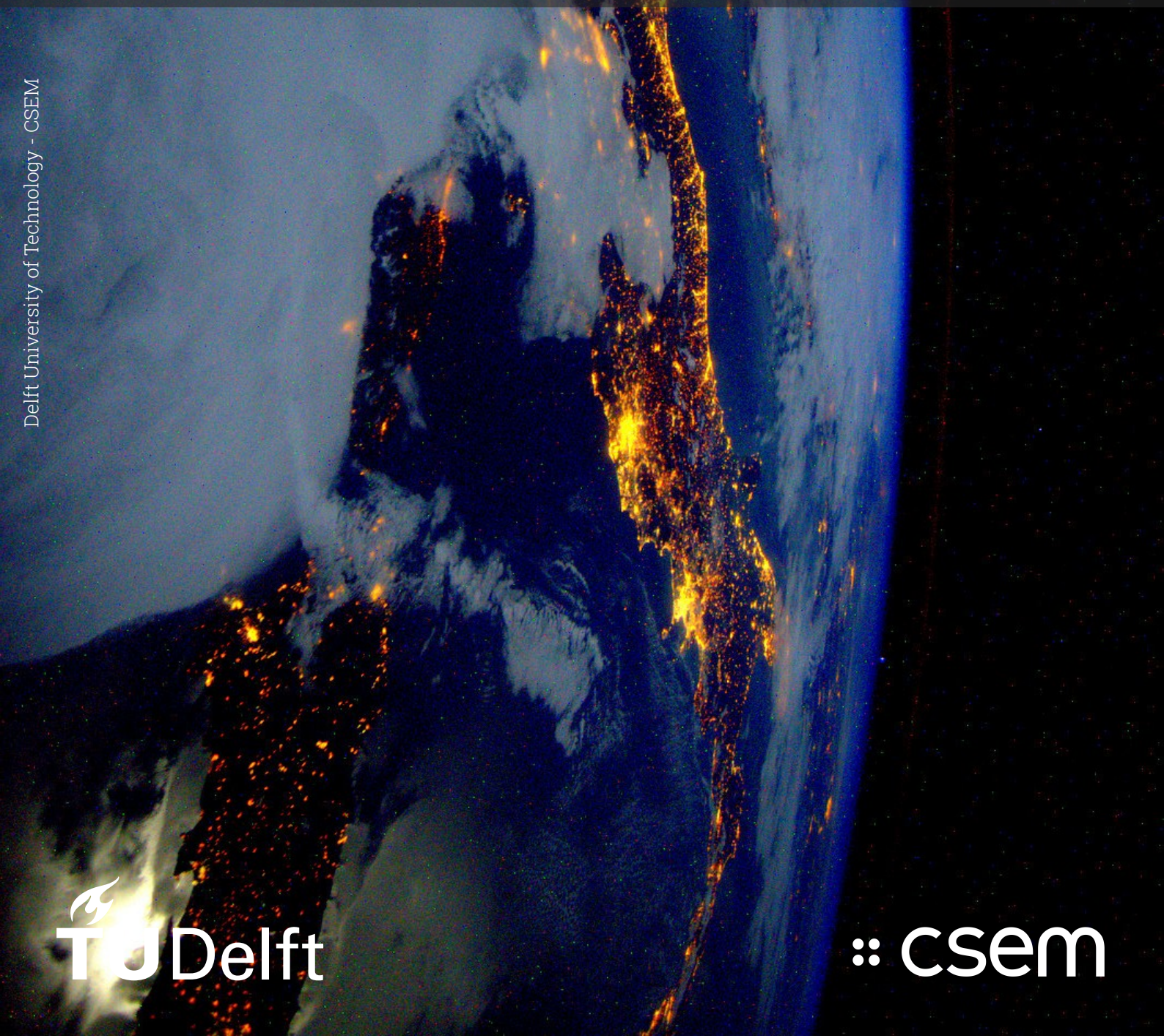
Grid Congestion Forecasting

Advanced Graph Neural Networks for Transmission Grid Congestion Forecasting

Samy VINCENT



Delft University of Technology - CSEM

TU Delft

:: csem

Grid Congestion Forecasting

Advanced Graph Neural Networks for Transmission Grid Congestion Forecasting

by

Samy VINCENT

to obtain the degree of Master of Data Science and Artificial Intelligence Technology
at the Delft University of Technology,
to be defended publicly on June 2nd 2026 at 15:00.

University Supervisor:	Elvin ISUFI
Company Supervisor:	Rafael Eduardo CARRILLO RANGEL
Project Duration:	October, 2025 - May, 2026
Faculty:	Electrical Engineering, Mathematics and Computer Science

Cover: By NASA / Scott Kelly - Public Domain

TU Delft

csem

Preface

This thesis presents the work carried out over the past eight months as part of my Master's degree. This period has been both intellectually stimulating and personally enriching, marking the conclusion of an important chapter of my life. The research was conducted within the Digital Energy Solutions team at CSEM, where I had the opportunity to work in an inspiring environment.

I would first like to express my sincere gratitude to Pierre-Jean Alet and Rafael Carrillo for giving me the opportunity to join their team and contribute to this project. I am especially thankful to Rafael for his guidance, availability, and the time and energy he dedicated throughout this process. His feedback and the many insightful discussions we had were invaluable to the development of this work. I would also like to thank Baptiste Schubnel for his technical expertise and support.

My thanks extend to the entire team at CSEM : Corentin, Paul, Diya, Max, Renault, Claire, Tomasz, and Jelena, for the welcoming atmosphere and the many moments shared during this experience.

I am also deeply grateful to my university supervisor, Dr. Elvin Isufi, for his support and valuable advice throughout the thesis. I would like to thank Dr. Pedro Vergara Barrios for his interest in my work and for taking the time to review this manuscript and participate in my thesis defense.

I am also thankful to Emil Soltanov and Federico Pianoforte working at BayWa r.e. for the insightful discussions and their valuable input as stakeholders, which helped refine my understanding of the Italian energy market and improve the design of the trading experiment presented in this thesis.

I would like to acknowledge the team at Electricity Maps, and in particular Dragos Petria, for providing access to their 2025 hourly flow forecasts in Italy, which allowed me to benchmark my model against industry standards.

Beyond the technical aspects, this experience has been a meaningful introduction to the professional world. I greatly appreciated the collaborative environment and the willingness to share knowledge and provide guidance throughout the project.

Finally, as this thesis marks the end of my studies, spanning France, the Netherlands, and Switzerland, I would like to express my heartfelt thanks to Emilie for her constant support along the way. I conclude this journey with a strong sense of accomplishment, having learned a lot and eager to continue learning in the years to come. I hope that whoever reads this thesis will find it both interesting and insightful.

Samy VINCENT
Neuchâtel, May 2026

Summary

As renewable energy penetration grows across Europe, transmission networks face increasingly frequent and unpredictable congestion, a problem costing an estimated 4.2 billion euros per year in Europe alone. Yet most data-driven forecasting tools either focus on generation or demand in isolation, or require access to proprietary grid models unavailable to market participants. This thesis addresses the question of whether interzonal power flows, and by extension, congestion, can be forecast purely from publicly available market and weather data.

The proposed approach is a spatio-temporal graph neural network that operates directly on transmission edges rather than nodes. It combines an LSTM encoder for temporal dynamics, a Transformer-based graph message-passing module with updatable edge representations, and a future-aware decoder that ingests day-ahead prices and weather forecasts. The model is trained and evaluated on Italy's seven-zone electricity market over a 2025 test year.

Against baselines ranging from naive persistence to gradient-boosted trees and LSTM, the model achieves the best normalized absolute error, directional accuracy, and congestion detection F1 score, with advantages that persist across all six forecast horizons. Critically, the proposed model achieves the best AUROC, demonstrating its ability to rank truly congested hours above non-congested ones regardless of where the threshold defining congestion is placed. Through comprehensive experiments and analyzes, the model proves to be more accurate than standard industry methods and domain-specific model. Further experiments demonstrate the quality of the forecast per edge and horizon, and break down the contribution of the different choices regarding its design. Moreover, the impact of future features is assessed, showing significant performance increase in congestion detection.

The central contribution to the renewable energy field is a reproducible, open-data framework that transforms observable market and weather signals into physically grounded flow forecasts, without access to network topology or TSO-proprietary models.

Contents

1

Introduction

This thesis addresses a growing problem: as renewable energy sources become a significant part of national electricity mixes, power systems face new patterns of congestion that are not fully captured by typical forecasting and risk-assessment approaches. Variable renewable energy generation, particularly photovoltaic and wind power, brings substantial environmental benefits and is essential for meeting decarbonisation goals, yet it also increases variability, uncertainty, and the likelihood of stress on transmission networks. This study focuses on the transmission system operator level, where interzonal flows, market coupling, and large-scale balancing are coordinated. The project is part of Task 5.5 of the European Union's Supernova project, which focuses on the interactions between photovoltaic power plants and the rest of the power system.

1.1. Grid failures and cascading outages

Interdependent infrastructure networks are inherently vulnerable to cascading effects: the failure of a small set of nodes in one network can propagate recursively and fragment coupled systems. This phenomenon was dramatically illustrated by the Italian blackout of 28 September 2003. The event originated from the loss of a few transmission lines in Switzerland, which rapidly propagated through the Italian grid, ultimately disconnecting the entire country from the European system [1].

As reported in [2], large-scale blackouts can often be initiated by the outage of a single transmission or generation element which, if not properly managed by automatic control systems or operator intervention, gradually leads to cascading outages and eventually to the collapse of the entire system. The electric blackout that affected Italy in 2003 induced severe degradation in several critical infrastructures, including the railway network, healthcare systems, financial services, and communication networks. At the same time, the partial failure of communication systems reduced the capability of the Supervisory Control and Data Acquisition (SCADA) network, responsible for managing the electric grid, to perform its function, producing a negative feedback loop that complicated and delayed the restoration phase [3, 2]. The dynamics of such "concurrent malfunctions" remain a canonical example of interdependent network cascades.

A more recent case occurred on 28 April 2025, when Spain and Portugal experienced a near-nationwide blackout. According to the ENTSO-E Expert Panel's final investigation [4], the event was classified as a Scale 3 event, the highest severity level. The crisis began at a substation in Granada, followed by rapid failures in Badajoz and Sevilla, eventually leading to the decoupling of the France–Spain interconnection. While initial data was obscured by communication failures similar to the 2003 event, the final report identifies several stressors, like inadequate voltage control (gaps in reactive power control and differing regulation practices across regions), system oscillations (unstable interactions that led to fast voltage increases), and rapid output reductions (sudden disconnections of generators in Spain that outpaced the system's stabilization capabilities). This 2025 blackout represents a "first of its kind" event. It underscores that while the triggers remain local, the implications are now increasingly dictated by the physical limits of the system and the variability of modern generation. The investigation concludes

that regulatory frameworks must evolve to support closer data exchange and improved monitoring of system behavior to mitigate these fast-developing cascades.

These two events highlight the intrinsic vulnerability of highly interconnected power systems. They show that cascading failures can develop within seconds, making them difficult to anticipate and contain. As the share of renewable energy sources increases, the variability and unpredictability of generation may further complicate this dynamic, introducing new stress patterns and congestion risks that can trigger or amplify such cascades. Understanding this link between renewable variability and grid stability is therefore essential for developing reliable forecasting and mitigation strategies, as discussed in the next section.

1.2. Renewable energy sources and congestion

The rapid deployment of renewable energy sources (RES) is essential to achieve decarbonisation goals and to mitigate climate change [5]. However, this transition also increases the variability and uncertainty of electricity supply across multiple time scales, from minutes to entire seasons. Because RES output is weather-dependent, system balancing increasingly relies on accurate meteorological information and the availability of flexible resources such as storage and demand response [6, 7, 8]. Being able to forecast renewable generation accurately is therefore necessary to maintain equilibrium in energy systems with high renewable penetration [9].

At the same time, high local concentrations of photovoltaic (PV) and wind generation can produce large power flows toward transmission exit points and overload lines or transformers that were not originally designed for such flow patterns: this phenomenon is commonly referred to as *congestion*. When power flows exceed line limits, conductors can overheat, leading to potential damage and automatic disconnection. Many existing transmission lines were built decades ago, and their physical limits make them increasingly vulnerable to overloads under high renewable penetration [10]. As a result, congestion is becoming more frequent in both national and regional grids. When congestion forces renewable curtailment or creates pronounced zonal imbalances, market signals and physical stress can combine to raise blackout risks if not adequately forecasted and managed.

The economic impact of congestion is already large and growing. Recent analyses estimate multi-billion-euro losses each year due to transmission constraints. In Europe, congestion costs were estimated at approximately 4.2 billion euros in 2023 [11], while in the United States they reached about $20.8 billion in 2022 [12]. These costs are accompanied by significant curtailments of renewable generation: in 2018, approximately 6,500 GWh of solar energy was curtailed across the United States, Germany, Chile, and China due to grid congestion and limited flexibility [13].

Moreover, congestion costs in the U.S. doubled between 2020 and 2021 and rose again by 56% between 2021 and 2022 [12], contributing to higher wholesale electricity prices and highlighting the urgent need for congestion solutions. These figures demonstrate that congestion is not a marginal operational issue but a major system-level stressor.

The increasing integration of intermittent RES thus introduces both opportunities and challenges: while essential for climate objectives, it also demands new methods for monitoring, forecasting, and managing grid congestion. The next section examines how these challenges are currently addressed through detailed power-flow modelling and highlights the limitations of such approaches in data-constrained contexts.

1.3. Motivation

Most academic and operational forecasting work in power systems focuses on either electricity demand or renewable generation in isolation, often using synthetic networks (built to match statistical characteristics found in actual power grids), or assuming complete knowledge of the physical grid and power flows [14, 15]. While these studies provide valuable methodological insights, they rarely address the practical constraints faced by real-world market participants, such as limited access to network data or the need for actionable forecasts under uncertainty.

There is comparatively little empirical research that explicitly links market observables, such as zonal prices, to physical grid conditions. Yet, these signals provide indirect but insightful information about

transmission congestion. Integrating them with exogenous variables like weather forecasts, renewable generation, or load data could enable early detection of congestion risk even when detailed grid models are unavailable [16, 17].

In practice, full AC power-flow modeling (relationship between voltages and power injections at nodes in an electric power system) remains the gold standard for transmission system analysis, but such models depend on confidential data that are typically restricted to TSOs and are seldom accessible to external researchers or market actors. The Day Ahead Congestion Forecast procedure is conducted by TSOs on an international scale, in which each of them generates a model of their network that is later shared with partners. Based on these models, TSOs have the right to change cross border power exchange schedule or internal production schedule for a particular period of the day in case of security constraints [18].

Consequently, there is a growing need for data-driven approaches capable of inferring congestion patterns from publicly available information. Market-based signals such as zonal price differentials offer a scalable and transparent alternative to physical flow models for congestion forecasting in data-limited settings [19, 20].

This thesis positions itself within this commercial and market-oriented perspective, focusing on methods that can support energy traders, aggregators, and market analysts who must operate without full visibility of the grid. By leveraging observable market and weather data, the aim is to provide a framework for anticipating congestion and assessing its market impact close to real time.

1.4. Research questions and contributions

In zonal markets with interzonal transmission lines, such as Italy's seven-zone market, price imbalances between zones act as indirect indicators of transmission congestion. Coupling these market-based signals with external predictors, such as numerical weather observations and historical flows, offers a path toward data-driven forecasting of congestion patterns without requiring access to proprietary grid models. This approach contributes to operational decision-making by transforming publicly observable data into physically grounded flow forecasts.

Why Italy? Italy's electricity market offers a particularly compelling test case for this research. Its seven-zone structure, with bidding zones that range from mainland regions to island interconnections, closely mirrors the multi-country architecture of the broader European electricity market, where each country plays a role analogous to an Italian zone and cross-border ENTSO-E flows serve as the congestion signal of interest. The combination of high and unevenly distributed renewable penetration, persistent north-to-south transmission bottlenecks, and a well-documented history of zonal price divergence makes Italy both a challenging and representative setting. Beyond its local relevance, results obtained on the Italian market can therefore inform the design of congestion forecasting tools at the European scale, which is one of the broader objectives of the Supernova project.

To address this objective, we pose the following research questions:

RQ1: Can physical interzonal flows be forecast from public market and weather data? We evaluate whether market prices and ERA5 weather observations, without any access to network topology or injection data, carry sufficient information to produce operationally useful multi-step flow forecasts.

RQ2: Does a graph-based architecture add value over per-edge models? We assess whether explicitly representing the Italian grid topology via message passing improves flow prediction compared to independent per-edge forecasting models, and how performance varies with the depth of the graph.

RQ3: Does the model translate into actionable congestion signals? We examine whether the regression output of the model can serve as a congestion score, and whether it carries tradeable predictive value beyond what intraday market prices already reflect.

The main contributions of this thesis are :

- **Flow-level forecasting from market data.**
  Existing congestion forecasting methods either require high network observability (physics-based) or target day-ahead price proxies that are not close to real-time events. We directly forecast interzonal physical flows using only publicly available market prices, NWP weather data, and historical flows, in order to obtain operationally relevant outputs.
- **Spatio-temporal graph architecture for edge-level targets.**
  Existing STGNNs in power systems focus on node-level tasks. We propose an encoder-decoder architecture combining LSTM temporal encoding with graph transformer message passing with explicit edge representation updates, operating directly on edge-level targets (interzonal flows) across all zone pairs simultaneously.
- **Open and reproducible framework scalable to the EU.**
  The pipeline relies exclusively on publicly available data sources (Terna, ENTSO-E, GME, ERA5) and open-source tooling. The trained model is evaluated on a full test year, and results are visualized in an interactive interface displaying predicted versus actual interzonal flows across all Italian zone pairs and neighboring countries (available at: `https://congestion-forecasting.portal.csem.ch/`). The framework generalizes directly to the EU-wide setting where each country acts as a zone, providing a scalable foundation for continent-level congestion forecasting.

1.5. Thesis outline

Chapter 2 provides the technical background: it covers Italy's zonal electricity market and its energy mix, the day-ahead and intraday market mechanisms that motivate the choice of input features, and the machine learning foundations underpinning the proposed architecture, including graph neural networks, graph attention, edge-level prediction, spatio-temporal extensions, and multivariate time series forecasting.

Chapter 3 surveys the related work across five methodological families: physics-based methods, econometric approaches, hybrid and deep learning architectures, spatio-temporal GNNs, and dedicated edge-level prediction methods. For each family the chapter identifies what has been achieved and why a gap remains for forecasting physical interzonal flows from market-observable data in a multi-zone system.

