K., Andriopoulos, N., Birbas, A., Moraitis, I., & Papalexopoulos, A. (2023b). A Forecasting Model for the Prediction of System Imbalance in the Greek Power System. *ITISE 2023*, 18. https://doi.org/10.3390/engproc2023039018

REN DataHub. (2025). *REN - Redes Energéticas Nacionais* [Dataset]. https://www.datahub.ren.pt

REUTERS. (2025, April 29). *Spain, Portugal switch back on, seek answers after biggest ever blackout*. https://www.reuters.com/world/europe/spains-power-generation-nearly-back-normal-after-monday-blackout-says-grid-2025-04-29/

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., & García-Herrera, R. (2018). Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*, *90*, 728–741. https://doi.org/10.1016/j.rser.2018.04.008

Shahiduzzaman, K. M., Jamal, M. N., & Nawab, Md. R. I. (2021). Renewable Energy Production Forecasting: A Comparative Machine Learning Analysis. *International Journal of Engineering and Advanced Technology*, *10*(6), 11–18. https://doi.org/10.35940/ijeat.E2689.0810621

SMARD Platform. (2025). *Bundesnetzagentur* [Dataset]. https://www.smard.de

Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. *North American Power Symposium 2010*, 1–8. https://doi.org/10.1109/NAPS.2010.5619586

Song, K.-M., Kim, T.-G., Cho, S.-M., Song, K.-B., & Yoon, S.-G. (2025). XGBoost-Based Very Short-Term Load Forecasting Using Day-Ahead Load Forecasting Results. *Electronics*, *14*(18), 3747. https://doi.org/10.3390/electronics14183747

Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews*, *16*(2), 1223–1240. https://doi.org/10.1016/j.rser.2011.08.014

Sweeney, C., Bessa, R. J., Browell, J., & Pinson, P. (2020). The future of forecasting for renewable energy. *WIREs Energy and Environment*, *9*(2), e365. https://doi.org/10.1002/wene.365

Ullah, K., Ahsan, M., Hasanat, S. M., Haris, M., Yousaf, H., Raza, S. F., Tandon, R., Abid, S., & Ullah, Z. (2024). Short-Term Load Forecasting: A Comprehensive Review and Simulation Study With CNN-LSTM Hybrids Approach. *IEEE Access*, *12*, 111858–111881. https://doi.org/10.1109/ACCESS.2024.3440631

Zhao, H.-X., & Magoulès, F. (2012). Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method. *Journal of Algorithms & Computational Technology*, *6*(1), 59–77. https://doi.org/10.1260/1748-3018.6.1.59

Zhao, Z., Xia, C., Chi, L., Chang, X., Li, W., Yang, T., & Zomaya, A. Y. (2021). Short-Term Load Forecasting Based on the Transformer Model. *Information*, *12*(12), 516. https://doi.org/10.3390/info12120516

Appendix A - Weather Data Aggregation and Weighting

To construct a single representative weather time series for each country, hourly meteorological observations were retrieved from the Open-Meteo Archive API for a fixed set of cities. For each weather variable, national-level values were computed as weighted averages across these locations. The same cities are used consistently for all variables within a country, while the applied weights differ depending on the variable.

Three weighting schemes are employed. General meteorological variables (temperature, relative humidity, precipitation, snowfall, and cloud cover) use equal weights across all cities. For shortwave radiation, weights are adjusted to better reflect regions with higher relevance for solar generation. For wind speed at 100 m, weights are adjusted to reflect regions with higher relevance for wind generation. All weights are fixed over the full study period.

A.1 Germany - Locations and Weights

For Germany, five cities were selected to represent major climatic zones and regions relevant for electricity generation: Munich, Berlin, Hamburg, Cologne, and Frankfurt. Table A.1 summarises the weights applied under each weighting scheme.

Table A.1: Germany - City Weights by Variable Group

Location	General Variables	Shortwave Radiation	Wind Speed (100 m)
Munich	0.20	0.35	0.10
Berlin	0.20	0.15	0.20
Hamburg	0.20	0.10	0.40
Cologne	0.20	0.15	0.20
Frankfurt	0.20	0.25	0.10

General variables include temperature, relative humidity, cloud cover, rain, and snowfall. Shortwave radiation weights place greater emphasis on southern and south-western regions relevant for solar generation, while wind speed weights emphasise northern and coastal regions relevant for wind generation.

A.2 Portugal - Locations and Weights

For Portugal, five cities were selected to capture key climatic and generation-relevant regions: Lisbon, Porto, Faro, Coimbra, and Évora. Table A.2 summarises the weights applied under each weighting scheme.

Table A.2: Portugal - City Weights by Variable Group

Location	General Variables	Shortwave Radiation	Wind Speed (100 m)
Lisbon	0.20	0.20	0.25
Porto	0.20	0.10	0.30
Faro	0.20	0.30	0.15
Coimbra	0.20	0.15	0.20
Évora	0.20	0.25	0.10

General variables include temperature, relative humidity, cloud cover, rain, and snowfall. Shortwave radiation weights give greater emphasis to southern and inland locations relevant for solar generation, while wind speed weights emphasise coastal and northern/central locations relevant for wind generation.

A.3 Rationale and Limitations

The adopted weighting approach represents a practical compromise between physical relevance and data availability. It allows the construction of country-level weather signals that are informative for short-term electricity forecasting while remaining transparent and reproducible.

The weights are not statistically optimised and do not capture fine-scale spatial heterogeneity or highly localised extreme weather events. More detailed spatial aggregation strategies, such as gridded reanalysis data or capacity-weighted schemes, could improve accuracy but were beyond the scope of this thesis.