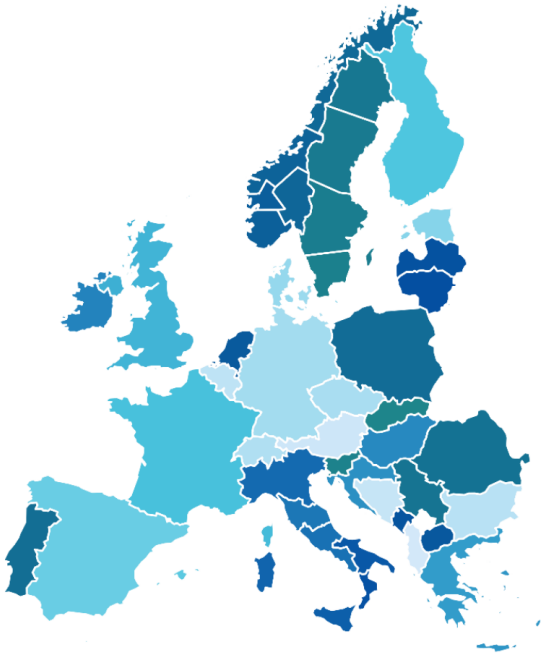
Chapter 4 describes the problem formulation, the target variable construction, and details the proposed encoder-decoder architecture, including the LSTM encoder, the graph message-passing block, the decoder design, and the training procedure.

Chapter 5 presents the experimental setup : data sources, preprocessing pipeline, the graph representation of the Italian grid. The evaluation of the model is presented as well: baseline comparisons, ablation studies on graph depth and future features, congestion classification analysis.

Chapter 6 discusses the findings in relation to the research questions, acknowledges the limitations of the study, and outlines directions for future work.

2

Background



Source: JRC based on (Electricity Maps, 2023)

Figure 2.1: Zonal electricity markets in Europe.

2.1. Italian electricity market & 7-zone structure

2.1.1. Zonal market concepts

The European Internal Electricity Market is organized according to a zonal market design, in which electricity supply and demand are matched at the level of predefined market zones, also referred to as bidding zones. Within each zone, the transmission network is assumed to be unconstrained for market clearing purposes, while transmission constraints are explicitly taken into account only at the borders between zones [21, 22].

Market clearing is performed independently for each zone, subject to available cross-zonal transfer capacities. As a result, a single wholesale electricity price is formed per zone, applying uniformly to all accepted bids within that zone. This design significantly simplifies market operations compared to

nodal pricing, but at the cost of neglecting internal network constraints during the clearing process [21].

Historically, market zones have largely coincided with national borders, reflecting the former national organization of European power systems. However, several countries operate multiple bidding zones. In the Nordic region, Denmark, Sweden, and Norway are subdivided into multiple zones, and it is also the case in Italy, see Figure 2.1. These zonal configurations either reflect natural geographical separations or persistent structural transmission constraints within the grid [21].

The European regulatory framework explicitly links the definition of bidding zones to congestion patterns. According to Article 14 of the EU Electricity Regulation, bidding zone borders shall be based on long-term, structural congestion in the transmission network. Structural congestion should not persist within zones, and corrective measures are intended to be temporary rather than a permanent substitute for appropriate zone delineation [21].

When actual power flows resulting from the market outcome are not physically feasible, Transmission System Operators apply remedial actions to restore system security. These actions include redispatch, whereby generation schedules are adjusted upward or downward after market clearing, and counter-trading, where energy is procured across zones to relieve congestion. In practice, redispatch is often used as a general term encompassing both mechanisms [21, 23].

2.1.2. Italy's zonal market

From the outset of market liberalization, the Italian electricity market has been modeled through a zonal structure. As described by Terna, Italy's Transmission System Operator, "the electrical system is divided into areas where producers and consumers can sell and buy electricity freely, while there are limitations on the buying and selling of energy between different zones" [24].

The adoption of market zones in Italy reflects both geographical characteristics and structural constraints of the transmission network. Natural separations, such as islands, as well as persistent bottlenecks in the mainland grid, motivated the subdivision of the territory. In addition, zonal pricing was introduced to differentiate electricity prices according to the local balance between generation capacity and demand, thereby providing location-specific price signals. Zones with relatively abundant generation tend to experience lower prices, whereas zones with higher demand and limited local generation face higher prices [23].

Italy applies an interzonal pricing mechanism, under which electricity prices are determined separately for each zone based on accepted bids and available transmission capacity between zones. This mechanism allows prices to reflect not only local supply and demand conditions but also the net physical power exchanges between zones. Compared to other liberalized electricity markets, Italy stands out for its extensive use of zonal pricing at the national level, featuring one of the highest numbers of bidding zones in Europe [23].

The current configuration consists of seven geographical market zones (Fig. 2.2): North (NORD), Center North (CNOR), Center South (CSUD), South (SUD), Calabria (CALA), Sicily (SICI), and Sardinia (SARD). These zones are defined by Terna and approved by the Italian Regulatory Authority for Energy, Networks and Environment, based on grid topology, demand distribution, and the spatial allocation of major generation assets [25]. Sicily and Sardinia form separate zones due to their electrical isolation from the mainland, while the mainland is subdivided to reflect internal transmission constraints and demand concentration patterns.

The introduction of the seventh zone, Calabria (CALA), represents a recent refinement of the zonal structure. This separation was motivated by persistent structural congestion between Southern Italy and the Sicilian interconnection corridor, as well as by the growing penetration of renewable generation in the southern regions. By isolating Calabria as a distinct zone, the market design more accurately captures local congestion patterns and improves the transparency of price signals, while reducing reliance on remedial actions such as redispatch [23, 25].

In addition to geographical zones, the Italian market includes virtual zones corresponding to interconnection points with neighboring countries. These virtual zones facilitate cross-border electricity trade and integrate Italy into the wider European electricity market, while maintaining control over congestion at national borders [25].

Figure 2.2: Zonal electricity markets in Italy.

2.1.3. Italy's energy mix

Italy provides a particularly relevant test case for studying congestion risks in systems with high renewable penetration. The country's power system combines a large share of PV generation, a zonal market structure, and transmission constraints that often manifest as price differentials between zones.

According to the most recent statistical data published by Terna (Italian TSO) for 2025, Italy's total installed renewable capacity reached approximately 80 GW, distributed as follows [26]:

- Photovoltaics: 41,1 GW
- Wind: 13.4 GW
- Hydropower: 21.3 GW
- Bioenergy and Geothermal: 4.9 GW

These figures highlight the central role of PV, which alone accounts for nearly half of Italy's renewable capacity.

PV capacity has grown rapidly over the last decade, supported by abundant solar resources and national incentives such as the *Conto Energia* schemes and declining module costs [27]. Between 2020 and 2023, the installed PV capacity increased by more than 30%, with over 310,000 new small and medium-sized installations connected to the grid during that period [28, 29]. The majority of these new plants are located in southern and central regions, where both irradiation levels and available land make these projects economically attractive.

This geographical pattern is illustrated with Italy's Global Horizontal Irradiance (GHI) in [30]. The highest solar resource potential is found in Sicily, Sardinia, Puglia, and Calabria, while northern regions have more limited but still viable solar potential. This uneven spatial distribution of generation, combined with historical transmission infrastructure designed for north-to-south power flows, contributes to localized congestion and persistent zonal price differences.

In Italy's zonal electricity market, these price differentials serve as a practical indicator of transmission bottlenecks. As a result, zonal price separation and curtailment events are increasingly common operational challenges for both system operators and market participants. These dynamics make Italy an ideal case study for exploring how renewable-driven congestion can be forecast and managed using market indicators.

2.1.4. Day-ahead and Intraday markets

Electricity trading in Europe is organized across multiple sequential markets, primarily the Day-Ahead, Intraday, and Balancing markets. Each market operates at a different time horizon and exhibits distinct characteristics in terms of time resolution, information availability, and price formation mechanisms. These differences result in unique challenges for market participants and for price analysis and forecasting [31].

While the day-ahead market aggregates expectations about next day system conditions, the intraday market allows participants to revise their positions closer to real time, when uncertainty about demand, renewable generation, and network conditions is significantly reduced. As a result, intraday prices should display stronger volatility and sharper congestion signals than day-ahead prices.

Day-ahead market

In the Italian Day-Ahead Market (Mercato del Giorno Prima, MGP), electricity is traded for delivery on the following day. Market participants, including generators, retailers, and large consumers, submit supply offers and demand bids specifying quantities and corresponding price limits. Supply offers indicate the minimum price at which a given quantity of electricity is willing to be sold, while demand bids reflect the maximum price buyers are willing to pay.

To formulate their bids, participants rely on internal forecasts and operational constraints. These typically include expectations about generation costs (such as fuel prices, start-up costs, and ramping constraints), plant availability, anticipated renewable production from wind, solar, and hydro resources, interconnection capacities, and expected electricity demand. In addition, strategic considerations related to competition and anticipated bidding behavior of other market participants influence bid formation.

Market clearing in the MGP results in zonal prices determined by the intersection of aggregated supply and demand curves, subject to interzonal transmission constraints. Importantly, the day-ahead market itself does not constitute a forecasting model. Rather, it is a market mechanism that aggregates the heterogeneous private forecasts and expectations of all participants into a single competitive equilibrium outcome. Consequently, the day-ahead zonal price can be interpreted as an implicit forecast: it represents the collective expectation of system conditions for the following day, rather than a prediction of real-time operational prices.

Intraday market

The Intraday Market allows participants to adjust their positions after the day-ahead clearing and closer to the delivery hour. This market exists precisely to address the residual uncertainty inherent in day-ahead trading, particularly related to electricity demand, renewable generation, and network constraints. As delivery approaches, updated forecasts and operational information become available, enabling market participants to refine their bids accordingly.

Intraday prices therefore reflect real-time corrections to the expectations embedded in day-ahead prices. At this stage, bids incorporate updated demand forecasts, improved predictions of renewable output, and emerging congestion conditions within the transmission network. The resulting price formation captures a more accurate and informed view of the actual system state.

Empirical studies indicate that, in intraday markets, historical price information from previous trading sessions often constitutes the most relevant explanatory input, while exogenous variables play a comparatively limited role in improving forecast quality [32]. This highlights the endogenous nature of intraday price dynamics, where market participants continuously react to prior price signals and revised expectations.

From an interpretative perspective, the distinction between the two markets can be summarized as follows:

- the day-ahead price represents an implicit forecast formed by market participants under significant uncertainty;
- the intraday price reflects the realized adjustment of these expectations and can be regarded as a closer approximation to the operational ground truth.

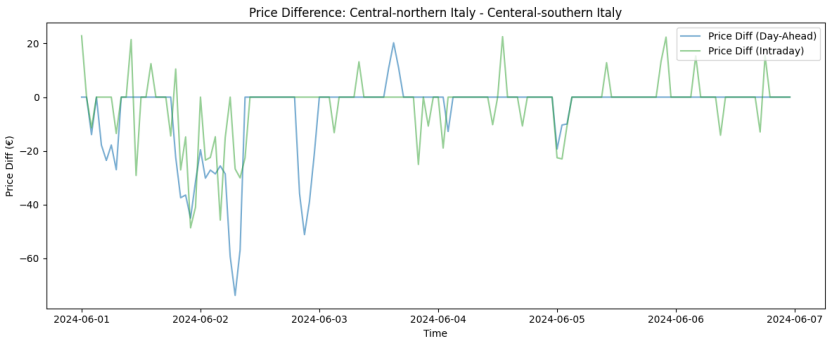


Figure 2.3: Volatility of smoothed Intraday and Day-ahead price differences between Central North and Central South zones. Absolute differences below 10€ were set to 0€ for readability.

This difference is particularly evident in zonal price spreads. While day-ahead prices often exhibit limited congestion signals, intraday prices frequently reveal pronounced zonal divergences as actual bottlenecks materialize closer to real time. Figure 2.3 illustrates this effect for the Central North and Central South zones, where intraday price differences display substantially higher volatility than their day-ahead counterparts.

2.2. Power system congestion and zonal market concepts

2.2.1. Congestion

In power systems, congestion arises when transmission constraints prevent electricity from being transported freely from generation areas to consumption areas. In economic terms, congestion reflects a situation where the desired power flows implied by market outcomes exceed the physical limits of the transmission network, requiring corrective actions to maintain system security.

From an operational perspective, congestion is fundamentally linked to the physical laws governing power flows and to the thermal, voltage, and stability limits of transmission lines. System operators therefore rely on technical indicators such as transmission capacity and line loading to assess and manage congestion. However, these indicators are not directly observable to market participants and are often difficult to forecast.

Several measures of transmission capability are used in practice. Total Transfer Capability (TTC) represents the maximum power that can be exchanged between two areas while respecting all operational constraints. Available Transfer Capability (ATC) is the remaining portion of TTC that is still available for commercial transactions after accounting for already allocated capacity and security margins. More recently, Flow-Based Market Coupling (FBMC) has been introduced in parts of Europe, relying on Power Transfer Distribution Factors (PTDFs) to model how transactions affect flows on critical network elements. This approach enables a more accurate representation of network constraints but requires detailed grid data and complex system modeling.

For parties other than Transmission System Operators, it is practically impossible to compute TTC, ATC, or flow-based capacities accurately. Although common European regulations define both the ATC and FBMC methodologies, their concrete implementation is carried out by TSOs and is not fully disclosed. TSOs retain degrees of freedom in the choice of models, parameter settings, and security margins to ensure system reliability, which limits the transparency of capacity calculation procedures. As a result, these quantities depend on a large set of variables, including line limits, network topology, generation dispatch, and PTDF matrices, that are not observable to market participants. [33] emphasizes that this lack of transparency makes it extremely difficult for market participants to forecast the transmission capacities that TSOs will make available for future trading intervals, thereby constituting a significant obstacle to the formulation of effective trading strategies.

Given these limitations, congestion can be approached from different conceptual perspectives. A physically grounded definition relies on actual power flows relative to line capacities. For instance, Løland et al. [34] define the Net Capacity Utilization (NCU) between two zones at time t as the ratio of

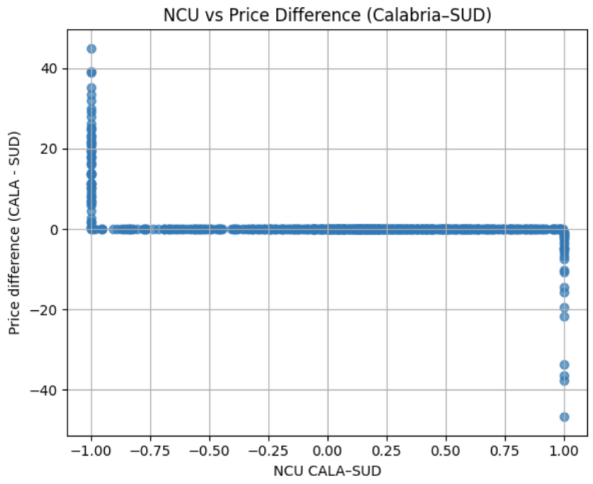


Figure 2.4: Net Capacity Utilization versus Day-Ahead price difference between Calabria and South zones.

realized power flow to available transmission capacity:

$$\text{NetCapacityUtilization}_{(A,B),t} = k \times \frac{\text{flow}_{(A,B),t}}{\text{capacity}_{(A,B),t}}, \qquad k = \begin{cases} +1, & \text{if zone } A \text{ exports to } B, \\ -1, & \text{if zone } A \text{ imports from } B. \end{cases} \tag{2.1}$$

This measure directly captures how intensively a transmission line is used and whether it operates close to its physical limits. An alternative and widely adopted approach defines congestion in economic terms, using price signals. In zonal electricity markets, congestion manifests itself through price differences between zones that exchange electricity. When transmission capacity between two zones becomes binding, the zonal prices diverge, reflecting the marginal cost of delivering electricity across the constrained interface. This definition aligns directly with the design of the Italian electricity market, where zonal pricing is explicitly intended to signal interzonal congestion.

Recent empirical studies on the Italian market therefore define congestion as the occurrence or magnitude of price spreads between connected zones [35]. This price-based definition is particularly suitable when the analysis focuses on market outcomes rather than on system operation. It leverages publicly available price data and captures the effective economic impact of congestion as perceived by market participants.

2.2.2. Price difference and congestion relationship

The relationship between transmission congestion and zonal price differences can be empirically illustrated by comparing price spreads with physical congestion indicators. Figure 2.4 shows the relationship between the Net Capacity Utilization (NCU) and the day-ahead price difference between the Calabria and South zones.

A striking correlation emerges in the day-ahead market. As soon as the NCU reaches its extreme values (NCU = ±1), indicating a binding transmission constraint, a zonal price difference systematically appears. When Calabria exports electricity (NCU > 0), the price difference decreases, corresponding to a lower price in Calabria relative to the South. Conversely, when Calabria imports electricity (NCU < 0), the price difference increases, reflecting a higher price in Calabria. Importantly, price differences are observed almost exclusively during congestion events, and congestion events are consistently associated with non-zero price spreads.

This observation provides empirical support for using zonal price differences as indicators of congestion in a zonal market framework. In the day-ahead market, price spreads appear tightly linked to physical transmission constraints, validating the interpretation of price divergence as a congestion signal.

However, while day-ahead prices capture anticipated congestion, they do not fully reflect real-time system conditions. Intraday markets are designed to incorporate updated information closer to delivery, including revised demand forecasts, renewable generation updates, and emerging grid constraints. Intraday prices are therefore expected to better reflect actual congestion conditions in real time.

In practice, intraday prices, particularly those observed shortly before delivery (for example XBID60 : market prices 60 minutes before physical delivery), exhibit substantial volatility and noise. This raises the question of whether observed intraday price variations reflect meaningful system information or merely stochastic trading behavior. Two main drivers can explain deviations of intraday prices from day-ahead prices. First, unexpected physical events such as demand peaks, forecast errors in renewable generation, or unanticipated congestion can induce genuine price adjustments. Second, intraday trading itself introduces noise: continuous trading and speculative behavior may generate price movements unrelated to underlying physical constraints, similarly to short-term fluctuations observed in financial markets.

Disentangling these two effects is challenging. The literature emphasizes that intraday electricity prices are influenced by both fundamental shocks and speculative trading dynamics, and that raw intraday prices may therefore be poorly suited as direct modeling targets [36]. As a consequence, the use of engineered targets, such as aggregated price measures or transformations designed to extract structural information, has been proposed to improve interpretability and predictability.

A further limitation arises from data availability. While NCU can be computed for the day-ahead market using day-ahead flows and transmission capacities, equivalent information is not available at the intraday time scale. Real-time physical flows, continuous intraday transfer capacities, and their temporal alignment with XBID prices are not directly observable or synchronized. Although offered intraday transfer capacities are published by ENTSO-E, they do not necessarily reflect realized physical constraints at the time of trading. As a result, replicating the NCU-based analysis for intraday prices is not feasible.