Appendix B - Additional Model Results

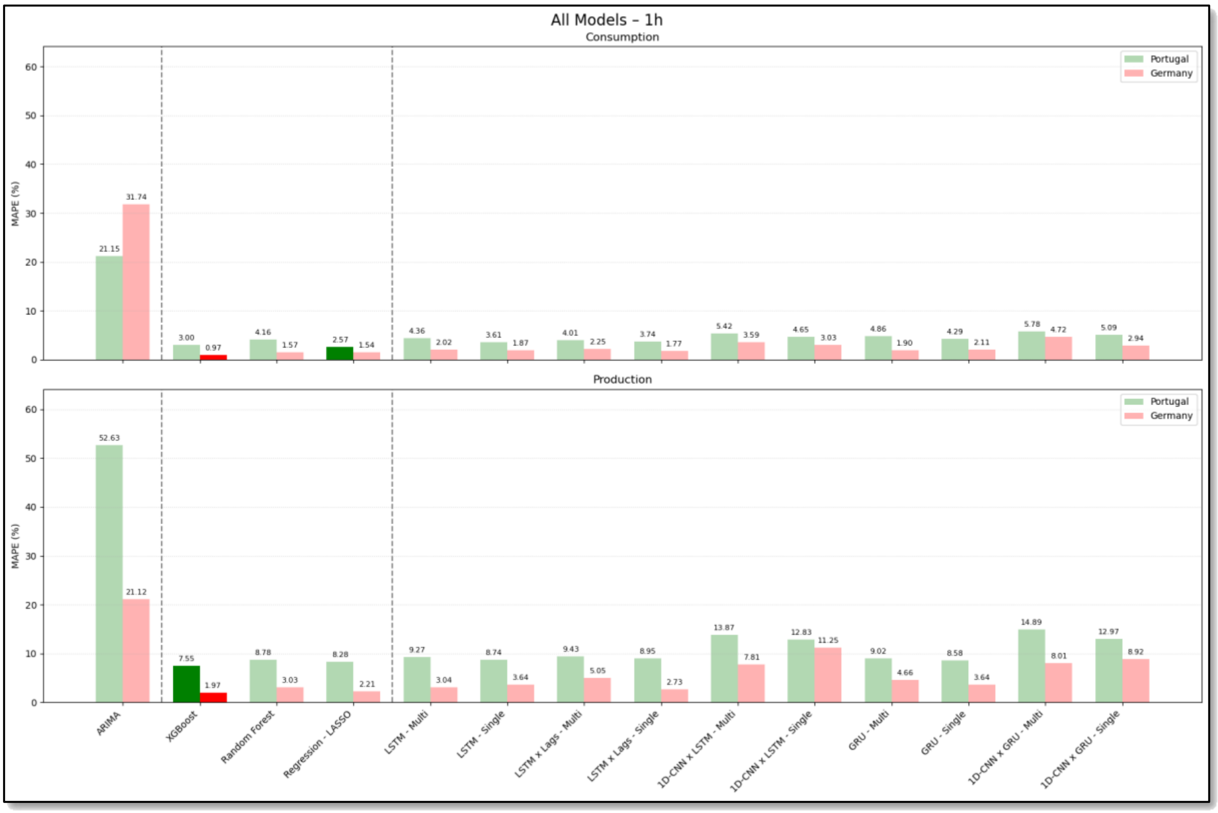
Appendix B.1 Table MAPE Results of all Models

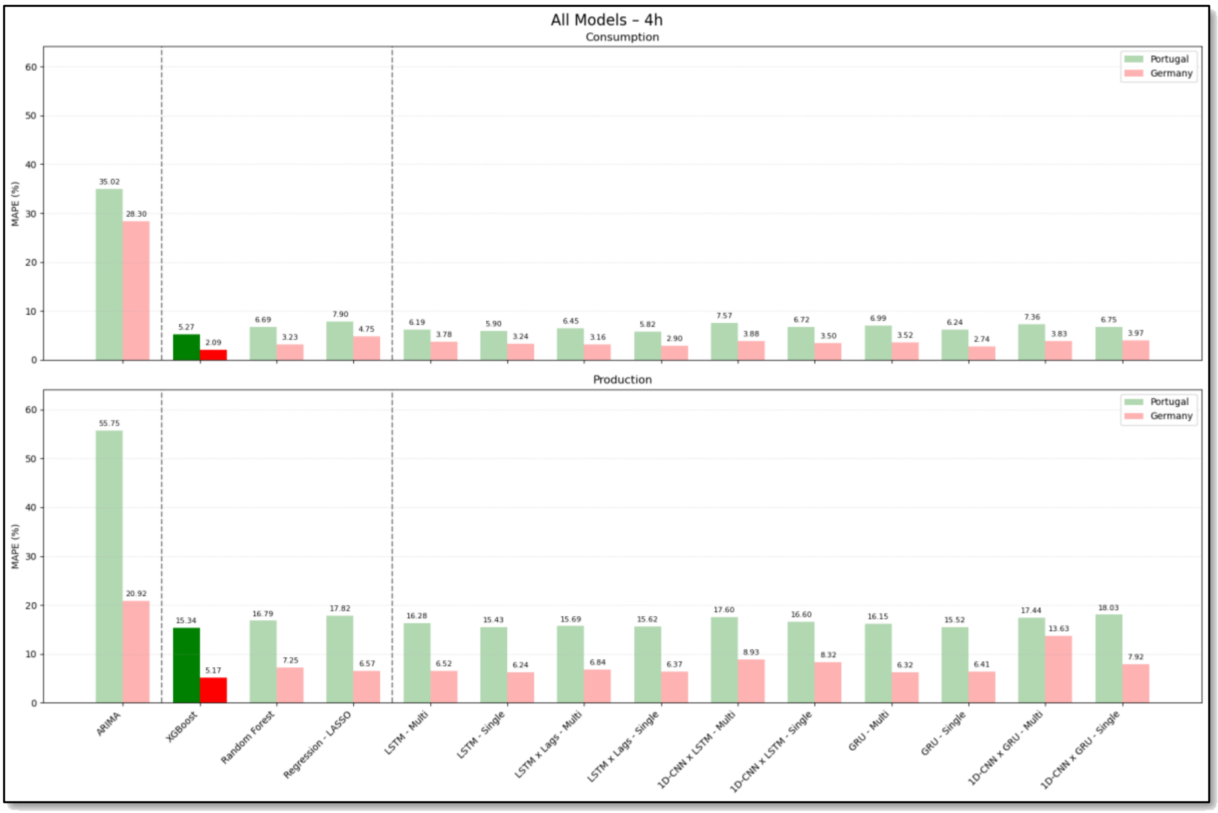
Horizon	Countries	Models	MAPE Cons. In %	MAPE Prod. in %
1h	Portugal	ARIMA	21,15	52,63
1h	Portugal	XGBoost	3,00	7,55
1h	Portugal	Random Forest	4,16	8,78
1h	Portugal	Regression - LASSO	2,57	8,28
1h	Portugal	LSTM - Multi	4,36	9,27
1h	Portugal	LSTM - Single	3,61	8,74
1h	Portugal	LSTM x Lags - Multi	4,01	9,43
1h	Portugal	LSTM x Lags - Single	3,74	8,95
1h	Portugal	1D - CNN x LSTM - Multi	5,42	13,87
1h	Portugal	1D - CNN x LSTM - Single	4,65	12,83
1h	Portugal	GRU - Multi	4,86	9,02
1h	Portugal	GRU - Single	4,29	8,58
1h	Portugal	1D-CNN x GRU - Multi	5,78	14,89
1h	Portugal	1D-CNN x GRU - Single	5,09	12,97
1h	Germany	ARIMA	31,74	21,12
1h	Germany	XGBoost	0,97	1,97
1h	Germany	Random Forest	1,57	3,03
1h	Germany	Regression - LASSO	1,54	2,21
1h	Germany	LSTM - Multi	2,02	3,04
1h	Germany	LSTM - Single	1,87	3,64
1h	Germany	LSTM x Lags - Multi	2,25	5,05
1h	Germany	LSTM x Lags - Single	1,77	2,73
1h	Germany	1D - CNN x LSTM - Multi	3,59	7,81
1h	Germany	1D - CNN x LSTM - Single	3,03	11,25
1h	Germany	GRU - Multi	1,90	4,66
1h	Germany	GRU - Single	2,11	3,64