Given these constraints, congestion analysis in the intraday market cannot rely on price-based indicators alone. This approach inevitably aggregates both physical congestion effects and noise induced by trading, and preliminary tests showed significant hardships to overcome the noise and predict valuable information. They remain excellent features for the model to rely on, but the target must be physically grounded. After several trials to forecast intraday prices differences, it was decided to rather predict the interzonal flows. The challenge is therefore to design models capable of extracting the congestion-related signal embedded in intraday price dynamics.

2.2.3. Fundamental price drivers and input variables

Electricity spot prices are commonly modeled as the outcome of a wide set of fundamental drivers related to supply, demand, and system conditions. In addition to these drivers, electricity prices exhibit pronounced seasonal patterns at multiple time scales. Daily and weekly seasonalities are particularly relevant in short-term markets, while longer-term horizons are influenced by annual seasonality and trend-cycle components. The relevance of these patterns depends on the forecasting horizon: short-term models must account for intraday and weekly structures, whereas longer-term analyses may largely abstract from them [37].

Beyond seasonal effects, the literature identifies several key categories of explanatory variables, including system load (electricity demand and consumption), weather related variables (temperature, wind speed, solar radiation, precipitation), fuel costs (primarily natural gas and oil, and to a lesser extent coal), reserve margins (available generation capacity relative to expected demand), and planned or unplanned outages of generation units or critical grid components. Both historical realizations of these variables and their forecasts over the relevant horizon are typically considered essential inputs for electricity price forecasting models [37].

Weather variables play a dual role in modern electricity systems. On the demand side, temperature influences heating and cooling loads, while on the supply side it directly affects renewable generation, particularly wind and solar output. Fuel prices influence the marginal cost of thermal generation and are therefore expected to impact price formation, especially in systems where gas-fired units frequently set the marginal price. Reserve margins and outages capture system tightness and scarcity conditions,

which are known to be associated with price spikes and increased volatility.

However, the influence of fundamental drivers on electricity prices is neither constant over time nor uniform across markets. Empirical evidence suggests that, in certain periods, spot price dynamics may be dominated by factors that are only weakly related to observable fundamentals. For example, [36] shows that in the UK market during 2011–2012, traditional fundamental variables such as wind generation, demand, and gas prices explained only a small fraction of price variability, while speculative or price-driven shocks accounted for up to 95% of observed volatility. The distinction between fundamental shocks, originating from changes in demand, generation, or fuel costs, and speculative shocks driven by trading behavior, liquidity, and market microstructure is emphasized [36]. The relative importance of these components varies across markets and trading stages, with intraday markets in particular being more exposed to speculative dynamics due to continuous trading and shorter time-to-delivery.

As observed in the literature, the selection of input variables is therefore often guided by heuristics, prior experience, and data availability, rather than by a universally optimal rule. While pure price-based models exist, most short-term forecasting applications rely on a combination of price history and fundamental inputs. The optimal choice of variables remains an open question, and it is unlikely that a single universal set of inputs can be identified across different electricity markets and time horizons [37]. Consequently, in the present study, fundamental variables are included not under the assumption that they fully explain congestion dynamics, but to test empirically the extent to which they contribute to explaining and predicting congestion on the grid.

2.3. Machine and Deep learning for grid forecasting

2.3.1. What is a Graph Neural Network

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ is defined by a set of nodes $\mathcal{V}$ with $|\mathcal{V}| = N$, a set of edges $\mathcal{E}$, and a weighted adjacency matrix $A \in \mathbb{R}^{N \times N}$, where $A_{ij} > 0$ if nodes i and j are connected ($i \neq j$) and $A_{ij} = 0$ otherwise. Three types of features can be defined on a graph: node features $\mathbf{h}_i$ for $i \in \mathcal{V}$, edge features $\mathbf{e}_{ij}$ for $ij \in \mathcal{E}$, and a global graph feature vector $\mathbf{g}$.

A Graph Neural Network (GNN) is a class of neural networks designed to operate on graph-structured data. GNNs learn node, edge, and graph-level representations by iteratively propagating and transforming information along the edges of the graph, a process known as message passing.

In the general message passing framework, at each layer k, node, edge, and graph representations are updated as follows:

$$\mathbf{h}_i^{(k)} = f_v\left(\mathbf{h}_i^{(k-1)},\ \left\{\mathbf{h}_j^{(k-1)},\ \mathbf{e}_{ij}^{(k-1)} : j \in \mathcal{N}_i\right\}\right), \tag{2.2}$$

$$\mathbf{e}_{ij}^{(k)} = f_e\left(\mathbf{e}_{ij}^{(k-1)},\ \mathbf{h}_i^{(k)},\ \mathbf{h}_j^{(k)}\right), \tag{2.3}$$

$$\mathbf{g}^{(k)} = f_g\left(\mathbf{g}^{(k-1)},\ \left\{\mathbf{h}_i^{(k)} : i \in \mathcal{V}\right\},\ \left\{\mathbf{e}_{ij}^{(k)} : ij \in \mathcal{E}\right\}\right), \tag{2.4}$$

where $\mathcal{N}_i$ denotes the neighborhood of node i, and f_v, f_e, f_g are learnable functions. This formulation generalizes to a wide range of GNN architectures and allows simultaneous learning of representations at different levels of the graph hierarchy.

An early and widely used instantiation of this framework is the Graph Convolutional Network (GCN), which simplifies the node update by discarding edge features and using degree-based normalization:

$$\mathbf{h}_i' = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{\circ(i)\circ(j)}} \left(W\mathbf{h}_j\right) + \mathbf{b}, \tag{2.5}$$

where $\circ(\cdot)$ denotes the node degree, W is a learnable weight matrix, and $\mathbf{b}$ is a bias term. For the entire graph, this operation can be expressed as

$$H' = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}HW\right), \tag{2.6}$$

where $\hat{A} = A + I$ includes self-loops, $\hat{D}$ is the corresponding degree matrix, and $\sigma(\cdot)$ is a nonlinear activation function.

GNNs introduce a relational inductive bias by explicitly modeling interactions between connected components. This property is particularly suitable for power grids, where phenomena such as congestion propagation and cascading failures follow the physical network topology.

2.3.2. Graph attention

Graph Attention Networks (GATs) [38] extend the message passing framework by introducing a learnable attention mechanism to weight the contributions of neighboring nodes. Rather than relying on fixed degree-based normalization as in GCN, GAT computes attention coefficients that reflect the relative importance of each neighbor.

Given node features $\mathbf{h}_i \in \mathbb{R}^F$, the updated node representation is computed as

$$\mathbf{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W \mathbf{h}_j\right), \tag{2.7}$$

where $W \in \mathbb{R}^{F' \times F}$ is a shared learnable weight matrix and $\sigma(\cdot)$ is a nonlinear activation function. The attention coefficient α_{ij} is obtained by normalizing unnormalized scores over the neighborhood of node i:

$$\alpha_{ij} = \mathrm{softmax}_j\big(\mathrm{attention}(W\mathbf{h}_i,\, W\mathbf{h}_j)\big)\,. \tag{2.8}$$

In the original formulation, the attention function is implemented as a single-layer feedforward network applied to the concatenation of the transformed features:

$$\alpha_{ij} = \frac{\exp\big(\mathrm{LeakyReLU}\big(\mathbf{a}^\top [W\mathbf{h}_i \,\|\, W\mathbf{h}_j]\big)\big)}{\sum_{k \in \mathcal{N}_i} \exp\big(\mathrm{LeakyReLU}(\mathbf{a}^\top [W\mathbf{h}_i \,\|\, W\mathbf{h}_k])\big)}, \tag{2.9}$$

where $[\cdot \,\|\, \cdot]$ denotes concatenation and $\mathbf{a} \in \mathbb{R}^{2F'}$ is a learnable parameter vector.

To improve expressiveness and training stability, GATs commonly employ multi-head attention. Using K independent attention heads with weight matrices W_k, the update rule becomes

$$\mathbf{h}'_i = \sigma\left(\sum_{k=1}^{K} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W_k \mathbf{h}_j\right). \tag{2.10}$$

The attention mechanism allows the model to focus selectively on the most relevant neighbors, which is particularly valuable in heterogeneous graphs where nodes have varying degrees of influence on their surroundings.

2.3.3. Edge-level prediction

While many GNN architectures are designed for node-level tasks, several real-world problems require predictions defined on edges. In power systems, quantities such as interzonal flows, line loadings, and congestion states are inherently edge-centric: they describe the relationship between two connected nodes rather than a property of a single node.

The general message passing formulation introduced in Section 2.3.1 naturally supports edge-level reasoning. By maintaining and updating explicit edge representations $\mathbf{e}_{ij}^{(k)}$ throughout the forward pass, the model can produce edge-level outputs directly, rather than deriving them as a post-hoc combination of node embeddings. This is particularly important for signed quantities such as directed power flows, where the order of the incident nodes matters and symmetric aggregation would wash out directional information.

In practice, deriving edge predictions from node embeddings alone, typically by concatenating source and destination representations, remains a common but limited approach. More principled architectures maintain first-class edge representations and handle orientation explicitly. A detailed discussion of dedicated edge-level methods and their relation to this work is provided in Section 3.3.3 of the related work chapter.

2.3.4. Multivariate Time Series Forecasting

Time series forecasting aims to predict future values based on historical observations. In power system applications, forecasting tasks include load, generation, and congestion prediction, and can be formulated as either value-based regression or classification problems. In both cases, predictions are made from past time series observations.

Day-ahead forecasting methods can be broadly categorized into three approaches [39]: (i) *rolling forecasting*, where one-step-ahead predictions are iteratively fed back as inputs; (ii) *direct forecasting*, where separate models are trained for each future time step; and (iii) *multi-step forecasting*, where a single model directly predicts a sequence of future values over a fixed horizon. The third approach is commonly adopted in modern deep learning frameworks due to its efficiency and ability to capture temporal dependencies across the prediction horizon.

In many real-world settings, time series do not exist independently. For instance, in power grids or weather systems, multiple correlated entities evolve simultaneously. Instead of forecasting a single sequence, the goal is to predict multiple time series jointly while exploiting their interdependencies. This setting is referred to as *multivariate time series forecasting*.

Let N denote the number of time series and F the number of features per series. At time step t, the observation matrix is

$$X_t \in \mathbb{R}^{N \times F}. \tag{2.11}$$

Using a sliding window of length w, the input sequence is defined as

$$X_{t-w:t} = [X_{t-w}, \ldots, X_{t-1}, X_t] \in \mathbb{R}^{N \times w \times F}. \tag{2.12}$$

The objective is to predict a future sequence over a horizon of h time steps:

$$\hat{Y}_{t+1:t+h} = [\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+h}] \in \mathbb{R}^{h \times N \times d_0}, \tag{2.13}$$

where d_0 denotes the output dimension. In the common case where a single value is predicted per time series, i.e., $d_0 = 1$, the prediction reduces to

$$\hat{Y}_{t+1:t+h} \in \mathbb{R}^{h \times N}. \tag{2.14}$$

By modeling multiple time series jointly, multivariate forecasting methods can exploit shared structure and correlations between entities. When combined with graph-based representations, this formulation naturally enables the integration of spatial dependencies, which is particularly relevant for power grid forecasting.

2.3.5. Spatio-temporal extension of GNNs

Many real-world forecasting problems involve both spatial and temporal dependencies. Graph neural networks are effective at modeling spatial structure through message passing, while time series models capture temporal dynamics. Power systems forecasting inherently requires both: electrical interactions are constrained by grid topology, and system states evolve over time due to demand patterns, weather conditions, and operational decisions.

Spatio-temporal graph neural networks (STGNNs) address this challenge by combining graph-based representations with temporal modeling. A systematic literature review by [40] categorizes STGNNs according to how spatial and temporal components are integrated, highlighting their effectiveness for time series forecasting and classification tasks on structured networks.

A common approach consists of combining GNNs with sequence models such as recurrent neural networks. In this setting, graph convolutions are used to model spatial interactions at each time step, while temporal dependencies are captured using recurrent units such as LSTMs or GRUs. This modular design allows spatial and temporal patterns to be learned separately. Such architectures are widely adopted in power system applications, including multi-site photovoltaic power forecasting, where spatial correlations between geographically distributed sites are coupled with strong temporal dynamics [41].

Introducing an encoder-decoder model is an interesting improvement to better include temporal interactions in a GNN. An encoder takes as input a graph and outputs embeddings that maps nodes and edges features to hidden representations. A decoder takes as input the embeddings and makes predictions. What is interesting in our case is to input weather observations in the encoder, and weather forecast in the decoder. Thus, the encoder is in charge of processing the past whereas the decoder process the future. It is possible to train the two components end-to-end. In such cases, the parameters of the encoder and the decoder are typically initialized randomly. Then, until some criterion is met, several epochs of stochastic gradient descent are performed where in each epoch, the embedding function is produced by the encoder, predictions are made based on the embedding function by the decoder, the error in predictions is computed with respect to a loss function, and the parameters of the model are updated based on the loss. Inserting the GNN between the encoder and the decoder allows the final model to be aware of the past temporal interactions, of the spatial interactions, and of the future temporal interactions.

3

Related Work

Transmission congestion forecasting sits at the intersection of power systems engineering, market analysis, and machine learning. This chapter surveys the main methodological families: physics-based approaches, econometric and price-based methods, deep learning architectures for time series, spatio-temporal graph neural networks, and edge-level prediction. For each family we identify what has been achieved and why a gap remains for forecasting physical interzonal flows from publicly observable market data in a multi-zone system such as Italy.

3.1. Physics-based methods

Physics-based methods model the network explicitly (nodes, lines, impedances, and generator constraints) and compute power flows under different operating scenarios [14].

Probabilistic power-flow and stochastic security-constrained optimal power flow (SCOPF) variants extend this to estimate the distribution of flows and hence the probability that a given interface becomes congested under uncertain renewable injection or demand [15]. These methods allow TSOs to quantify congestion risk under different weather or generation scenarios with high physical fidelity.

Ji et al. [42] introduced a probabilistic framework using multiparametric programming, partitioning the uncertainty space into critical regions to derive conditional distributions of real-time locational marginal prices (LMPs) and congestion states. The approach efficiently separates offline and online computations, making probabilistic congestion forecasting computationally tractable. Building on this idea, Hernandez-Matheus et al. [43] proposed a hybrid method for DSOs that integrates probabilistic power flow with machine learning to forecast congestion in smart distribution grids. By sampling uncertain operating conditions and training ML models on the resulting probabilistic flow data, they demonstrated faster and accurate congestion risk estimation in networks with high renewable penetration.

Recent work has extended these approaches by combining physical models with deep learning. For instance, a cascaded LSTM–DNN framework [44] uses an optimization-based label generation process to train a surrogate model capable of real-time congestion prediction. This hybrid method retains the interpretability of physics-based models while achieving instantaneous runtime performance. Similarly, the probabilistic co-planning approach of distributed series reactors and dynamic line ratings [45] expands probabilistic SCOPF to the planning level, jointly optimizing infrastructure upgrades and operational constraints to manage congestion under high renewable integration.

However, these solutions require complete and accurate network topology data, impedance parameters, and real-time unit-commitment information that are rarely accessible to market participants or external researchers. Because our setting is explicitly data-limited, and relying only on publicly available market prices, weather forecasts, and reported flows, physics-based methods are not directly applicable, so a data-driven alternative is needed.

3.2. Econometric approaches

Zonal price differences are widely used as a proxy for interzonal congestion in markets where physical flow data are unavailable or coarse [35]. Statistical and econometric methods exploit correlations between these price spreads and observable variables such as renewable generation, demand, and calendar effects [46, 47].

For the Italian market specifically, a multivariate regression was used in [46] to quantify how wind and solar generation influence zonal prices in Italy over 2015-2019. Their results confirm the merit order effect: more renewables tend to reduce average prices, though with heterogeneous effects across zones. Similarly, [47] directly models interzonal congestion states, defined by price differentials between zones, as a function of renewable generation, local demand, and calendar variables. They find that increased renewable generation tends to lower congestion probability for importing zones but can raise it for exporting ones, reflecting direction-dependent market responses. More advanced econometric models extend this line of work by addressing volatility and uncertainty. [48] employed a heteroscedastic additive model to jointly forecast zonal prices and demand in Italian zones, capturing both mean and variance dynamics for one-day ahead forecasting

In the Nordic market, traditional time series models such as ARIMA and VAR have been widely used to estimate congestion probabilities from demand and production indicators [34], demonstrating that market-level factors can signal interzonal bottlenecks without detailed grid data. Electricity price forecasting more broadly has been approached with ARIMA, GARCH, and factor models [49, 48]. Simple approaches such as exponential smoothing remain effective baselines for short-term forecasting [50, 51, 52].

Recent comparative studies bridge econometric and machine learning methods. For example, [53] compare statistical models (SARIMA, VAR) with deep neural networks (LSTM, CNN-LSTM) for day-ahead price forecasting in Germany, showing that neural models often outperform classical methods at shorter horizons, especially when external regressors such as weather and fuel variables are included. Likewise, [54] propose an interpretable hybrid framework combining seasonal trend decomposition, Gated recurrent unit, Light gradient boosting machine, and Shapley additive explanations to explain factor contributions, achieving improvements in volatility and extreme price event forecasting in the United States and Australia.

These econometric approaches are interpretable and feasible under limited observability, but they are designed around price outcomes rather than physical flows, rely on mostly linear temporal structures, and do not capture spatial correlations between zones. Our work builds on the insight that price signals contain congestion information, but targets physical flows directly and uses a spatially-aware model.

3.3. Machine and Deep Learning methods

3.3.1. Hybrid and Deep Learning Architectures

With the increasing availability of high-frequency market and weather data, data-driven and deep learning methods have become key tools for forecasting tasks in energy systems. Hybrid architectures combining convolutional and recurrent components have been particularly successful in modeling the spatio-temporal complexity of these systems. Convolutional Neural Networks (CNNs) extract spatial or local correlations such as dependencies among neighboring grid areas or weather features, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, capture temporal dynamics [55, 56].

Even if deep learning has only recently been applied to congestion forecasting, early attempts such as [57] already explored neural network models for real-time congestion management in zonal markets. These studies used historical demand and generation data to approximate total transfer capacity between zones, offering an initial demonstration of data-driven congestion prediction, albeit with limited spatial modeling capabilities. Later works on cross-border transmission forecasting using Artificial Neural Networks (ANNs) in Central Western Europe also reported mixed performance [33], underscoring the challenge of learning accurate congestion patterns from limited or noisy data. The authors stressed the need for more detailed understanding of the structure and its relation to the underlying market characteristics in order to obtain accurate forecasts.

A representative example, although not in congestion, is PVNet [58], which integrates convolutional and LSTM layers to process Numerical Weather Prediction (NWP) maps for large-scale photovoltaic power forecasting. The CNN component learns spatial weather patterns, which are then passed to the LSTM to model their temporal evolution, resulting in improved accuracy for country-level PV output forecasts. Similarly, CNN–LSTM hybrids have been applied to load and demand forecasting [59, 60], where convolutional layers extract local feature patterns from historical demand and exogenous inputs (temperature, holidays), and recurrent layers handle longer-term dependencies such as daily or weekly cycles.

Encoder–decoder frameworks, initially developed in natural language processing, have also found application in time-series forecasting within the energy sector [61, 56]. In these architectures, the encoder (often an LSTM or GRU) summarizes past information into a latent representation, while the decoder generates future predictions, sometimes conditioned on exogenous variables. When combined with CNN or attention mechanisms, such architectures can effectively model multi-step forecasts and complex spatio-temporal correlations [62].

Similarly, the Temporal Fusion Transformer (TFT) [63] combines recurrent encoders, attention layers, and gating mechanisms to capture long-term dependencies while maintaining interpretability. The model explicitly learns which features (such as demand, renewable generation, or interzonal capacity) are most influential at different forecast horizons, making it well-suited to congestion prediction tasks where feature relevance varies dynamically with system conditions.

While these architectures achieve strong performance for demand or RES generation forecasting, they do not explicitly represent the spatial structure of the transmission network. It is a key limitation when the forecasting target is an interzonal flow that couples multiple zones simultaneously.

3.3.2. Spatio-temporal Graph Neural Networks

Graph Neural Networks for Power Systems

Graph Neural Networks (GNNs) are a natural modeling choice for power system tasks with inherent spatial structure: instead of treating each measurement or zone as independent, they represent the system as a graph where nodes denote substations, regions, or generation sites, and edges capture physical, geographical, or logical linkages [64, 65, 66]. Through iterative message passing, each node aggregates information from its neighbors, thereby capturing spatial interdependencies that standard time-series models ignore.

In power system applications, hybrid architectures combining graph convolutions with temporal models further extend this representational capability. Line-graph convolutions coupled with LSTMs have been used for probabilistic transmission line rating forecasts [67], while PowerGNN [68] integrates a GraphSAGE encoder with GRU modules to predict grid operating states under high renewable penetration, jointly exploiting topology and temporal consistency. At a broader scale, [69] demonstrated graph-based representations supporting multiple simultaneous tasks (consumption, generation, and asset condition forecasting) in heterogeneous energy networks, with probabilistic voltage outputs that can be used to estimate violation likelihoods directly related to congestion risk.

GNN-based approaches have also demonstrated accurate AC power flow prediction by learning flow distributions directly from grid topology [70, 71], showing that spatial structure alone is sufficient to recover system states with high fidelity without solving the full optimal power flow problem. However, these methods operate as static snapshot predictors: given complete nodal injections and topology, they estimate the resulting flows. This assumes full network observability and does not address temporal forecasting or the data-limited regime where topology and injection data are unavailable. Our work targets precisely this gap: rather than emulating a power flow solver, we forecast interzonal flows from publicly observable market prices and weather signals, using an edge-level spatio-temporal GNN that operates without access to network parameters.

Spatio-Temporal GNNs for Energy Forecasting

A major strand of recent research focuses on the fusion of heterogeneous spatio-temporal data sources, particularly Numerical Weather Prediction (NWP) outputs, satellite imagery, and historical generation or demand data, to improve forecast accuracy at longer horizons where single-source models struggle

[72, 73]. Because renewable generation and electricity demand are inherently spatio-temporal processes driven by weather dynamics, their accurate forecasting requires models that integrate both spatial correlations and temporal dependencies across multiple data sources.

Photovoltaic power forecasting illustrates the benefits of this fusion clearly. [74] showed that NWP information becomes increasingly valuable beyond 12-hour horizons, while locally observed or satellite-based features dominate short-term forecasts, hybrid models fusing NWP, satellite imagery, and historical PV generation further improve robustness under variable irradiance conditions [75]. Spatial correlation is equally important: wind generation exhibits dependencies over distances of several hundred kilometers [76], meaning that spatially correlated generation surges or deficits (precisely the situations that trigger congestion) can be captured by models that explicitly represent geographic structure. To address this, [77] proposed treating geographically distributed PV plants as nodes in a graph and learning to propagate information across them through graph convolutions coupled with LSTMs or Transformers, capturing both spatial interdependencies and temporal evolution. For data-limited settings, compressive spatio-temporal models [78] leverage spatial correlations across weather stations to maintain predictive accuracy even when observations are incomplete, which is a relevant consideration for Italy's unevenly distributed renewable production. More recently, diffusion-based GNN architectures such as GraphDiffusion [79] and SAGDFN [80] integrate GNN encoders with diffusion processes to jointly model spatial correlations and temporal uncertainty, offering promising directions for probabilistic forecasting where both the likelihood and intensity of events must be estimated. A systematic review by [81] confirms that spatio-temporal GNNs have driven substantial improvements across domains including renewable energy, traffic, and meteorology, highlighting the potential of these models for grid congestion forecasting.

These methods establish that graph-based spatio-temporal fusion yields clear accuracy gains for generation and demand forecasting. Yet their application to interzonal congestion forecasting remains largely unexplored. Existing studies target physical quantities like nodal voltages, line loadings, generation output, rather than market congestion signals driven by the interplay of prices, flows, and zonal capacities. The fusion of market-level variables with spatio-temporal weather data is rare, and edge-level prediction (forecasting the flow on a specific interconnection rather than a nodal quantity) receives little attention. Our work addresses these gaps by building an edge-level spatio-temporal GNN that combines market and weather inputs to forecast physical interzonal flows in Italy's multi-zone system without requiring access to network topology or injection data.

3.3.3. Edge-level Prediction

While most GNN research focuses on node-level tasks, several recent works address learning and prediction directly on edges. This is particularly relevant in domains where the quantities of interest are associated with relationships between nodes rather than the nodes themselves. Link prediction and edge regression have been explored in traffic networks where edge flows correspond to road segment volumes and in knowledge graphs. In power systems, edge-level tasks include line flow estimation and congestion state classification, but these are typically solved as post-hoc analysis problems with full network observability rather than as forecasting problems from market data. This mismatch has begun to attract dedicated methodological attention, though the field remains emerging.

A foundational challenge for edge-level prediction is directionality. Physical flows on a transmission interface are signed quantities: the same edge carries different information depending on whether power flows from north to south or south to north. Continuous Edge Direction [82] addresses this by assigning a learnable continuous phase parameter to each edge, allowing the effective direction of information propagation to adapt during training. Through a complex-valued Laplacian representation, the model separates incoming and outgoing signals, avoiding the feature homogenization that standard symmetric message passing produces in deeper networks. This is directly relevant to interzonal flow forecasting, where the sign of the flow determines whether a zone is importing or exporting, and where symmetric aggregation would wash out that distinction.

Beyond directionality, a second challenge is that standard GNN message passing is designed to produce node embeddings, and deriving edge predictions from those embeddings is typically a post-hoc operation (like concatenating the source and destination node representations) rather than a first-class architectural concern. [83] formalize this gap by studying GNNs for edge-valued signals that depend

on edge orientation, introducing the notions of orientation equivariance and invariance, and proposing message-passing schemes that aggregate information from neighboring edges while respecting their relative orientation. Their work provides a theoretical grounding for why existing node-centric architectures are not well suited to learning signed flow quantities, and motivates architectures that maintain and update explicit edge representations throughout the forward pass.

The temporal dimension adds further complexity. [84] study Temporal Edge Regression in dynamic graphs, benchmarking heuristic methods, static GNNs (GraphSAGE, GAT), and temporal graph models (TGN) on continuously evolving edge values. Their main finding is that existing architectures perform similarly and that none is specifically optimized for edge regression, leaving the task as an open design problem.

Taken together, these works identify three properties that an architecture for interzonal flow forecasting must satisfy: explicit and updatable edge representations, orientation awareness for signed flows, and a temporal model capable of multi-step ahead inductive prediction. Our encoder-decoder architecture addresses all three: the LSTM encoder produces directed edge embeddings initialized separately for each flow direction, the explicit edge update propagates and refines edge representations through graph message passing, and the decoder produces a full forecast horizon conditioned on future exogenous inputs. To our knowledge, no prior work combines these properties for edge-level regression from market-observable data in a power system graph.

3.4. Congestion forecasting in zonal markets

Research on transmission congestion forecasting and zonal electricity market dynamics has been pursued across several power systems worldwide, though with varying focus, data availability, and methodological depth. Most early studies emerged in nodal markets such as the Nordic and U.S. systems, where congestion pricing and locational marginal pricing (LMP) provide direct visibility into grid bottlenecks. In contrast, studies focusing on Italy's zonal market, while increasing in recent years, remain relatively limited in number and often rely on traditional econometric or machine learning techniques.

In the Nord Pool market, one of the earliest contributions is [34], which developed statistical models to forecast transmission congestion probabilities between bidding zones. Their approach used demand and production indicators to estimate congestion occurrence, illustrating how market-level factors could signal bottlenecks in interzonal transmission.

In the United States, various studies have addressed congestion forecasting from market perspective. For example, [85] proposed a machine learning ensemble framework for grid congestion price forecasting. Their model combines gradient boosting, random forests, and neural networks to produce day-ahead price forecasts that more closely match real-time market outcomes than the official market predictions. By incorporating weather variables, load fluctuations, and historical nodal prices, their ensemble approach captures both the temporal and spatial variability underlying congestion-driven price differences. Similarly, [86] analyzed locational marginal prices across U.S. power markets to quantify the economic value of transmission and identify its key drivers. Since LMPs represent the marginal cost of serving an incremental unit of demand at a given node, including components for energy, losses, and congestion, the study decomposed these elements to understand how transmission constraints shape regional price disparities. Together, these works demonstrate how data-driven approaches can anticipate congestion and quantify its economic implications within nodal electricity markets.

In China, studies have explored probabilistic congestion warning systems under rapidly evolving renewable integration. For instance, [87] proposed a data-driven early warning method for grid congestion probability based on multi-timescale features. Although promising, this work lacks detailed data disclosure and focuses mainly on high-level correlation patterns rather than precise spatio-temporal modeling.

Turning to Italy, research on congestion forecasting and zonal market dynamics is more limited, despite the Italian grid's unique structure of seven interconnected bidding zones and substantial international transmission constraints. Early work such as [88] introduced a neural network approach to estimate Transmission Transfer Capability (TTC) between zones, aiming to support real-time congestion

management and reduce redispatching costs. Later, [89] applied neural networks and support vector regression (SVR) to forecast Italian day-ahead electricity prices, incorporating zonal prices, projected RES generation, and demand forecasts as explanatory variables. While not explicitly focused on congestion, their model implicitly captured zonal imbalances driven by renewable fluctuations.

More recent work, such as [35], conducted an empirical analysis of interzonal congestion in the Italian electricity market using a multinomial logistic regression framework. Their study quantitatively linked zonal renewable generation, demand, and transmission capacities to congestion probabilities between zone pairs, providing one of the first systematic congestion modeling efforts for the Italian market.

In addition, research on demand and renewable forecasting in Italy, such as [90], developed and tested a LSTM model capable of predicting national hourly electricity demand with a Root Mean Squared Error (RMSE) consistently below 2%. The model demonstrated strong alignment with official Terna data, accurately reproducing both seasonal patterns and daily peaks over short- and medium-term horizons. Although primarily designed for load forecasting, the study also examined how the model handled anomalies and extreme events such as sudden load drops or spikes associated with renewable fluctuations, showing reasonable resilience but without explicitly modeling network congestion. Hence, while informative for system-level demand prediction, the approach remains limited in its ability to infer spatial congestion propagation.

Overall, while there is a growing body of work addressing market forecasting in Italy, significant research gaps remain. Most existing models rely on traditional machine learning rather than modern deep learning or graph-based architectures capable of representing the spatial interdependence among zones. Furthermore, rare events such as congestion spikes and their relationships with market prices remain underexplored. Addressing these challenges is particularly relevant given the Italian system's structure, characterized by high renewable penetration, zonal transmission bottlenecks, and evolving market coupling mechanisms.

Importantly, Italy provides a valuable proxy for broader European congestion forecasting. Its zonal configuration matches the multi-country structure of the European electricity market, where cross-border exchanges play a role analogous to interzonal flows. As part of the Supernova project, the Italian case thus offers a scalable framework for extending congestion forecasting methodologies to the entire European context, with each country treated as a zone. Studying Italy in detail therefore represents not only a step toward improving national market efficiency but also a pathway to developing data-driven tools for congestion prediction in the EU.

3.5. Conclusion

Taken together, the literature surveyed across five methodological families leaves a clearly defined open problem. Physics-based methods [42, 43] achieve high physical fidelity but depend on confidential bus, load, and generator data that are inaccessible to market participants. Econometric and statistical approaches [34, 35] are feasible under limited observability, but model hours and zone pairs in isolation, rely on largely linear structures, and target price proxies rather than physical flows. Deep learning and hybrid architectures [63, 53] improve on temporal expressiveness yet still treat the problem as a collection of independent time series, ignoring the spatial interdependence imposed by the transmission network. Spatio-temporal GNN methods [41, 40] successfully incorporate topology but focus on node-level quantities (generation, load, voltage) and are typically evaluated with full network observability on the TSO side. Finally, the emerging literature on edge-level prediction [84, 83] establishes that existing architectures are not specifically designed for direct regression of signed, directed edge quantities, and leaves the multi-step inductive setting as an open design problem.

No existing work simultaneously satisfies all three requirements that physically grounded interzonal flow forecasting imposes: *(i)* a spatially aware model that respects the topology of the zonal network, *(ii)* edge-level outputs that treat directed flows as first-class regression targets rather than post-hoc combinations of node embeddings, and *(iii)* exclusive reliance on publicly observable market and weather inputs, making the approach actionable for market participants operating without access to proprietary TSO data.

4

Method

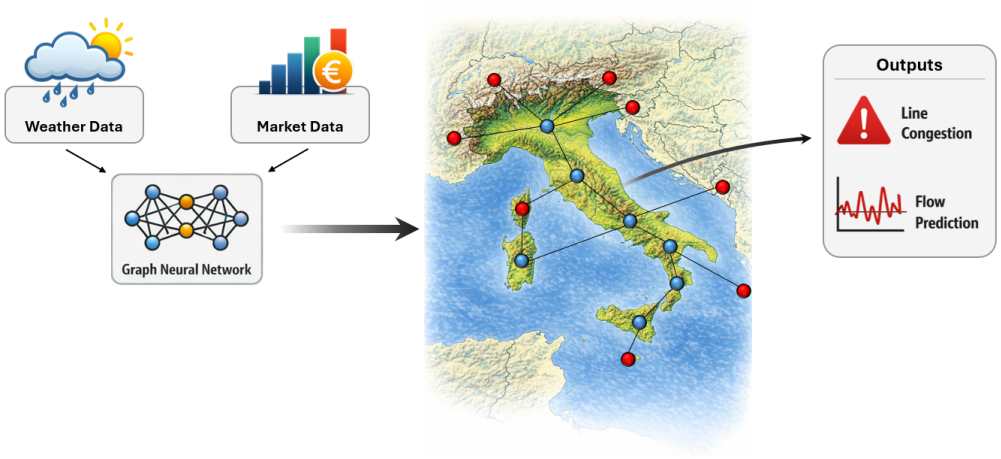


Figure 4.1: Overall problem formulation.
Blue nodes on the graph are the Italian zones and red nodes are neighboring countries. Base map generated with Copilot.

4.1. Problem formulation

The initial goal of this project was to predict interzonal congestion as a price difference. However, after careful consideration and preliminary experiments, the objective was reframed as interzonal flow forecasting. Price time series exhibit high noise levels that hindered model performance, whereas physical flows are governed by more stable underlying dynamics. Once flows are predicted, they can be compared against the day-ahead interzonal capacities published by Terna to yield a congestion ratio signal.

Formally, let $\mathbf{f}_{ij}^t \in \mathbb{R}$ denote the net observed flow on interconnection (i, j) at time t, and let C_{ij}^t be the corresponding day-ahead published capacity. We define the congestion indicator as:

$$r_{ij}^t = \frac{f_{ij}^t}{C_{ij}^t} \tag{4.1}$$

Our objective is to forecast $\hat{f}_{ij}^{t+h}$ for forecast horizon h, using historical observations of flows, weather, and market prices as inputs. The model does not replicate market participants' forecasts, it rather learns from their implicit output (day-ahead and intraday prices) alongside weather observations to detect situations where market forecasts under or overestimate realized congestion.

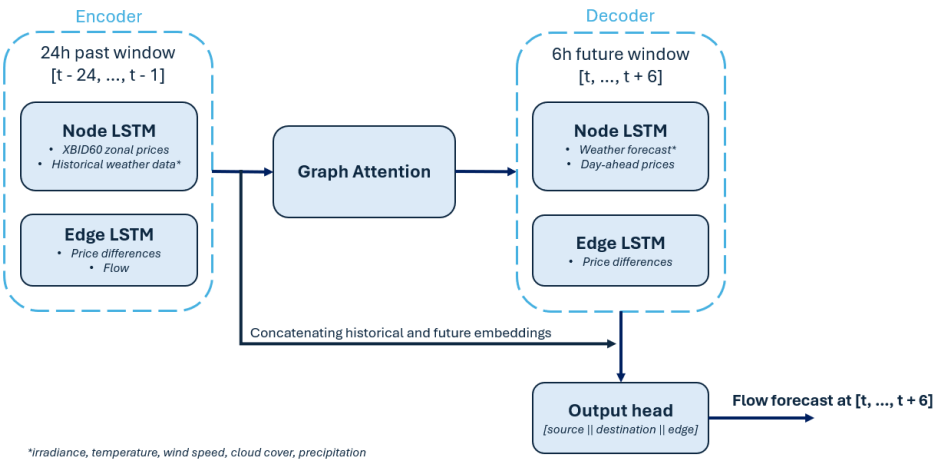


Figure 4.2: Diagram of the model architecture.

4.2. Model Architecture

The proposed model follows an encoder-decoder architecture built on a spatio-temporal graph neural network. An overview is provided in Figure 4.2. The intuition is straightforward: we first summarize each node's and edge's recent history into a compact context vector, then let the graph propagate information across connected zones so that, for example, a wind surge in the north can inform the expected flow on a southern interconnection. Finally, a decoder processes available future information (weather forecasts, day-ahead prices) before a lightweight output head produces the flow forecast.