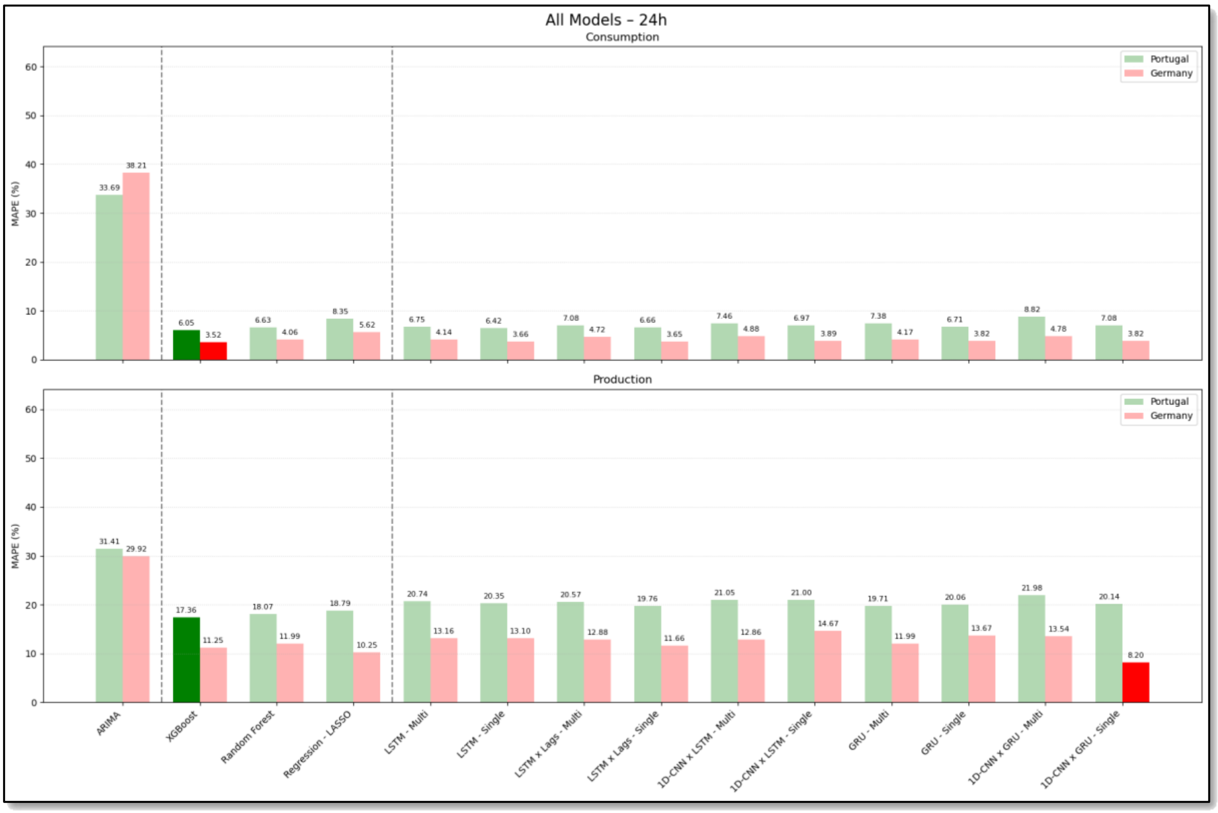
1h	Germany	1D-CNN x GRU - Multi	4,72	8,01
1h	Germany	1D-CNN x GRU - Single	2,94	8,92
4h	Portugal	ARIMA	35,02	55,75
4h	Portugal	XGBoost	5,27	15,34
4h	Portugal	Random Forest	6,69	16,79
4h	Portugal	Regression - LASSO	7,90	17,82
4h	Portugal	LSTM - Multi	6,19	16,28
4h	Portugal	LSTM - Single	5,90	15,43
4h	Portugal	LSTM x Lags - Multi	6,45	15,69
4h	Portugal	LSTM x Lags - Single	5,82	15,62
4h	Portugal	1D - CNN x LSTM - Multi	7,57	17,60
4h	Portugal	1D - CNN x LSTM - Single	6,72	16,60
4h	Portugal	GRU - Multi	6,99	16,15
4h	Portugal	GRU - Single	6,24	15,52
4h	Portugal	1D-CNN x GRU - Multi	7,36	17,44
4h	Portugal	1D-CNN x GRU - Single	6,75	18,03
4h	Germany	ARIMA	28,30	20,92
4h	Germany	XGBoost	2,09	5,17
4h	Germany	Random Forest	3,23	7,25
4h	Germany	Regression - LASSO	4,75	6,57
4h	Germany	LSTM - Multi	3,78	6,52
4h	Germany	LSTM - Single	3,24	6,24
4h	Germany	LSTM x Lags - Multi	3,16	6,84
4h	Germany	LSTM x Lags - Single	2,90	6,37
4h	Germany	1D - CNN x LSTM - Multi	3,88	8,93
4h	Germany	1D - CNN x LSTM - Single	3,50	8,32
4h	Germany	GRU - Multi	3,52	6,32
4h	Germany	GRU - Single	2,74	6,41
4h	Germany	1D-CNN x GRU - Multi	3,83	13,63
4h	Germany	1D-CNN x GRU - Single	3,97	7,92
24h	Portugal	ARIMA	33,69	31,41

24h	Portugal	XGBoost	6,05	17,36
24h	Portugal	Random Forest	6,63	18,07
24h	Portugal	Regression - LASSO	8,35	18,79
24h	Portugal	LSTM - Multi	6,75	20,74
24h	Portugal	LSTM - Single	6,42	20,35
24h	Portugal	LSTM x Lags - Multi	7,08	20,57
24h	Portugal	LSTM x Lags - Single	6,66	19,76
24h	Portugal	1D - CNN x LSTM - Multi	7,46	21,05
24h	Portugal	1D - CNN x LSTM - Single	6,97	21,00
24h	Portugal	GRU - Multi	7,38	19,71
24h	Portugal	GRU - Single	6,71	20,06
24h	Portugal	1D-CNN x GRU - Multi	8,82	21,98
24h	Portugal	1D-CNN x GRU - Single	7,08	20,14
24h	Germany	ARIMA	38,21	29,92
24h	Germany	XGBoost	3,52	11,25
24h	Germany	Random Forest	4,06	11,99
24h	Germany	Regression - LASSO	5,62	10,25
24h	Germany	LSTM - Multi	4,14	13,16
24h	Germany	LSTM - Single	3,66	13,10
24h	Germany	LSTM x Lags - Multi	4,72	12,88
24h	Germany	LSTM x Lags - Single	3,65	11,66
24h	Germany	1D - CNN x LSTM - Multi	4,88	12,86
24h	Germany	1D - CNN x LSTM - Single	3,89	14,67
24h	Germany	GRU - Multi	4,17	11,99
24h	Germany	GRU - Single	3,82	13,67
24h	Germany	1D-CNN x GRU - Multi	4,78	13,54
24h	Germany	1D-CNN x GRU - Single	3,82	8,20

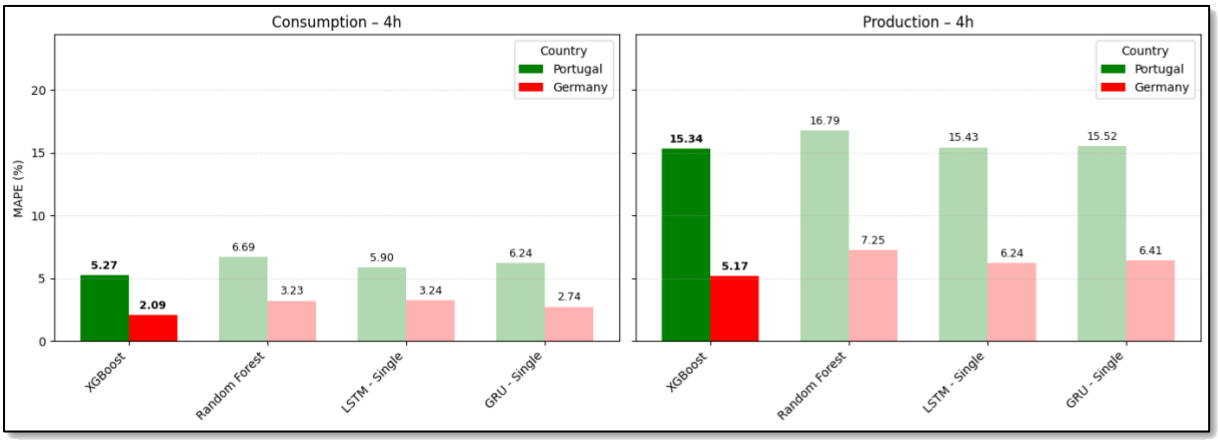
Appendix B.2 Plots of MAPE Results of all Models


All Models – 1h
Consumption
Production
MAPE (%)
Portugal
Germany
ARIMA
XGBoost
Random Forest
Regression - LASSO
LSTM - Multi
LSTM - Single
LSTM x Lags - Multi
LSTM x Lags - Single
1D-CNN x LSTM - Multi
1D-CNN x LSTM - Single
GRU - Multi
GRU - Single
1D-CNN x GRU - Multi
1D-CNN x GRU - Single
21.15
31.74
3.00
0.97
4.16
1.57
2.57
1.54
4.36
2.02
3.61
1.87
4.01
2.25
3.74
1.77
5.42
3.59
4.65
3.03
4.86
1.90
4.29
2.11
5.78
4.72
5.09
2.94
52.63
21.12
7.55
1.97
8.78
3.03
8.28
2.21
9.27
3.04
8.74
3.64
9.43
5.05
8.95
2.73
13.87
7.81
12.83
11.25
9.02
4.66
8.58
3.64
14.89
8.01
12.97
8.92


All Models – 4h
Consumption
Production
MAPE (%)
Portugal
Germany
ARIMA
XGBoost
Random Forest
Regression - LASSO
LSTM - Multi
LSTM - Single
LSTM x Lags - Multi
LSTM x Lags - Single
1D-CNN x LSTM - Multi
1D-CNN x LSTM - Single
GRU - Multi
GRU - Single
1D-CNN x GRU - Multi
1D-CNN x GRU - Single
35.02
28.30
5.27
2.09
6.69
3.23
7.90
4.75
6.19
3.78
5.90
3.24
6.45
3.16
5.82
2.90
7.57
3.88
6.72
3.50
6.99
3.52
6.24
2.74
7.36
3.83
6.75
3.97
55.75
20.92
15.34
5.17
16.79
7.25
17.82
6.57
16.28
6.52
15.43
6.24
15.69
6.84
15.62
6.37
17.60
8.93
16.60
8.32
16.15
6.32
15.52
6.41
17.44
13.63
18.03
7.92



Appendix B.3 Plots of Best Models 4h Horizon



Appendix C - Sensitivity Analysis of Severity Thresholds

To test how sensitive the results are to the choice of severity thresholds, a robustness analysis was carried out using lower-bound and upper-bound variants of the relative severity definitions. In the lower-bound case, stricter percentage-of-load thresholds are applied, meaning that forecast errors are more easily classified as severe or critical. In the upper-bound case, the thresholds are relaxed, shifting borderline cases into lower severity categories. Importantly, the forecasts and forecast errors themselves remain unchanged; only their classification into severity levels is affected.