Concretely, the model is composed of four stages:

1. a dual LSTM encoder that compresses historical node and edge sequences into fixed-size context vectors,
2. a Transformer-based graph message-passing module that propagates spatial context across the network,
3. a dual LSTM decoder, initialized from the GNN-updated representations, that processes the weather and prices forecasts,
4. and an MLP output head that maps the fused representation to a scalar flow forecast.

4.2.1. Encoder

The encoder's role is to summarize each node's and each edge's recent history into a single fixed-size vector that captures the most relevant temporal patterns. An LSTM is a natural choice here because it is well-suited to variable-length sequences with long-range dependencies, such as the 24-hour look-back window used in this work.

Each node and edge sequence is passed independently through its respective LSTM. The final hidden state (the vector the LSTM produces after seeing the entire input sequence) serves as the context vector for that entity:

$$\mathbf{h}_v^{\text{enc}}, \mathbf{c}_v^{\text{enc}} = \text{LSTM}_{\text{node}}(\mathbf{X}_v) \tag{4.2}$$

$$\mathbf{h}_e^{\text{enc}}, \mathbf{c}_e^{\text{enc}} = \text{LSTM}_{\text{edge}}(\mathbf{X}_e) \tag{4.3}$$

where $\mathbf{X}_v \in \mathbb{R}^{T_{\text{enc}} \times F_v}$ and $\mathbf{X}_e \in \mathbb{R}^{T_{\text{enc}} \times F_e}$ are the historical node and edge feature sequences, T_{enc} is the look-back window length (24 hours in this work), F_v and F_e are the number of input features per time step for nodes and edges respectively, and $\mathbf{h}_v^{\text{enc}}, \mathbf{h}_e^{\text{enc}} \in \mathbb{R}^H$ are the resulting context vectors of hidden dimension H.

The graph contains two levels of granularity: Italian bidding zones and neighboring country nodes. Because these two entity types can have different feature sets and operate at different spatial scales,

separate LSTM encoders are used for each. This specialization allows each encoder to focus on the dynamics most relevant to its resolution:

$$\text{LSTM}_{\text{node}}(\cdot) = \begin{cases} \text{LSTM}_{\text{country}} & \text{if country node} \\ \text{LSTM}_{\text{zone}} & \text{otherwise} \end{cases} \tag{4.4}$$

Edge features such as net flow are inherently directional: a positive flow from zone A to zone B is physically a negative flow from B to A. To obtain a fully undirected graph representation while preserving this directionality information, both directions of each edge are encoded. The reverse direction is represented by simply negating the input features, so that the encoder sees a physically consistent antisymmetric signal:

$$\mathbf{h}_e^{\text{fwd}} = \text{LSTM}_{\text{edge}}(\mathbf{X}_e), \quad \mathbf{h}_e^{\text{bwd}} = \text{LSTM}_{\text{edge}}(-\mathbf{X}_e). \tag{4.5}$$

This produces a full undirected edge attribute tensor $\mathbf{A} = [\mathbf{h}^{\text{fwd}} \,\|\, \mathbf{h}^{\text{bwd}}] \in \mathbb{R}^{2E \times H}$ that is passed to the graph message-passing module, with H the hidden dimension and E the number of edges.

4.2.2. Graph Neural Network

After encoding, each node holds a summary of its own recent history, but nodes are not yet aware of what is happening at neighboring zones. The graph message-passing module presented in Eq. 2.2 addresses this by allowing each node to attend to its neighbors and update its representation across L successive layers.

Specifically, L layers of Transformer-based graph convolution are applied, where edge embeddings modulate the attention weights between connected nodes. The mechanism implemented is an extension of the original additive attention described in 2.3.2, but instead using dot-product attention.

The following three paragraphs go into the details of the GNN implementation.

Pre-LayerNorm Residual Updates

A critical design choice concerns how the GNN update is combined with the LSTM-encoded representation. The standard approach applies layer normalization after adding the update to the current state. We denote by $\mathbf{h}^{(l)}$ the node representation at layer l, with $\mathbf{h}^{(0)} = \mathbf{h}^{\text{enc}}$ initialized from the encoder output. Standard post-LayerNorm residual connections take the form:

$$\mathbf{h}^{(l+1)} = \text{LN}\Big(\mathbf{h}^{(l)} + \text{Update}(\mathbf{h}^{(l)})\Big) \tag{4.6}$$

However, because normalization is applied outside the addition, it rescales the entire representation after every layer. During development, we monitored the update significance ratio at each layer, defined as:

$$\rho^{(l)} = \frac{\|\mathbf{h}^{(l+1)} - \mathbf{h}^{(l)}\|_2}{\|\mathbf{h}^{(l)}\|_2} \tag{4.7}$$

In practice, this ratio, which is the ratio of the update magnitude to the current state magnitude, reached values around 500% in early layers under this scheme, meaning the GNN was effectively discarding the LSTM-encoded history entirely and rewriting it from scratch. This is undesirable: the LSTM context is the primary source of temporal information and should not be overwritten.

To preserve the representational fidelity of the encoder, Pre-LayerNorm is adopted instead. Here, the normalization is applied to the input before computing the update, and the result is added back to the original, unmodified state. Formally, the node update at layer l is:

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + g_n \cdot \text{Update}\Big(\text{LN}(\mathbf{h}^{(l)})\Big), \tag{4.8}$$

where LN denotes layer normalization. The raw LSTM state flows unchanged through the residual path, so the GNN receives only a normalized view of $\mathbf{h}^{(l)}$ when computing messages, but its output is added back to the unmodified state.

Learnable Residual Gates
Even with Pre-LayerNorm, it is important to control how strongly the GNN is allowed to modify the encoded representations, particularly in early training when the graph convolution weights are not yet meaningful. To achieve this, two scalar learnable gates are introduced, one for nodes and one for edges, that scale the GNN update before it is added to the residual. Each gate is the sigmoid of a learned scalar parameter, so it is bounded between zero and one:

$$g_n = \sigma(\gamma_n), \quad g_e = \sigma(\gamma_e), \quad \gamma_n, \gamma_e \in \mathbb{R}. \tag{4.9}$$

Both parameters are initialized to small negative values ($\gamma = -0.5$, corresponding to $\sigma \approx 0.38$), ensuring that early in training the GNN updates are small perturbations of the LSTM representations. As training progresses, the gates can open up to allow stronger spatial mixing. The update significance ratio $\rho^{(l)}$ (Eq. 4.7) was monitored throughout training to verify that the gates converged to stable values rather than saturating towards zero or one.

Edge Updates
Standard message-passing layers treat edge attributes as static conditioning variables that influence node updates but are never themselves updated. Since the prediction target lies on edges, it is beneficial to also update edge representations after each node update step, so that edge embeddings can incorporate information about the current state of both endpoints.

Concretely, after each convolution layer, the edge embedding for the directed pair $(i \to j)$ is updated by concatenating the updated source and destination node embeddings with a normalized version of the current edge embedding, and passing the result through a linear layer. The residual connection and gate ensure that the update is a controlled perturbation rather than a complete rewrite:

$$\mathbf{e}_{ij}^{(l+1)} = \mathbf{e}_{ij}^{(l)} + g_e \cdot \mathrm{Linear}\left(\left[\mathbf{h}_i^{(l+1)} \,\middle\|\, \mathbf{h}_j^{(l+1)} \,\middle\|\, \mathrm{LN}\left(\mathbf{e}_{ij}^{(l)}\right)\right]\right), \tag{4.10}$$

where $[\,\cdot \,\|\, \cdot\,]$ denotes concatenation. Because the undirected edge index contains both $u \to v$ and $v \to u$, concatenating source, destination, and edge embeddings preserves directionality: the model can learn that the congestion ratio approaching saturation when flow runs from zone A to zone B may differ from the ratio when flow runs from B to A, since Terna publishes distinct capacity limits for each direction.

4.2.3. Decoder

After message passing, each node and edge holds a GNN-updated context that fuses both its own temporal history and the spatial context from neighboring zones. The decoder's role is to incorporate future exogenous information, weather forecasts and day-ahead prices, into this context before making a prediction.

The decoder LSTMs are initialized with GNN-updated representations, so that each decoder step benefits from the full graph context gathered during message passing. As in the encoder, separate decoder LSTMs are used for country and Italian zone entities.

For edges, the forward and backward GNN-updated embeddings need to be combined into a single initialization vector. Max pooling is used rather than averaging, as it better preserves strong activations from either direction and is more robust to asymmetric congestion signals:

$$\mathbf{e}_{ij}^{\mathrm{init}} = \max\left(\mathbf{e}_{ij}^{(\mathrm{fwd})}, \mathbf{e}_{ij}^{(\mathrm{bwd})}\right) \in \mathbb{R}^H. \tag{4.11}$$

The decoder then unrolls over the forecast horizon T_{dec}, processing the future exogenous features at each step:

$$\mathbf{D}_e = \mathrm{LSTM}_{\mathrm{dec}}\left(\mathbf{X}_e^{\mathrm{fut}}, \mathbf{e}_{ij}^{\mathrm{init}}\right) \in \mathbb{R}^{T_{\mathrm{dec}} \times H}, \tag{4.12}$$

where $\mathbf{X}_e^{\mathrm{fut}}$ contains future exogenous edge features (weather forecasts and interzonal price differences from the day-ahead market). An analogous initialization is used for node decoders using the GNN-updated node embedding $\mathbf{h}_v^{\mathrm{gnn}}$ as the initial hidden state, with a zero cell state.

If no future features are available ($T_{\mathrm{dec}} = 0$), the model bypasses the decoder and the GNN-updated context is passed directly to the output head, collapsing the architecture to a direct multi-step output MLP.

4.2.4. Output Head

The output head produces a scalar flow forecast for each edge and each decoder time step. To do so, it fuses three sources of information: the encoder embeddings for the source node, the destination node, and the edge itself. These are combined with the corresponding decoder outputs, which carry the future exogenous signal. Bringing together both the historical spatial context and the future exogenous signal at the very last stage ensures that neither source is discarded.

Formally, for edge (i, j) at decoder step t, the input representation concatenates the encoder context for both endpoints and the edge, together with the decoder outputs for the same entities:

$$\mathbf{r}_{ij}^{t} = \left[\mathbf{h}_{i}^{\text{enc}} \,\middle\|\, \mathbf{h}_{j}^{\text{enc}} \,\middle\|\, \mathbf{e}_{ij}^{\text{enc}} \,\middle\|\, \mathbf{d}_{i}^{t} \,\middle\|\, \mathbf{d}_{j}^{t} \,\middle\|\, \mathbf{D}_{ij}^{t} \right] \in \mathbb{R}^{6H}. \tag{4.13}$$

This six-component vector is passed through a two-layer MLP to produce the scalar forecast:

$$\hat{f}_{ij}^{t+h} = \text{MLP}(\mathbf{r}_{ij}^{t}). \tag{4.14}$$

4.2.5. Training

The model is trained to minimize the difference between predicted and observed flows. Mean Absolute Error (MAE) is used rather than Mean Squared Error because it provides more robust gradients for the spike-like congestion events that are of most practical interest, without disproportionately penalizing large errors.

Because flows are scaled globally (a single scaling factor shared across all edges), the relative magnitudes between edges are preserved. This is physically meaningful : a high-capacity interconnection naturally produces larger absolute errors than a low-capacity one. But it also means the loss is dominated by the busiest edges. To compensate, each edge is assigned an inverse-frequency weight that down-weights high-flow edges and gives more attention to lower-flow interconnections. Specifically, the weight for edge (i, j) is inversely proportional to the square root of its mean absolute flow $\bar{f}_{ij}$ over the training period:

$$w_{ij} = \frac{1}{\sqrt{\bar{f}_{ij} + \varepsilon}}, \tag{4.15}$$

where ε is a small constant for numerical stability. The weights are normalized to have unit mean across all edges to avoid changing the overall loss scale. The weighted training objective is then:

$$\mathcal{L} = \frac{1}{|\mathcal{E}| T_{\text{dec}}} \sum_{(i,j)\in\mathcal{E}} \sum_{t=1}^{T_{\text{dec}}} w_{ij} \cdot \left| \hat{f}_{ij}^{t} - f_{ij}^{t} \right|. \tag{4.16}$$

with $\mathcal{E}$ being the edge set.

Training uses the Adam optimizer with a cosine annealing learning rate schedule. A best-snapshot early stopping strategy is employed: model weights are saved whenever training loss improves by more than $\delta_{\min}$, and the best snapshot is restored at the end of training. This provides implicit regularization without requiring a dedicated validation set.

4.2.6. Model Scalability

While the current framework is evaluated on the Italian grid, the long-term objective is to scale the approach to a European setting. We discuss scalability along two dimensions: training and inference.

Training scalability. The computational cost of the Transformer graph convolution scales as $O(|\mathcal{E}| \cdot H^2)$ per layer per time step, where $|\mathcal{E}|$ is the number of edges and H is the hidden dimension. The Italian grid contains a moderate number of interzonal interconnections so scaling to Europe would increase $|\mathcal{E}|$ by roughly one order of magnitude. Because each graph snapshot is processed independently (no cross-time graph dependency), the architecture is trivially parallelizable across time windows and naturally supports mini-batch training via PyTorch Geometric's `DataLoader`.

The dual-LSTM design (separate encoders for countries and zone level entities) also scales gracefully: adding new countries requires only that their nodes be assigned to the appropriate granularity level, with no architectural change.

Inference scalability. For real-time intraday use, inference on a single graph snapshot involves one forward pass through the encoder LSTMs, L GNN layers, and the decoder LSTMs. The overall inference latency scales linearly with the number of nodes and edges, and is dominated by the LSTM encoding steps rather than the graph convolution. In practice, this makes the model suitable for deployment at the intraday gate closure horizon (typically 15–60 minutes before delivery), provided that input features (weather forecasts, day-ahead prices) are available at that horizon.

Graph construction for Europe. Extending the graph to Europe requires defining provinces for each country and interzonal edges corresponding to cross-border interconnections as published by ENTSO-E. The hierarchical super-node mechanism already present in the architecture would accommodate multi-country zone aggregation with no modification to the message-passing logic.

4.2.7. Model Limitations

Absence of a validation set. Due to the limited size of the available dataset, no dedicated validation set was held out. Indeed, data spans from November 2021 to December 2025 (the Calabria zone was only introduced in 2021). Around 3 years are used for training and 1 year for testing. We considered that using one year for validation would really hinder the model performance, as keeping only 2 years of training data would be really low. The best-snapshot early stopping strategy provides implicit regularization, but it cannot detect overfitting to the training distribution. Expanding the dataset to cover multiple years would allow proper train/validation/test splits.

5

Experiments

This chapter evaluates the proposed architecture. Section 5.1 describes the shared experimental setup. Experiments 5.2–5.3 validate that the model does what it claims. Experiment 5.5 investigate where the model struggles. Experiment 5.4 describes the interesting effect of adding future features on congestion detection. Additional experiments and analysis are in Appendix A : how performance degrades with horizon in section A.1, which components in the architecture matter in section A.2 and A.3, and discussion on practical applicability in section A.4.

5.1. Experimental Setup

5.1.1. Dataset

The dataset covers the Italian transmission network from November 2021 to December 2025, yielding approximately four years of hourly observations. Data from 2021–2024 are used for training and 2025 for testing, giving a clean temporal split with no look-ahead leakage.

Four data sources are combined:

- **Terna**: physical and scheduled hourly interzonal flows for Italy's seven bidding zones (NORD, CNOR, CSUD, SUD, CALA, SICI, SARD).
- **ENTSO-E Transparency Platform**: hourly day-ahead prices, day-ahead interzonal capacities, and cross-border flows for neighboring countries (France, Switzerland, Austria, Slovenia, Greece, Montenegro, Malta).
- **GME**: hourly prices for the intraday auction market (MI), and the continuous trading session (MI-XBID) in Italy.
- **Era5**: hourly weather reanalysis (temperature, solar irradiance, wind speed, surface pressure, precipitation, cloud cover) at the province level, aggregated to the bidding-zone level (see Section 5.1.2).

Missing data. Hours with missing interzonal flow values were dropped. Montenegro and Malta are missing several years of price data, therefore results on the corresponding border connections should be interpreted with caution, as the model received a price signal for only a fraction of the training period.

Known data quality issue. A non-trivial discrepancy was discovered in the 2025 Terna dataset: flows are reported at 15-minute resolution and labeled as power (MW), but the values are in fact energy (MWh per quarter-hour). Treating them as power values therefore yields flows that are a factor of four too low. The issue was identified by comparing the magnitude of the 2025 test set against the 2021-2024 training set. The correction applied was to sum the four quarter-hourly energy values within each hour to obtain the hourly energy in MWh, then divide by one hour to recover the mean power in MW, rather than averaging as one would do for true power readings. This correction was critical for producing consistent training and test distributions.

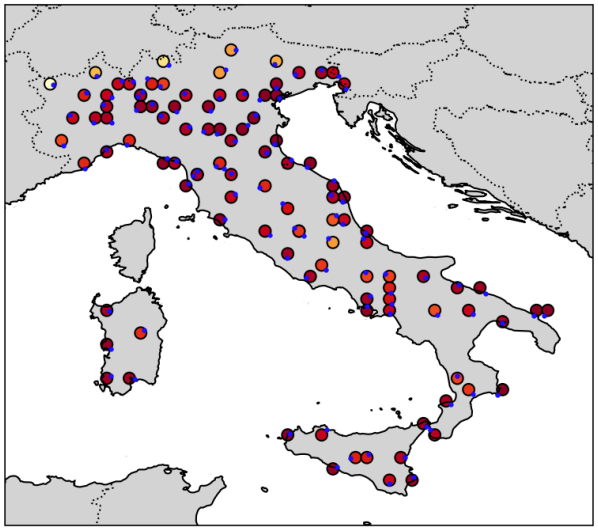


Figure 5.1: Italian provinces (in reddish, proportionate to the irradiance around noon) used to aggregate the weather data per zone, and the corresponding grid coordinates from Era5 (in blue).

Data availability. The Calabria (CALA) bidding zone was only introduced as a separate zone in 2021, meaning it has limited historical coverage relative to the other zones. Experiments are therefore based on a training period starting in 2021, even though data for previous years is available for other zones.

Scaling. All features are scaled using `RobustScaler`, which removes the median and divides by the interquartile range rather than the mean and standard deviation used by `StandardScaler`. This choice is motivated by the presence of outliers in both flow and price series (for instance, price spikes during the 2022 energy crisis), which would disproportionately inflate the standard deviation and compress the bulk of the distribution into a narrow range.