Apendix C.1 - Lower-Bound Thresholds

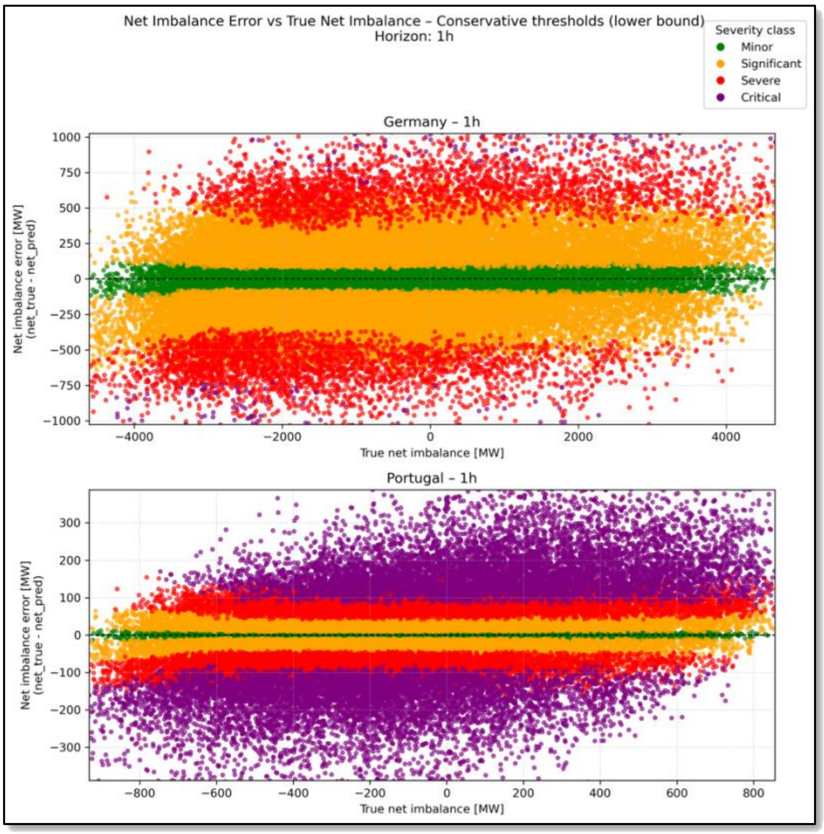
Under the lower-bound thresholds, forecast errors are more frequently classified as severe or critical across both countries and forecasting horizons. This is visible in the scatter plots as an expansion of the high-severity regions, particularly at larger imbalance values. In this scenario, the baseline severity thresholds of **1%**, **5%**, and **10%** of system load are tightened to **0.8%**, **4%**, and **8%**, respectively. As a result, smaller absolute errors are sufficient to trigger higher severity classifications. However, the overall shape and structure of the error distributions remain unchanged.

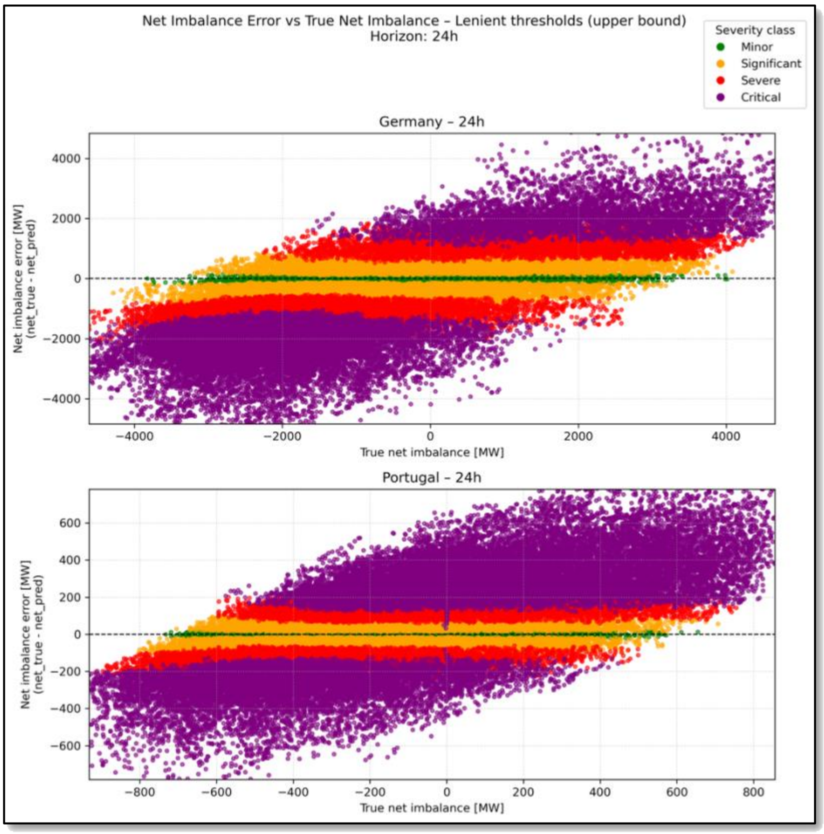
For Germany at the 1-hour horizon, the share of severe and critical errors increases compared to the baseline, reaching around 10% of all time steps. Even under these stricter thresholds, critical errors remain very rare, accounting for well below 1%. The error distribution stays tightly centered around zero, indicating that short-term net-imbalance forecasting in Germany remains stable even under conservative severity definitions.

At the 24-hour horizon, Germany shows a higher share of severe and critical classifications, with critical errors exceeding 50% under the lower-bound thresholds. This increase reflects the expected rise in uncertainty at longer forecast horizons rather than a fundamental change in model behavior. The scatter plot remains symmetric and shows no signs of instability or systematic bias.

For Portugal, the lower-bound thresholds further highlight an already challenging forecasting environment. At the 1-hour horizon, more than half of all forecasts are classified as severe or critical, with critical errors making up nearly one quarter of all time steps. At the 24-hour horizon, critical errors become dominant, exceeding 60%. These shifts do not introduce new patterns but instead strengthen those already observed under the baseline thresholds. Error dispersion remains much wider than in Germany, and forecast errors continue to increase with imbalance magnitude.

Overall, the lower-bound sensitivity analysis shows that the main conclusions of the thesis are robust to stricter severity definitions. Germany consistently shows stronger short-term predictability, while Portugal shows a higher exposure to high-severity forecast errors. The relative comparison across countries and horizons remains unchanged.