Flows are scaled globally across all edges using a single scale factor, rather than independently per edge. This is a deliberate design choice: preserving relative magnitudes between edges allows the model to learn that a 50 MW error on the NORD–France connection is physically very different from a 50 MW error on the Sicily–Malta cable. For context, flows between NORD and France routinely reach several thousand MW, whereas the Sicily–Malta interconnector operates in the range of tens of MW. Scaling per edge would map both to the unit interval and erase this physically meaningful contrast.

5.1.2. Weather Data

Weather observations were collected from ERA5 [91] as GRIB files covering the full study period. ERA5 provides a regular spatial grid of reanalysis data. From this grid, the points closest to each Italian province were extracted (see Fig 5.1). The provinces were chosen as spatial representatives because Terna regularly publishes electricity capacity and production statistics at the province level [92], making them a natural unit for associating weather conditions with local generation. A similar extraction was applied for the neighboring countries, using grid points near major cities as proxies.

The weather variables selected are those most commonly used in the energy forecasting literature [78]: temperature, wind speed, surface pressure, solar irradiance, cloud cover, and precipitation. Some ERA5 variables are provided as accumulated quantities (like solar irradiance in $J.m^{-2}$), and these were converted to instantaneous power fluxes ($W.m^{-2}$) by differencing consecutive accumulation values and dividing by the timestep.

Province-level observations were then averaged across all provinces belonging to a given Italian bidding zone to obtain a single value per zone, hour, and weather feature. Other spatial aggregation schemes were considered (like capacity-weighted averages), but as noted in [93], no single transformation consistently outperforms the others across forecasting tasks. Given this lack of consensus, simple averaging was adopted as the simplest baseline.

Important caveat. The model uses ERA5 *reanalysis* (past observations) rather than numerical weather predictions (NWP). In a live deployment, weather forecasts for the delivery period would need to be fetched from an NWP API. The results reported here therefore represent an upper bound on what is achievable with real forecast inputs, since reanalysis values are free of the forecast errors that a deployed system would face.

5.1.3. Model Configuration

The model was optimized running Optuna for 50 trials using the hyperparameters listed in Table 5.1.

Hyperparameter	Value
Encoder look-back window	24 h
Forecast horizon	6 h
Hidden dimension	[32, 512]
Number of GNN layers	[1, 6]
Attention heads	{2, 4, 8, 10}
Dropout	[0.0, 0.5]
Learning rate	$[10^{-5}, 10^{-3}]$ (log scale)
Batch size	{64, 128, 256}
Epochs	150
LR schedule	Cosine annealing

Table 5.1: Default model hyperparameters ranges used in all experiments unless stated otherwise.

5.1.4. Evaluation Metrics

Three complementary metrics are reported throughout this chapter.

Mean Absolute Error (MAE). The average absolute deviation between predicted and observed flows. *Normalized* MAE (NMAE) is introduced, expressed as a ratio of the edge MAE and the edge average flow over the training set, so that errors are observable compared to the edge usual flow. NMAE is the primary optimization target and the most directly interpretable metric for market participant.

Directional Accuracy (DA). The fraction of time steps where the predicted sign of the flow matches the observed sign: $\mathrm{DA} = 100 \cdot \mathbb{E}[\mathbf{1}(\mathrm{sign}(\hat{f}) = \mathrm{sign}(f))]$. A wrong sign prediction is a qualitatively different error for congestion monitoring: it implies that the model believes flow is running in the opposite direction, which would invert the congestion ratio.

Congestion F1 (CF1). Binary F1 score for the task of detecting congestion events, defined as $r_{ij}^t = \frac{f_{ij}^t}{C_{ij}^t} > 0.9$. This metric directly measures performance on the downstream task of interest and is reported alongside precision and recall to expose any precision–recall trade-off.

5.1.5. Baseline models

Several baselines of diverse sophistication are evaluated, ranging from naive statistical predictors to ML models and a domain-specific model. Together they provide a solid benchmark against which each marginal complexity increase in the proposed model can be assessed.

Short persistence. The forecast at horizon h is the value observed exactly 24 hours earlier: $\hat{f}^{t+h} = f^{t+h-24}$. This exploits the strong diurnal periodicity of electricity flows and is the cheapest possible model which sets a lower bound on useful performance.

Long persistence. For each edge, hour-of-day, day-of-week, and month-of-year combination, the forecast is the mean computed over the training set. This baseline captures the average seasonal and weekly pattern but has no ability to respond to day-to-day variability driven by weather or market conditions.

Scheduled flow (Terna). Every day, Terna publishes the quarter-hourly interzonal trade program resulting from the implicit and explicit capacity allocation process, covering all time horizons (yearly, monthly, day-ahead, intraday) and incorporating Trans-European Replacement Reserves Exchange (TERRE) values. This schedule[1] represents the market's best collective forecast of interzonal flows at the day-ahead stage and is a strong domain-specific baseline: any learned model that cannot beat the schedule is unlikely to add value in practice.

ARIMA. AutoRegressive Integrated Moving Average model fitted per edge. ARIMA captures linear temporal dependencies without requiring feature engineering, serving as a classical multivariate time-series reference.

Linear regression. A simple linear model trained per edge, using the same lagged flow, weather, and price features as the proposed model. This tests whether the temporal and spatial patterns in the data are well described by a simple linear relationship, before resorting to non-linear models.

XGBoost*. A gradient-boosted tree ensemble trained per edge on the same lagged flow, weather, and price features as the proposed model. XGBoost is a strong non-linear baseline that routinely outperforms deep learning on tabular data : it serves as the primary benchmark for the non-graph deep learning components.

Electricity Maps. Electricity Maps[2] is a company providing open carbon intensity and power flow estimates derived from a fundamental flow-tracing model. Although interzonal flow prediction is not the primary objective of Electricity Maps, its outputs are publicly available and represent high-quality flow estimation for a worldwide production environment. Including it contextualizes the proposed model against a real-world deployed system, although one with a different design goal.

LSTM (flat)*. A standard LSTM regressor trained independently per edge, using the same input features (historical flows, weather, prices) as the proposed model but with no graph structure. Any gap between this baseline and the proposed model is attributable to the GNN's ability to propagate spatial context across the network.

GNN*. A GNN without the encoder-decoder structure presented in Figure 4.2 is trained in order to assess the importance of the LSTMs temporal processing.

*Optimized using Optuna for 50 trials on a NVIDIA GeForce RTX 4090.

5.2. Experiment 1: Does the model produce meaningful flow forecasts?

Can the proposed architecture outperform simple baselines on the interzonal flow forecasting task?

Before investigating architectural choices, a basic sanity check is needed: can the model beat the simplest possible predictors? A model that fails to outperform persistence or Terna's scheduled flows provides no practical value, regardless of its complexity.

5.2.1. Results

Table 5.2 reports the three primary metrics averaged across all 16 directional edges and 6 forecast horizons on the 2025 test set. Detailed standard deviations are reported in Appendix B.1, and discussed here only when relevant.

5.2.2. Analysis

The proposed model achieves the lowest NMAE (52.1%) and the highest DA (89.8%) and CF1 (60.2%) among all models, confirming that the architecture produces meaningful forecasts that go beyond what

[1] https://dati.terna.it/en/transmission#scheduled-internal-exchange
[2] https://app.electricitymaps.com/map/live/fifteen_minutes?signal=electricity-price

Model	NMAE ↓	DA ↑	CF1 ↑
Short persistence	69.8%	86.1%	56.9%
Long persistence	135.0%	76.5%	16.2%
Scheduled flow*	204.6%	86.4 %	37.6 %
ARIMA	<u>61.2%</u>	88.3%	<u>59.5%</u>
LSTM	77.5%	88.3%	51.1%
XGBoost	81.4%	<u>89.1%</u>	51.4%
Linear regression	77.4%	87.6%	46.9%
GNN	76.9%	83.8%	40.0%
Proposed model	**52.1%**	**89.8%**	**60.2%**

Table 5.2: Comparison with baselines on the test set (2025). NMAE is expressed per edge as a percentage of the edge training mean flow, then averaged across all edges. DA and CF1 are expressed as percentages. **Bold**: best, <u>underline</u>: second-best. *Scheduled flow NMAE is high because the day-ahead schedule is not designed to minimize MAE but to ensure feasibility so it is included as an operational reference.

simple heuristics can capture. All models exhibit substantial variability in performance, with standard deviations often comparable to the mean. Importantly, this variability is consistent across baselines and does not alter the relative ranking of models: the proposed architecture remains dominant across metrics. It's variability is comparable to that of the strongest baselines and does not indicate instability. The superior average performance therefore reflects consistent gains rather than improvements driven by a small subset of favorable cases. Detailed mean ± standard deviation results are reported in Appendix B.1 for completeness. The proposed model was trained using 5 different seeds and the variance for these runs on the NMAE turned out to be insignificant, around 0.4%.

Among the baselines, short persistence is the strongest single competitor on NMAE (69.8%) and CF1 (59.5%, tied with ARIMA), which is expected: electricity flows exhibit strong diurnal regularity, so copying the value from 24 hours prior is already a non-trivial predictor. Long persistence is substantially worse across all metrics, indicating that the average seasonal pattern alone, without any response to day-to-day variability, is insufficient.

The scheduled flow baseline deserves separate interpretation. Its NMAE (295.8%) is high because the day-ahead schedule is not optimized to minimize absolute flow error, it is a feasibility-constrained allocation that systematically approaches physical limits. Despite this, it achieves the second-highest DA (89.1%), confirming that the schedule correctly captures the direction of flow most of the time. Its low CF1 (12.0%) reflects very low precision: the schedule is deliberately conservative and flags congestion far more often than it actually occurs (see Experiment 5.3 for a detailed discussion).

Among the supervised baselines, ARIMA achieves the second-lowest NMAE (61.2%) and second-highest CF1 (59.5%), performing surprisingly well given that it uses only the target series with no exogenous features. This suggests that the flow time series itself carries a strong autoregressive structure that simpler models like LSTM and XGBoost, which receive many additional features, fail to exploit as efficiently on this dataset. The LSTM (77.5%) and bare GNN (76.9%) both underperform short persistence on NMAE (69.8%), but for opposite reasons: the LSTM processes each edge independently with no spatial context, while the bare GNN has graph structure but no temporal encoding. Only the proposed model combines both, and the gap confirms that neither component alone is sufficient.

The proposed model improves upon the best supervised baseline (ARIMA) by approximately 9 pp in NMAE while also leading on DA and CF1, suggesting that both the richer feature set and the graph-based spatial propagation contribute beyond pure autoregressive modeling. Per-horizon and per-edge breakdowns, as well as ablation of individual components, are provided in Experiments A.1 and 5.5.

5.3. Experiment 2: Does the predicted congestion ratio align with observed congestion events?

Does a low flow-forecasting error translate into correct identification of congestion events?

Even with a low MAE, a model could systematically under-predict peak flows and thereby miss congestion events. This experiment evaluates the downstream congestion detection task directly, treating the predicted congestion ratio $\hat{r}_{ij}^{t} = \hat{f}_{ij}^{t}/C_{ij}^{\mathrm{t}}$ as a soft score and applying a threshold of 0.9.

5.3.1. Results

Table 5.3 reports Precision, Recall, CF1, and AUROC averaged across all 16 directional edges and 6 forecast horizons. Bold marks the best value per column and underline marks the second best.

Model	Precision ↑	Recall ↑	CF1 ↑	AUROC ↑
Short persistence	56.3	57.6	56.9	79.4
Long persistence	42.5	15.4	16.2	65.4
Scheduled flow	30.3	**85.4**	37.6	75.7
ARIMA	57.7	<u>62.5</u>	<u>59.5</u>	81.3
LSTM	71.1	44.0	51.1	86.9
XGBoost	**74.1**	43.9	51.4	<u>88.2</u>
Linear regression	59.8	41.3	46.9	82.0
GNN	58.1	36.0	40.0	76.4
Proposed model	<u>73.5</u>	53.7	**60.2**	**89.8**

Table 5.3: Congestion detection performance at threshold $r = 0.9$. All scores are expressed as percentages. Bold: best per column, underline: second best.

Threshold-fixed metrics (Precision, Recall, CF1). The proposed model achieves the highest CF1 (60.2%), combining competitive precision (73.5%, second only to XGBoost at 74.1%) with the best recall among the high-precision models (53.7%). ARIMA is the strongest baseline on CF1 (59.5%), benefiting from a more balanced precision–recall trade-off (57.7 / 62.5), while LSTM and XGBoost favor precision at the expense of recall, resulting in lower CF1 despite higher precision.

The scheduled-flow baseline is a special case: it achieves by far the highest recall (87.3%) but collapses to very low precision (6.4%) and CF1 (12.0%). This is consistent with the operational logic of the Terna day-ahead schedule, which is deliberately conservative : capacity is pre-reserved to anticipate potential congestion, so the scheduled flow systematically approaches or exceeds the threshold even when no congestion materializes. From a grid-operator perspective this behavior reflects a deliberate design choice: a false alarm (spurious congestion flag) is far less costly than a missed event that would require real-time re-dispatch.

Long persistence performs poorly across all metrics (CF1 16.2%, recall 15.4%), while short persistence shows surprisingly high CF1 (56.9%), confirming that copying the flow from 24 hours prior is a simple yet effective way to capture the sharp peaks that characterize congestion events.

Threshold-independent discrimination (AUROC). AUROC measures the model's ability to rank truly congested hours above non-congested ones, regardless of where the previous 0.9 threshold is placed. The proposed model achieves the best AUROC (89.8%), followed by XGBoost (88.2%) and LSTM (86.9%). The gap between AUROC and CF1 rankings is informative: ARIMA ranks second on CF1 but only fifth on AUROC (81.3%), suggesting that its strong threshold-fixed performance stems from well-calibrated predictions near $r = 0.9$ rather than globally superior discrimination. Conversely, XGBoost ranks second on AUROC but only fifth on CF1, indicating that while it separates the two classes well in ranking terms, its hard predictions at the 0.9 threshold are skewed.

5.3.2. Analysis

Precision–recall trade-off. For a transmission system operator, the asymmetry between error types is critical: a missed congestion event requires costly real-time re-dispatch or, in the worst case, a network constraint violation, whereas a false alarm leads only to unnecessary precautionary actions. This asymmetry motivates preferring higher recall over higher precision. Among the models with meaningful precision (> 50%), the proposed model achieves the best recall (53.7%), making it the

most operationally suitable choice at the default threshold of 0.9. If a softer threshold were adopted ($r = 0.85$), the AUROC advantage of the proposed model suggests its recall would increase further while maintaining competitive precision.

Flow regression vs. direct classification. A natural question is whether the flow-regression approach is fundamentally limited for congestion detection, and whether training a direct binary classifier (one that outputs a congestion probability rather than a flow value) would perform better. The argument in favor of a classifier is intuitive: if the ultimate goal is a binary label (congested / not congested), why not optimize for that label directly using a cross-entropy loss, rather than hoping that a regression loss on raw flows incidentally produces good threshold behavior?

There are two reasons why this argument does not hold here.

First, the regression output already provides a well-ordered soft score, and thus captures essentially all the separability available in the inputs. The proposed model's AUROC of 89.8% means that, in 89.8% of all congested/non-congested pairs of hours, the model assigns a higher predicted congestion ratio $\hat{r}_{ij}^t$ to the truly congested hour. This is precisely the objective of any classifier at the ranking level. A direct binary classifier could only improve upon this if it were able to *re-rank* these pairs, that is, learn a representation that separates the two classes better than the predicted flow does. Given that the flow predictor is itself a complex non-linear function of the same inputs, such gains are expected to be marginal.

Importantly, converting this soft score into a binary decision necessarily introduces a threshold, and thus does not eliminate uncertainty. For instance, if the congestion threshold is set at $r = 0.9$, then a congested hour with $r = 0.92$ and a non-congested hour with $r = 0.88$ are physically near-identical: small, unpredictable fluctuations in renewable generation or cross-border scheduling determine which side of the boundary is crossed. A classifier, lacking access to these fluctuations at gate closure, would assign similar predicted probabilities to both cases. The uncertainty is therefore intrinsic to the problem and cannot be removed by changing the modeling approach: it merely appears either as imperfect ranking (reflected in the AUROC gap for the regression model) or as ambiguity around the classification threshold.

Second, and perhaps most importantly for operational use, the flow value itself is more transparent and actionable than a binary flag. A grid operator who receives a predicted flow of 1,850 MW on a line with a 2,000 MW capacity can immediately gauge how close the system is to the limit, assess the margin for error, and decide whether precautionary re-dispatch is warranted. A classifier that outputs "congestion probability: 0.73" conveys less physical intuition and requires the operator to trust a black-box threshold internally. The regression framing thus aligns better with the mental model of market participants and provides a richer signal for decision-making.

In summary, the flow-regression framing is not a limitation: it provides a physically interpretable output (the predicted flow itself, not just a binary flag), it naturally produces a well-calibrated soft score for congestion ranking, and it does not sacrifice discrimination performance relative to what a direct classifier could achieve on the same inputs.

Consistency with Experiment 1. The ranking of models on CF1 is broadly consistent with their NMAE ranking from Experiment 1, confirming that lower flow error does translate into better congestion detection. However, the correspondence is not perfect: ARIMA achieves a CF1 close to the proposed model despite a substantially higher NMAE, while long persistence has both the worst NMAE and the worst CF1. This suggests that congestion detection depends not only on overall error magnitude but also on how errors are distributed near the capacity boundary.

5.4. Experiment 3: Does future information improve performance?

How much does access to day-ahead prices and weather forecasts improve the forecast?

At intraday gate closure, a market operator has access to day-ahead prices and numerical weather forecasts for the delivery period. The decoder is designed to ingest these future exogenous features.

This experiment quantifies the information gain from each source by ablating them individually, which also helps identify which features are most worth investing in (like higher-resolution weather data).

5.4.1. Results

Table 5.4 reports the three primary metrics averaged across all 16 directional edges and 6 forecast horizons.

Variant	NMAE ↓	DA ↑	CF1 ↑
No future features	51.36	89.53	56.44
Weather forecasts only	51.74	89.42	56.87
Price forecasts only	52.72	89.50	59.01
Weather + prices forecasts	52.08	89.83	60.16

Table 5.4: Mean metrics across 16 edges and 6 horizons for each ablation variant. NMAE is normalized by the per-edge mean absolute training flow, DA is directional accuracy, CF1 is the F1-score for congestion events (|flow|/capacity > 0.9).

Flow magnitude and direction. NMAE remains within a narrow band of 51–53% regardless of which future features are provided, and directional accuracy (DA) hovers around 89.4–89.8% across all variants. Paired Wilcoxon signed-rank tests against the *No future features* baseline (96 edge×horizon pairs) confirm that neither difference is statistically significant ($p > 0.05$ in all cases). This suggests that the encoder, which sees the full history of flows, prices, and weather up to gate closure, already captures most of the information needed to predict flow magnitude and direction, the additional look-ahead provided by the decoder features yields no measurable gain on these metrics.

Congestion detection. The picture changes substantially for CF1. All three variants that include at least one future feature outperform the no-future baseline (56.44), with day-ahead prices alone (+2.57 pp) already surpassing weather forecasts alone (+0.43 pp), and the combined variant achieving the highest score (60.16, +3.72 pp). Wilcoxon tests on the 90 edge×horizon pairs that contain at least one congestion event confirm that every improvement is highly significant ($p < 0.001$, $n = 90$). The asymmetry between prices and weather is noteworthy: day-ahead prices are set the evening before delivery and implicitly encode the market's expectation for the following day, effectively including a large part of the weather signal for congestion purposes. Weather forecasts, by contrast, provide only a marginal independent contribution (+0.43 pp alone, but +1.15 pp on top of prices), suggesting their value might be concentrated on a subset of connections where meteorological conditions directly drive congestion.

5.4.2. Analysis

Why do future features help congestion but not flow magnitude? Congestion is a threshold phenomenon: a small improvement in predicting the timing and direction of near-limit flows can disproportionately lift the F1-score, even if the overall MAE barely moves. The encoder already produces accurate mean-flow estimates, the decoder features appear to sharpen the model's confidence near the capacity boundary, reducing both false negatives (missed congestion) and false positives (spurious alarms).

Dominant signal: prices over weather. Day-ahead prices contributed more than weather forecasts in every comparison (+2.57 pp vs. +0.43 pp for CF1). This is consistent with the Italian power system, where cross-border scheduling, which is price-driven, are primary congestion drivers on the major north–south and island edges. Weather forecasts may matter most on wind-dominated edges (like Calabria–Sicilia or Southern-Italy–GR) where stochastic renewable injection directly loads the line, a finer-grained ablation at the per-edge level would be needed to confirm this.

Practical implication. Since day-ahead prices are always available at intraday gate closure and their inclusion is both cheap (a single scalar per node per hour) and statistically significant, they should be retained in any production deployment of the model. Higher-resolution numerical weather forecasts may be worth investing in for wind-heavy edges.

5.5. Experiment 4: Which edges are hardest to forecast?

Where does the model struggle most, and can the pattern be explained by physical or market characteristics?

Not all interzonal connections are equally predictable. Some are dominated by renewable variability (Sicily with high solar and wind penetration), while others are more load-driven and stable. Understanding per-edge performance identifies where the model is least reliable and may motivate future feature engineering or targeted data collection.

5.5.1. Results

Edge	Short pers.	Long pers.	Sched.	ARIMA	LSTM	XGBoost	Lin. reg.	GNN	Proposed
*CALA → SICI	70.3	94.3	432.0	65.0	49.2	46.7	45.1	76.8	**42.8**
*CALA → SUD	61.9	84.0	221.4	52.0	43.9	40.5	39.1	58.7	**38.4**
*CSUD → ME	40.1	74.6	41.0	59.8	42.8	**33.2**	34.6	58.9	37.3
*CSUD → SARD	58.0	68.7	283.1	60.8	44.6	39.5	40.0	70.1	**39.1**
*CSUD → SUD	36.1	45.7	175.8	29.4	22.4	21.3	**20.3**	36.1	20.7
*CNOR → CSUD	70.4	88.2	286.9	71.3	47.9	44.6	43.9	81.4	**43.8**
*CNOR → NORD	68.2	90.4	375.2	79.3	50.8	46.2	45.1	86.6	**44.8**
CNOR → SACODC	135.5	267.1	–	**62.5**	139.1	152.9	260.2	115.4	83.9
*NORD → AT	25.2	54.3	158.5	**14.5**	15.6	**14.5**	14.9	21.7	16.9
*NORD → CH	36.7	44.7	46.9	29.0	25.3	23.7	**23.6**	31.2	23.9
*NORD → FR	29.1	63.6	100.2	23.7	22.7	22.7	**19.4**	27.6	21.1
*NORD → SI	35.3	36.7	123.9	36.5	27.8	26.1	26.7	35.7	**25.6**
SACOAC → SARD	103.3	280.3	–	136.1	211.5	226.3	**91.2**	151.1	103.3
SACODC → SARD	160.0	374.0	–	**72.2**	192.6	229.4	369.8	138.4	95.0
*SICI → MT	154.7	410.6	385.7	150.2	258.0	304.3	**126.7**	200.8	149.3
*SUD → GR	31.7	82.2	29.1	36.7	45.5	30.6	37.7	40.2	**20.6**

Table 5.5: Per-edge NMAE (%, ↓) averaged over horizons h=1–6. * edges for which Scheduled flow is also available. **Bold** indicates best, underline second-best.

Edge	Short pers.	Long pers.	Sched.	ARIMA	LSTM	XGBoost	Lin. reg.	GNN	Proposed
*CALA → SICI	83.0	77.9	**91.3**	86.4	89.1	90.2	90.2	85.3	91.1
*CALA → SUD	73.8	62.3	**86.4**	78.8	81.3	82.9	83.3	75.0	83.7
*CSUD → ME	81.9	64.4	89.3	71.9	78.7	**84.3**	83.2	72.1	80.8
*CSUD → SARD	79.4	70.3	83.1	81.3	83.9	86.4	86.2	72.9	**87.4**
*CSUD → SUD	88.4	89.6	93.4	90.8	92.3	92.4	**93.7**	89.7	93.4
*CNOR → CSUD	76.1	64.3	**89.1**	76.5	84.1	85.6	85.6	68.0	85.8
*CNOR → NORD	82.8	74.1	**91.1**	80.6	87.1	89.0	89.0	77.3	88.9
CNOR → SACODC	80.2	52.1	–	**93.4**	78.6	79.3	64.3	84.5	88.6
*NORD → AT	97.1	98.2	68.1	98.2	98.5	**98.7**	98.6	98.5	98.5
*NORD → CH	90.9	92.4	93.9	92.6	94.4	**94.6**	94.5	92.3	**94.6**
*NORD → FR	96.6	97.9	99.1	98.3	99.5	**99.6**	99.4	99.2	99.4
*NORD → SI	90.1	92.5	59.1	91.1	93.3	93.6	93.1	93.8	**94.2**
SACOAC → SARD	94.6	93.2	–	95.3	**97.4**	97.1	96.9	97.5	96.8
SACODC → SARD	81.7	67.7	–	**95.0**	85.9	85.3	72.1	80.2	88.0
*SICI → MT	95.1	96.9	84.5	**97.5**	97.2	97.2	97.2	97.0	97.2
*SUD → GR	85.2	30.3	**94.9**	85.8	71.6	70.2	74.5	57.2	68.9

Table 5.6: Per-edge DA (%, ↑) averaged over horizons h=1–6. * edges for which Scheduled flow is also available. **Bold** indicates best, underline second-best.

5.5.2. Analysis

Three qualitatively different groups of edges emerge from the results, distinguished by their dominant flow mechanism.

Group 1: Well-behaved internal edges (NORD→AT/CH/FR/SI, CSUD→SUD). The northern cross-border connections and the CSUD→SUD edge are the easiest to forecast across all models, with NMAE

Edge (Congestion freq.)	Short pers.	Long pers.	Sched.	ARIMA	LSTM	XGBoost	Lin. reg.	GNN	Proposed
*CALA → SICI (4.5)	32.6	10.1	11.0	32.9	35.9	36.4	39.0	7.1	**51.2**
*CALA → SUD (3.1)	29.5	17.5	9.6	**35.8**	22.2	24.7	31.1	17.7	26.9
*CSUD → ME (23.3)	**55.0**	23.3	73.8	47.6	33.4	38.4	42.9	8.6	45.4
*CSUD → SARD (6.7)	29.2	8.1	18.4	33.7	26.8	28.2	**37.5**	3.2	34.4
*CSUD → SUD (2.9)	38.9	27.0	8.1	43.8	48.1	52.1	**54.6**	11.4	38.9
*CNOR → CSUD (4.5)	32.2	6.5	11.9	32.1	24.1	26.2	**38.1**	9.2	26.3
*CNOR → NORD (5.2)	30.4	2.9	12.8	28.2	19.3	19.6	32.1	1.0	**35.1**
CNOR → SACODC (62.2)	89.4	0.1	–	**94.7**	90.7	88.3	0.0	88.4	94.0
*NORD → AT (12.1)	63.9	25.0	21.8	**78.4**	73.5	75.2	74.6	64.8	71.5
*NORD → CH (25.7)	50.5	24.3	56.3	58.7	54.6	55.2	57.0	42.7	**65.2**
*NORD → FR (49.2)	73.3	12.7	69.0	78.7	71.6	70.6	**78.9**	68.8	74.9
*NORD → SI (70.2)	80.8	80.0	68.8	80.3	85.6	86.6	86.0	80.5	**87.5**
SACOAC → SARD (64.3)	86.2	3.6	–	83.1	66.9	43.8	**88.8**	82.1	86.3
SACODC → SARD (86.1)	90.1	1.0	–	**94.5**	89.0	90.3	0.0	87.0	93.2
*SICI → MT (-)	–	–	–	–	–	–	–	–	–
*SUD → GR (66.2)	71.9	0.8	**89.9**	70.4	25.0	35.9	43.2	28.3	71.4

Table 5.7: Per-edge CF1 (%, ↑) averaged over horizons h=1–6. * edges for which Scheduled flow is also available. *Congestion freq.* is the fraction of test hours with $r > 0.9$, included as context for interpreting CF1. SICI→MT is omitted (no capacity data available). **Bold** indicates best, underline indicates second-best.

values of 14–37%. These edges are driven primarily by relatively predictable market dynamics: Alpine hydro dispatch responding to day-ahead prices, and load-driven north-to-south transfers that follow a stable diurnal pattern. The proposed model is competitive on these edges but does not dominate: linear regression wins on NORD→FR and NORD→CH, and ARIMA wins on NORD→AT, suggesting that the flow dynamics on these edges are well-captured by linear autoregressive or feature-weighted models without requiring spatial context propagation. DA values above 94% on the NORD cross-border edges indicate that the direction of flow is almost always correctly predicted by all models, so the residual difficulty is purely in magnitude.

Group 2: Renewable-driven internal edges (CALA→SICI, CALA→SUD, CNOR→CSUD, CNOR→NORD). These edges connect zones with high wind and solar penetration (Calabria, Sicily) to the rest of the Italian grid. NMAE values of 43–71% are substantially higher than Group 1, and persistence is surprisingly competitive here, suggesting high temporal autocorrelation driven by persistent weather regimes (for example, several consecutive days of wind events). The proposed model achieves the best NMAE on all four of these edges, with margins of 2–7 pp over the nearest competitor, consistent with the hypothesis that spatial context from neighboring zones (knowing that wind in SUD is high helps predict the CALA→SUD flow) is beneficial precisely where renewable variability couples adjacent zones together.

The CF1 results on this group are mixed. The proposed model achieves its largest CF1 advantage on CALA→SICI (+12 pp over linear regression), which is the edge most affected by Sicilian wind and solar variability and where congestion events are likely clustered around specific weather-driven episodes. However, on CALA→SUD and CNOR→CSUD, linear regression outperforms all other models on CF1 despite having higher NMAE, indicating that precise magnitude estimation and congestion threshold detection are partially decoupled objectives on these edges.

Group 3: Structurally difficult edges (SACODC→SARD, SACOAC→SARD, SICI→MT, CNOR→SACODC). These four connections share a common characteristic: they are submarine cables with no physical redundancy, connecting island systems (Sardinia, Malta, Corsica). NMAE values of 83–160% exceed persistence for most models, indicating that the flow dynamics are fundamentally different from continental interconnectors.

The SACODC and SACOAC connections to Sardinia are the clearest examples of scheduling-dominated flows. ARIMA achieves NMAE of 72.2% on SACODC→SARD (best among all models), while LSTM (192.6%) and XGBoost (229.4%) collapse entirely, and linear regression degrades to 369.8%. This pattern strongly suggests that the flow on this cable is determined by an administrative or scheduling rule that ARIMA can partially recover through its autoregressive structure, but which the feature-rich models

cannot identify because the relevant signal (the TSO's dispatch decision) is not present in any of the available input features. The proposed model (95.0%) partially recovers the ARIMA result by leveraging graph context from adjacent edges, but cannot close the gap fully. Notably, linear regression assigns near-zero weight to all features on CNOR→SACODC (CF1 = 0.0), suggesting that it predicts a constant and the relevant signal is entirely absent from the feature set.

The SICI→MT cable presents a separate challenge: missing price data for Malta across most of the training period means the model receives no market signal for this edge, and the flow volumes are low and irregular. NMAE of 126.7–304.3% across all models reflects this data scarcity rather than a fundamental unpredictability of the physical flows. CF1 cannot be computed because the capacity of the line is not published on ENTSO-E.

SUD→GR is an outlier in a positive sense: the proposed model achieves NMAE of 20.6% (best by a large margin), while LSTM (45.5%) and XGBoost (30.6%) struggle. The GR connection is influenced by conditions across southern Italy, Montenegro, and the broader Balkan grid, and the graph-based spatial aggregation appears to successfully capture these multi-hop dependencies. On CF1, both short persistence (71.9%) and ARIMA (70.4%) match or exceed the proposed model (71.4%), which is explained by the high baseline predictability of the congestion state on this edge: the edge is frequently at or near capacity, so a model that simply predicts persistent congestion will achieve high recall at the cost of precision.

Summary. The proposed model provides the most consistent advantage on renewable-driven internal lines and edges with multi-hop spatial dependencies (CALA group, CNOR group, SUD→GR). It offers little advantage over simpler baselines on well-behaved load-driven edges (NORD cross-border), and no model provides reliable forecasts on submarine cable connections to island systems where the flow is probably administratively scheduled and the relevant dispatch signal is absent from the public data.

5.6. Case study : flow reversal propagation

A key claim of this work is that the graph structure enables the model to anticipate how flow changes propagate across the Italian grid. To illustrate this, we examine two episodes from the 2025 test set where a full reversal of the main Italian corridor (NORD → CNOR → CSUD → SUD → CALA) occurs and assess whether the model captures the timing and direction of each transition.
The reader is encouraged to explore these episodes interactively via the demo visualizer (link: `https://congestion-forecasting.portal.csem.ch/`), where the sequence of events is easier to follow than in static snapshots.

Episode 1: 17–18 February 2025

Figure 5.2 shows predicted and true flows for the four edges over the 03:00–19:00 window on 17 February.

Morning reversal (southward to northward, ≈04:00–10:00). The reversal propagates from south to north: CALA→SUD crosses zero at 04:00, followed by CSUD→CNOR at 08:00 and CNOR→NORD at 09:00. The model captures the two southern transitions in time. The northern edges (CSUD→CNOR) and (CNOR→NORD) are predicted one hour late.

Afternoon reversal (northward to southward, ≈15:00–18:00). Flows along the northern corridors peak around 13:00–14:00 and then decay. CNOR→NORD reverses at 15:00, the model captures this correctly. CSUD→CNOR follows at 18:00; the model flags the reversal one hour early, though the true flow was already near zero at 17:00, suggesting that the apparent error partly reflects the one-hour granularity of the data rather than a genuine forecasting mistake. This ambiguity would likely be resolved with 15-minute resolution data.

The pattern repeats on 18 February with broadly similar timing. All southward-to-northward reversals are captured: CNOR→CSUD is again one hour late around 09:00, but is predicted in time at 16:00 when the flow reverses back. One notable difference is a brief back-and-forth on CNOR→NORD around 10:00, where the true flow momentarily reverses direction before continuing northward. The model, already stabilized on a northward prediction at that point, does not capture this transient.

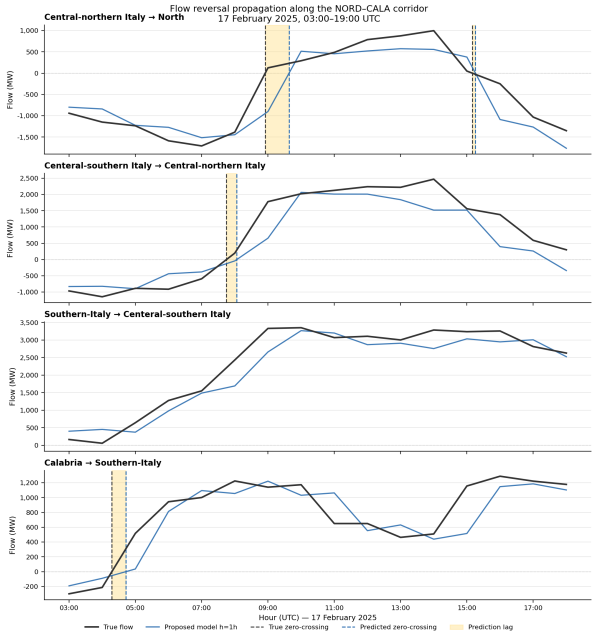


Figure 5.2: Predicted and true interzonal flows along the main Italian corridor on 17 February 2025. Dashed vertical lines mark the zero-crossing times for the true flow (dark) and the proposed model (blue), yellow shading indicates the prediction lag at each reversal.

Figure 5.3: Grid state at 05:00 on 17 February 2025: the two southern corridor edges (CALA→SUD, SUD→CSUD) have northward flows while CNOR→NORD and CSUD→CNOR flows are still toward south.

Episode 2: 17 March 2025

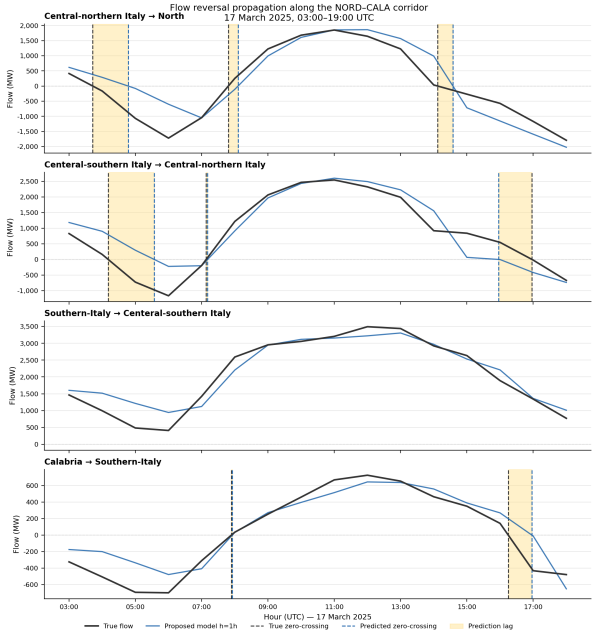


Figure 5.4: Predicted and true interzonal flows along the main Italian corridor on 17 March 2025. Dashed vertical lines mark the zero-crossing times for the true flow (dark) and the proposed model (blue); yellow shading indicates the prediction lag at each reversal.