Appendix C.2 - Upper-Bound Thresholds

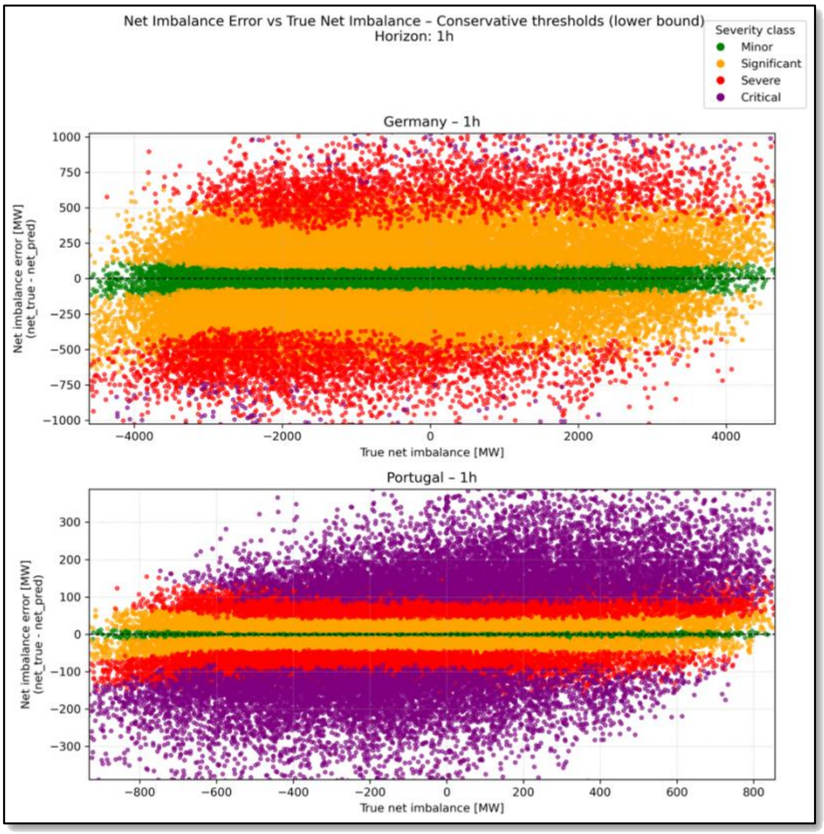
Under the upper-bound thresholds, severity classifications shift in the opposite direction compared to the lower-bound case. Marginal forecast errors are more often assigned to lower severity categories, which appears in the scatter plots as a reduction in severe and critical regions and an expansion of minor and significant classifications. In this scenario, the baseline thresholds of **1%**, **5%**, and **10%** of system load are relaxed to **1.2%**, **6%**, and **12%**, respectively, meaning that larger absolute errors are required to be classified as severe or critical. The underlying forecasts and error distributions remain unchanged.

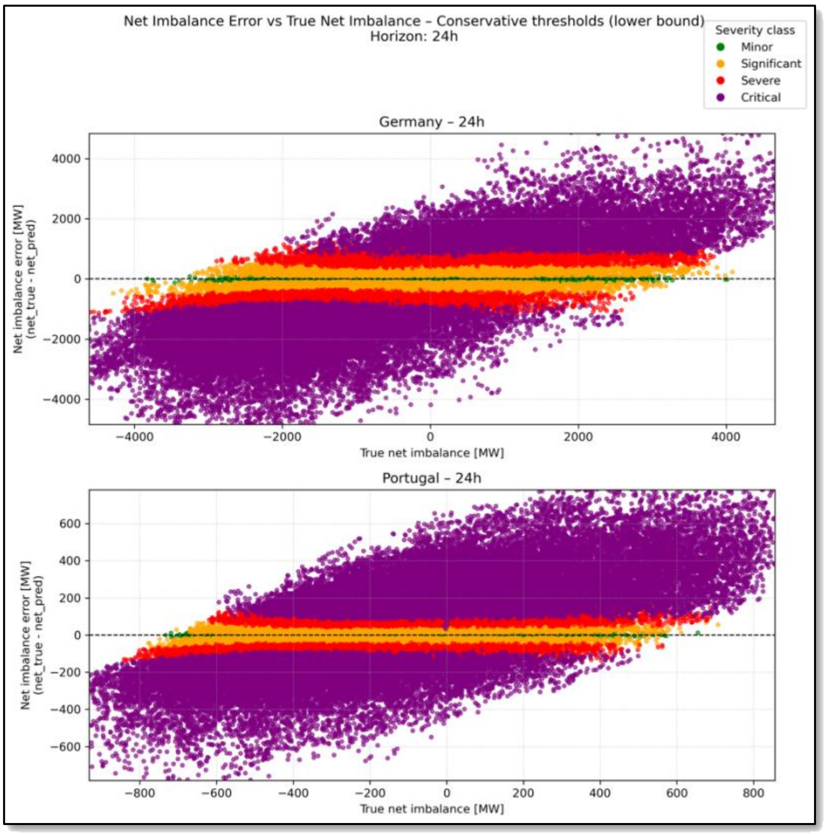
For Germany at the 1-hour horizon, the use of upper-bound thresholds reduces the share of severe and critical errors to only a small fraction of all time steps, with critical errors becoming almost none. This supports the conclusion that short-term net-imbalance forecasting in Germany is operationally reliable and that high-severity deviations occur only in rare situations. The shape and spread of the error distribution remain unchanged, confirming that the reduction in severity reflects classification choices rather than improved forecast accuracy.

At the 24-hour horizon, Germany continues to show elevated uncertainty, although the proportion of critical errors is lower than in the baseline case. Even under these lower thresholds, a substantial share of forecasts remains classified as severe or critical, indicating that day-ahead imbalance forecasting remains challenging regardless of how severity is defined.

For Portugal, the upper-bound thresholds reduce the share of high-severity outcomes but do not eliminate them. At the 1-hour horizon, a notable portion of forecasts is still classified as severe or critical. At the 24-hour horizon, nearly half of all forecasts continue to fall into the critical category. The scatter plots still show wide error dispersion and a strong relationship between imbalance magnitude and forecast error, particularly when compared to Germany.

Overall, the upper-bound sensitivity analysis confirms that the main findings of the thesis are robust. Even under more lenient severity definitions, Portugal remains significantly more difficult to forecast than Germany, especially at longer horizons.


Net Imbalance Error vs True Net Imbalance – Conservative thresholds (lower bound)
Horizon: 1h
Severity class
Minor
Significant
Severe
Critical
Germany – 1h
Net imbalance error [MW] (net_true - net_pred)
True net imbalance [MW]
Portugal – 1h
Net imbalance error [MW] (net_true - net_pred)
True net imbalance [MW]


Net Imbalance Error vs True Net Imbalance – Conservative thresholds (lower bound)
Horizon: 24h
Severity class
Minor
Significant
Severe
Critical
Germany – 24h
Net imbalance error [MW] (net_true - net_pred)
True net imbalance [MW]
Portugal – 24h
Net imbalance error [MW] (net_true - net_pred)
True net imbalance [MW]

Appendix D: SHAP Analysis Conditioned on Severe and Critical Forecast Errors

This appendix extends the main SHAP-based analysis by focusing only on forecast cases classified as severe or critical net-imbalance errors. These classifications are based on the relative severity thresholds defined as a percentage of contemporaneous system load. The aim is to examine whether extreme forecast errors arise from different underlying drivers than typical errors, or whether they reflect stronger versions of the same mechanisms observed in the full sample.

Across both countries, forecast targets, and horizons, the results show that severe and critical errors are not random outliers. Instead, they are linked to clear and interpretable feature patterns, which remain consistent across model configurations.

For electricity consumption, SHAP values continue to be dominated by temporal features such as hour of day, day of week, and seasonal indicators, even when considering only extreme errors. Weather variables like temperature and shortwave radiation play a secondary role, while market- and generation-related features contribute little. The relatively narrow SHAP distributions for consumption suggest that large consumption forecast errors mainly occur when regular demand patterns are disrupted, for example during holidays, unusual demand periods, or behavioural shifts. This confirms that consumption forecasting remains comparatively stable and is not the main source of high-severity net-imbalance events.

In contrast, production-side SHAP analyses for severe and critical cases show much wider and more uneven feature contributions. Renewable-related variables-especially wind, solar, and hydropower-display large SHAP magnitudes and broad distributions. This indicates that extreme imbalance events are closely linked to uncertainty in renewable generation forecasts. Dispatchable generation and cross-border flows also contribute, but their influence is generally smaller than that of weather-driven renewables.

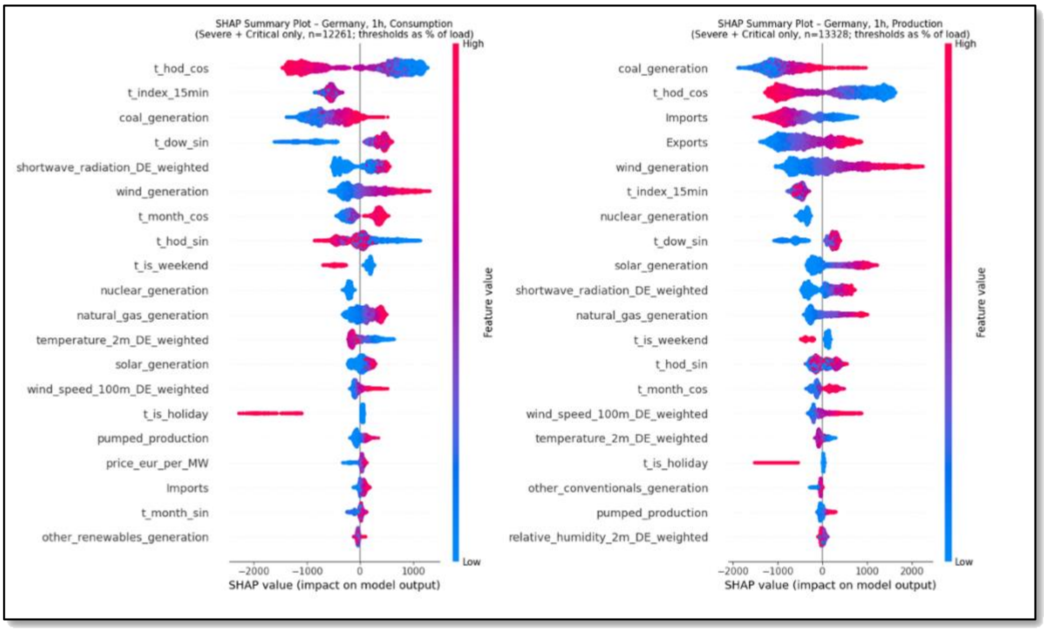
The differences between Germany and Portugal become clearer when focusing on high-severity events. In Germany, severe and critical production errors are associated with a mix of contributing factors, including renewable output, imports and exports, and conventional generation. No single variable consistently dominates the SHAP values, reflecting the stabilising effects of Germany’s large system size, geographic spread, and diversified generation mix. Even in extreme cases, forecast errors typically result from the interaction of several factors rather than from one dominant driver.

In Portugal, by contrast, severe and critical forecast errors are strongly concentrated around renewable-related features. Wind generation, hydropower output, and weather variables such as radiation and temperature dominate the SHAP explanations. Imports and exports become more important in these cases, reflecting the system's greater reliance on cross-border balancing under stress. Temporal features play a smaller role than in Germany. This concentration indicates that high-severity imbalance events in Portugal are mainly driven by renewable variability, combined with the system's smaller size, higher renewable share, and more limited internal balancing capacity.

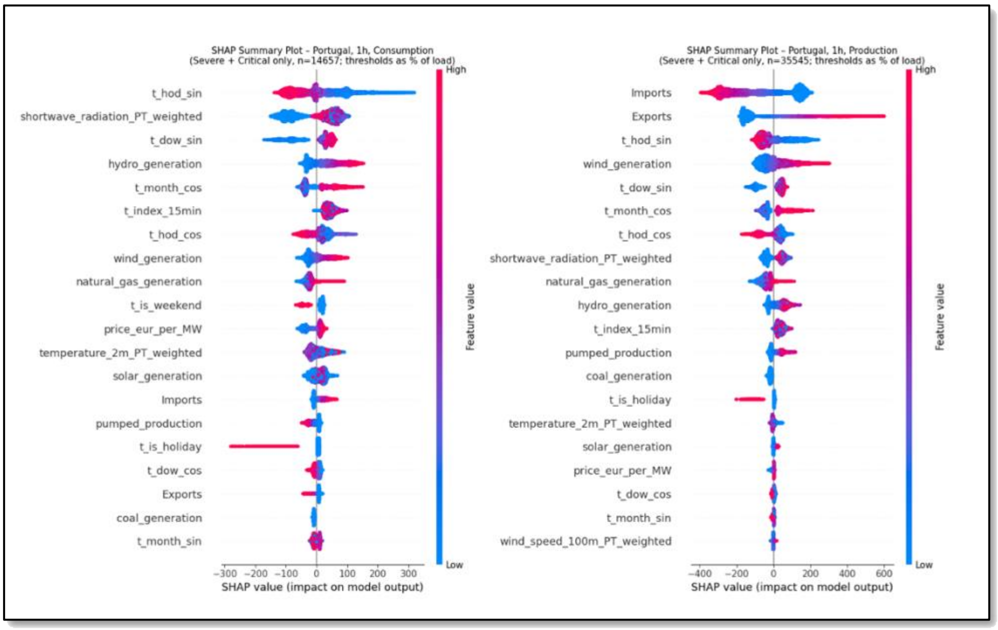
The impact of forecast horizon is also evident. At the 1-hour horizon, recent system conditions still limit forecast deviations, and SHAP distributions remain relatively compact. At the 24-hour horizon, renewable and weather-related features become more dominant, and SHAP distributions widen substantially, especially for production forecasts. Importantly, focusing on severe and critical errors reduces explanatory noise rather than increasing it, making the main drivers of extreme imbalance risk more clearly identifiable at longer horizons.

Overall, this severity-conditioned SHAP analysis supports the central conclusion of the thesis: high-risk net-imbalance forecast errors are primarily driven by uncertainty in renewable production, not by model instability. Germany's large and diversified system spreads this uncertainty across multiple factors, while Portugal's renewable-heavy and lower-inertia system concentrates forecast risk in a smaller set of weather-driven variables. These findings confirm that the observed severity patterns are rooted in physical system characteristics rather than in modelling artefacts.

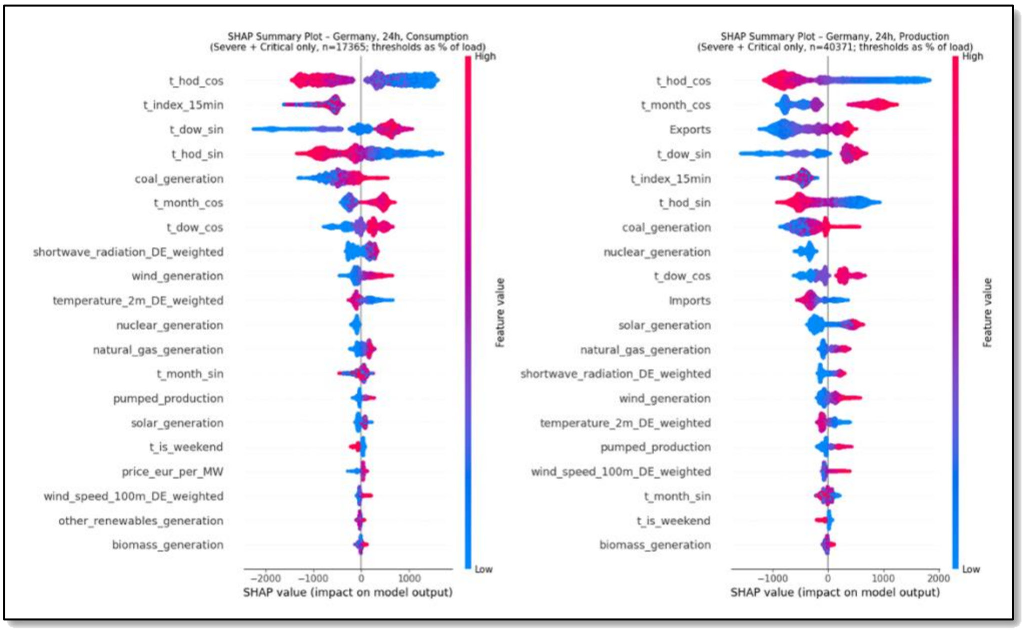
Germany 1h – Severe and Critical SHAP



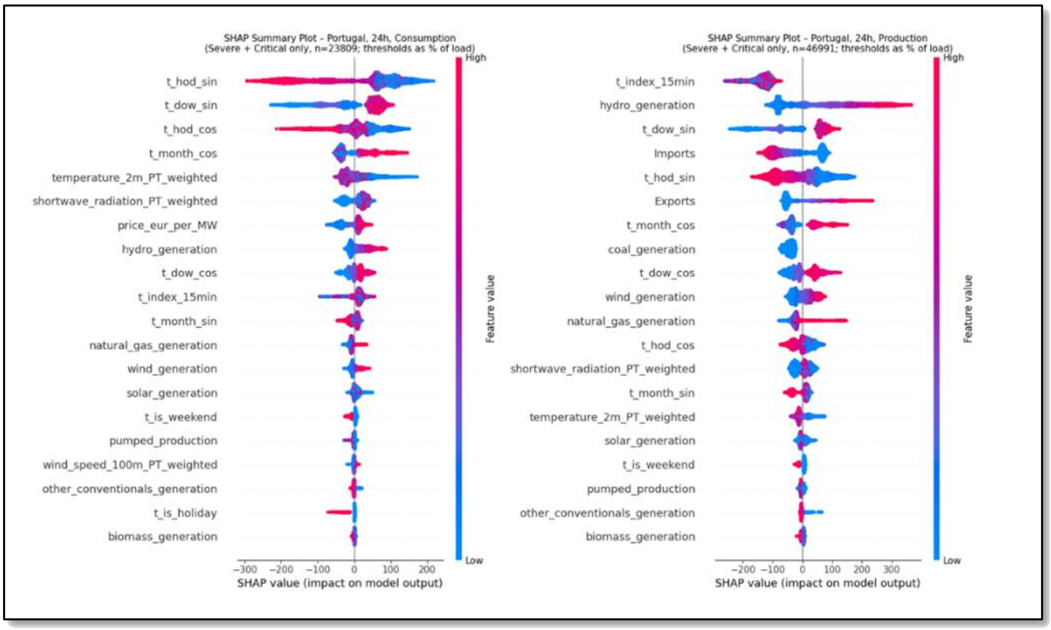
Portugal 1h – Severe and Critical SHAP



Germany 24h – Severe and Critical SHAP



Portugal 24h – Severe and Critical SHAP



Appendix E: Code Repository

All code used in this thesis, including data preprocessing, model implementation, and evaluation, is publicly available. The full repository can be accessed at:
https://github.com/PapaAryGit/MasterThesis

Appendix F: Use of AI

AI-based language assistance tools were used to support code completion, sentence rewriting, and brainstorming during the research and writing process. All methodological choices, data processing, model development, and interpretation of results were done independently by the author.