Figure 5.4 shows the edges on 17 March, a day with a richer sequence of reversals than the February episode.

Early morning (northward to southward, ≈04:00–05:00). At 03:00 the flows are already directed south-to-north. CNOR→NORD reverses southward at 04:00 and CSUD→CNOR follows at 05:00. The model is one hour late on both transitions, consistent with the lag observed on the northern edges on February 17.

Mid-morning reversal (southward to northward, ≈08:00). At 08:00, three edges reverse nearly simultaneously: CALA→SUD, CSUD→CNOR, and CNOR→NORD. The model captures CALA→SUD and CSUD→CNOR in time, but is again one hour late on CNOR→NORD : the furthest edge from the signal originating in the south.

Figure 5.5: Grid state at 03:00 UTC on 17 March 2025, before the morning reversal sequence.

Afternoon reversal (northward to southward, ≈15:00–17:00). The corridor remains in south-to-north configuration until 14:00–15:00, when flows decay and reverse. CNOR→NORD crosses zero at 15:00 and is well captured by the model. CSUD→CNOR follows at 17:00, as in the February case the model

flags this reversal one hour early, at a point when the true flow was already near zero : again likely a granularity artifact of the hourly resolution rather than a structural forecasting error.

Comparison with XGBoost. Figure 5.6 shows XGBoost predictions on the same day. Overall accuracy is competitive, consistent with the quantitative results in Section 5.2, yet a qualitative difference is visible across all four edges: XGBoost exhibits a near-constant one-hour lag throughout the entire window, not only at zero-crossings. The proposed model, by contrast, tracks the true flow in phase for most of the day and incurs a lag only at the specific moments of directional transition on the northern edges. This distinction matters operationally: a systematic time shift means XGBoost would consistently signal a congestion event one hour after it has already occurred, whereas the GNN's lag is confined to the propagation delay across the corridor rather than a global forecasting latency. The difference is attributable to the graph structure: XGBoost is trained independently per edge and has no mechanism to integrate contemporaneous information from neighboring zones, forcing it to rely on lagged features alone.

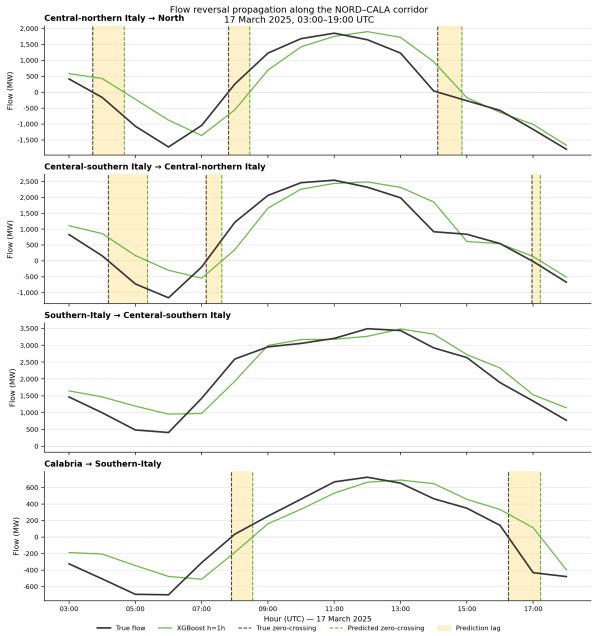


Figure 5.6: XGBoost predictions on 17 March 2025 for the same corridor edges as Figure 5.4. Unlike the proposed model, XGBoost exhibits a near-constant one-hour lag across the full window on all edges, rather than a lag confined to directional transitions on the northern edges.

Conclusion

Across both episodes, two patterns emerge consistently. First, reversals originating in the southern zones (CALA, SUD) are captured in time: they could be driven by local renewable ramps that are directly observable in the node features. Second, the propagation of these reversals to the northern edges (CSUD→CNOR, CNOR→NORD) is often one hour late. The one early prediction on CSUD→CNOR at the afternoon reversal is the only exception, and is plausibly explained by hourly data granularity. These results support the core claim that a spatially aware architecture captures the causal structure of flow propagation along the corridor, a behavior that a flat per-edge model, lacking any graph context, cannot reproduce.

6

Discussion

6.1. Summary of findings

This thesis investigated whether interzonal power flows in Italy's seven-zone electricity market can be forecast from publicly available market and weather data, using a spatio-temporal graph neural network operating directly on edges. The proposed encoder–decoder architecture combines LSTM temporal encoding, Transformer-based graph message passing with updatable edge representations, and a future-aware decoder that ingests day-ahead prices and weather forecasts.

The central finding is that the proposed model outperforms all baselines across every metric on the 2025 test set: NMAE of 52.1% (vs. 61.2% for the best baseline, ARIMA), DA of 89.8%, and CF1 of 60.2%. Crucially, this performance advantage is consistent across all six forecast horizons (h=1 to h=6).

Three secondary findings deserve emphasis. First, graph depth matters: removing the GNN entirely is statistically comparable to the weakest graph configurations, and the best results are achieved at 3–6 message-passing layers, consistent with the north–south diameter of Italy's grid topology. Second, future features help congestion detection but not flow magnitude: day-ahead prices alone add +2.57 pp to CF1 (significant at $p < 0.001$), while NMAE is unaffected. This asymmetry reveals that the encoder already captures most of the information relevant to mean flow magnitude, and the decoder's role is to sharpen the model's confidence near capacity boundaries. Third, the model's AUROC of 89.8% confirms that the flow-regression framing naturally produces a well-calibrated congestion score, making a dedicated binary classifier unnecessary for this task.

6.2. Were the research questions answered?

Can physical interzonal flows be forecast from public market and weather data? Yes. The proposed model achieves an NMAE of 52.1% on a test year, outperforming all baselines including Terna's own day-ahead schedule on magnitude metrics. This confirms that market prices and ERA5 weather observations carry sufficient information to produce operationally useful flow forecasts without access to network topology, injection data, or proprietary TSO models.

Does a graph-based architecture add value over per-edge models? Yes, but modestly and non-monotonically. The gain over the No-GNN baseline is statistically significant at 3, 4, and 6 layers ($p < 0.05$, Wilcoxon), but a single message-passing layer is actually significantly worse than no graph. This finding is practically important: naively adding graph structure does not help, the receptive field must be wide enough to cover the full grid diameter before spatial context becomes beneficial.

Does the model translate into actionable congestion signals? Yes, with caveats. The CF1 of 60.2% and AUROC of 89.8% demonstrate meaningful congestion discrimination, and the precision–recall analysis shows that the model achieves the best recall among high-precision models (73.5% precision, 53.7% recall), making it the most operationally suitable choice for a grid operator who prioritizes precision

while avoiding missed events.
The intraday trading simulation (Experiment A.4) provides a more direct and quantitative answer. At $h = 2$, the proposed model achieves a hit rate of 74% and a Sharpe ratio of 4.11, competitive with the best baselines. More tellingly, at $h = 6$ its average PnL rises to +0.84 €/MWh and its Sharpe ratio increases to 5.52, making it the strongest model at this horizon while most baselines deteriorate or stagnate. This improvement with lead time is the key operational result: the model captures structural physical information that is not yet reflected in intraday prices 6 h before delivery, whereas simpler baselines provide no additional predictive value beyond what the market already knows. The simulation represents a theoretical upper bound rather than a deployable strategy, but the directional finding is robust and motivates extending the model to longer horizons.

6.3. Limitations

ERA5 reanalysis instead of NWP forecasts. The most significant limitation is that all weather inputs come from ERA5 reanalysis, which uses observed data rather than forecasts. In a real deployment, weather forecasts for the delivery period would be available at gate closure, but these introduce forecast errors that the current evaluation does not account for. The reported metrics therefore represent an upper bound on deployed performance.

Absence of a validation set. With data spanning only from November 2021 (when the Calabria zone was introduced), a three-year training window leaves little room for a dedicated validation set. The best-snapshot early stopping strategy provides implicit regularization, but cannot detect overfitting to the training distribution. This is a structural data limitation rather than a modeling choice, and it will resolve naturally as more years of data become available.

Missing price data for Malta and Montenegro. Both connections are included in the graph, but the corresponding nodes lack price observations for a large fraction of the training period. The model compensates by relying more heavily on weather and flow signals for these edges, but the resulting predictions should be interpreted with caution. The SICI→MT edge in particular shows anomalously high NMAE (≈149%), likely attributable to this data scarcity combined with the low and irregular flow volumes on this cable.

Static graph topology. The graph structure is fixed over the entire study period. In practice, the Italian and European grids undergo planned outages, capacity revisions, and new interconnector commissioning that alter effective transmission limits. Incorporating dynamic topology (time-varying edge attributes or adjacency) would require significant architectural extension and is left for future work.

Zone-level weather aggregation. Averaging ERA5 observations across all provinces within a bidding zone discards intrazonal spatial heterogeneity. This is particularly consequential for large zones such as NORD, which spans both Alpine terrain (with distinct wind patterns) and the Po Valley (dominated by industrial load and PV). A hierarchical aggregation that first groups provinces by generation type before computing zone-level features could improve the weather signal for these connections.

Per-edge variability in performance. The proposed model does not uniformly outperform baselines on every edge. Linear regression is best on NORD→FR and CSUD→SUD, ARIMA is best on the SACODC edge, and XGBoost leads on NORD→AT. The SACODC and SACOAC submarine cable connections remain structurally difficult, with NMAE values of 95% and 103% respectively. These are edges where the model's learned representations are not expressive enough to recover the underlying pattern.

6.4. Societal and Temporal Context

Position within the broader DSAIT field

This work sits at the intersection of two active research fronts. Within spatio-temporal learning, the dominant trend applies increasingly expressive GNNs to node-level tasks such as traffic or renewable

generation forecasting, while edge-level regression has received comparatively little attention. This thesis contributes a concrete design principle with explicit edge representations updated during message passing, and sufficient graph depth for the network to make spatial context beneficial.

More broadly, the forecasting field is mainly about physics-informed hybrid architectures. This work takes a complementary position: rather than embedding physical equations, it uses network topology and market prices as implicit encoders of the physical state, operating without a calibrated power-flow model. The asymmetric finding that future features improve congestion detection but not flow magnitude reflects a general principle worth noting: threshold-crossing events respond to forecast sharpness near the decision boundary, not to average magnitude accuracy.

Stakeholder perspectives and real-world implications

The practical value of this work differs across stakeholders.
Transmission system operators (TSOs) gain not raw accuracy, as they already have proprietary grid models, but an independent, market congestion signal that can flag discrepancies between the market's implicit expectations and the physical system state, prompting earlier preventive re-dispatch.
Energy traders benefit most from the 6-hour horizon, where the model captures structural physical information not yet priced into intraday markets, though this represents an upper bound given the simple simulation and reliance on ERA5 reanalysis rather than NWP forecasts.
Regulators may note that inferring congestion from public data alone reduces information asymmetries between large and small market participants, but also raises the concern that widespread adoption of the same signal could induce correlated bidding and new intraday instabilities.
Consumers benefit indirectly: better congestion forecasting reduces costly emergency re-dispatch, the cost of which is ultimately socialized. Europe's estimated €4.2 billion congestion cost in 2023 [11] provides an order of magnitude on the value at stake.

Risks and responsible use

Three risks merit acknowledgment. First, point forecasts without uncertainty quantification encourage overconfident use near the capacity boundary, the quantile output extension proposed in future work is a prerequisite for responsible operational deployment. Second, the training data includes the 2022 energy crisis, an extreme regime poorly represented elsewhere, the model may behave unpredictably under future structural shifts outside its training distribution. Third, the 2025 Iberian blackout is a reminder that cascading failures can develop within seconds: a forecasting tool that reduces a TSO operator's vigilance without meeting a certified reliability standard could, in a worst case, contribute to delayed intervention during a fast-developing event. This work is intended as a decision-support signal, not as an autonomous control input.

Temporal outlook

As NWP forecast quality improves and intraday markets deepen across Europe, the gap between the ERA5 upper bound reported here and a deployed system is expected to narrow. The ENTSO-E bidding zone review and the expansion of smart metering infrastructure will progressively alter both the spatial structure of congestion and the observability of its drivers.

6.5. Future work

NWP forecast inputs. The most impactful near-term extension is replacing ERA5 reanalysis with operational NWP forecasts at the appropriate lead times. This would convert the current upper-bound evaluation into a realistic deployment assessment and reveal which lines are most sensitive to weather forecast uncertainty.

Intra-zonal heterogeneity. A two-level hierarchical graph, where province-level nodes are aggregated into bidding zones through a learned pooling mechanism, would allow the model to exploit spatial variation within zones, particularly for large zones like NORD where Alpine and Po Valley generation profiles are substantially different. This is directly motivated by the performance gap on the NORD→FR and NORD→CH corridors, where the current zone-averaged weather signal may be too coarse to resolve Alpine wind dynamics.

Additional features. Integrating explicit physical state variables from ENTSO-E, specifically zonal load and actual generation per energy source, could significantly refine the model's accuracy. While the current model uses weather as a proxy for renewable production, explicit generation data could better account for the non-linear relationship between meteorological conditions and actual grid injections. Similarly, incorporating load profiles would allow the model to directly observe the net-demand imbalances that drive interzonal flows, rather than relying solely on price as a latent signal for these physical requirements.

Extension to the European graph. Scaling from Italy's 7-zone graph to the ENTSO-E bidding zone graph (~30 nodes, ~100 cross-border edges) is computationally straightforward given the architecture's linear scaling in the number of edges. The main open question is data harmonization: cross-border flow data from ENTSO-E has varying completeness across countries and time periods, and a unified preprocessing pipeline would be needed.

Probabilistic outputs. The current model produces point forecasts. For congestion monitoring, a probabilistic output, and specifically, a predicted distribution over the congestion ratio, would allow operators to set risk-adjusted thresholds, which is more actionable than a hard binary flag. A natural and lightweight extension is to replace the scalar MLP output head with a multi-quantile head trained with the pinball loss. Rather than predicting a single value $\hat{f}_{ij}^{t+h}$, the model would simultaneously predict a set of quantiles $\hat{q}_\tau \in \{q_{0.1}, q_{0.5}, q_{0.9}, \ldots\}$, each penalized by its own asymmetric pinball loss. The total loss is summed across all target quantiles. This approach requires minimal architectural change as only the output head is modified, and preserves the full encoder-decoder-GNN structure.

References

[1] Union for the Co-ordination of Transmission of Electricity (UCTE). *Final Report of the Investigation Committee on the 28 September 2003 Blackout in Italy*. Tech. rep. Issued on behalf of the Investigation Committee. UCTE, Apr. 2004. URL: https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/ce/otherreports/20040427_UCTE_IC_Final_report.pdf.

[2] Roberto Pietrantuono et al. *Critical Infrastructure Protection: Threats, Attacks and Countermeasures*. 2014. URL: http://wpage.unina.it/roberto.pietrantuono/deliverables/Tenace-Deliverable1.pdf.

[3] Sergey V. Buldyrev et al. "Catastrophic cascade of failures in interdependent networks". In: *Nature* 464 (2010), pp. 1025–1028.

[4] ICS Investigation Expert Panel. *Grid Incident in Spain and Portugal on 28 April 2025, Final Report*. 2026. URL: https://www.entsoe.eu/publications/blackout/28-april-2025-iberian-blackout/#Publications_&_Documents.

[5] S. Clarke et al. "Renewable energy sources are essential to mitigate climate change". In: *Nature Climate Change* 12.6 (2022), pp. 505–512.

[6] Hannah Bloomfield et al. "Weather and climate information for renewable energy forecasting". In: *Renewable and Sustainable Energy Reviews* 152 (2021), p. 111699. DOI: 10.1016/j.rser.2021.111699.

[7] V. S. K. Harish and A. Kumar. "A review of renewable energy forecasting methods and applications". In: *Renewable and Sustainable Energy Reviews* 133 (2020).

[8] S. Vanting and P. Jorgensen. "Data-driven approaches for renewable energy forecasting: a review". In: *IEEE International Conference on Smart Energy Systems*. 2021.

[9] Oliver Neumann et al. "Using weather data in energy time series forecasting: the benefit of input data transformations". In: *Energy Informatics* 6 (Nov. 2023). DOI: 10.1186/s42162-023-00299-8.

[10] Tech Insights. *Solving Grid Congestion: How GETs Maximize Renewable Integration*. Technical white paper. 2023. URL: https://get.techinsights.com/blog/solving-grid-congestion.

[11] European Parliamentary Research Service. *Electricity Grid Congestion in the EU: Trends, Costs and Policy Options*. Tech. rep. European Parliament, 2025. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/772854/EPRS_BRI%282025%29772854_EN.pdf.

[12] Grid Strategies LLC. *Transmission Congestion Costs in the U.S. RTOs*. Tech. rep. July 2023. URL: https://gridstrategiesllc.com/wp-content/uploads/2023/07/GS_Transmission-Congestion-Costs-in-the-U.S.-RTOs1.pdf.

[13] International Energy Agency. *Renewables 2019: Market Analysis and Forecast from 2019 to 2024*. 2019. URL: https://www.iea.org/reports/renewables-2019.

[14] Daniel K. Molzahn et al. "A survey of relaxations and approximations of the power flow equations". In: *Foundations and Trends in Electric Energy Systems* 4.1-2 (2017), pp. 1–221. DOI: 10.1561/3100000004.

[15] Adam B. Birchfield et al. "Grid structural characteristics as validation criteria for synthetic networks". In: *IEEE Transactions on Power Systems* 32.4 (2017), pp. 3258–3265. DOI: 10.1109/TPWRS.2016.2616385.

[16] Hao Pan, Ning Zhang, and Chongqing Kang. "Data-driven probabilistic congestion forecasting in power systems". In: *IEEE Transactions on Power Systems* 34.4 (2019), pp. 2760–2770. DOI: 10.1109/TPWRS.2019.2899487.

[17] Yifan Xu, Qingyu Zhao, and Zhaoyang Dong. "Learning-based approaches for transmission congestion prediction". In: *Applied Energy* 295 (2021), p. 117022. DOI: 10.1016/j.apenergy.2021.117022.

[18] Miomir Kostic, Nenad Sijakovic, and Valeri Mladenov. "Automation of the day ahead congestion forecast procedure". In: July 2011, pp. 390–393.

[19] Florian Ziel and Rafal Weron. "Forecasting electricity price spreads: the impact of renewable generation and market coupling". In: *Energy Economics* 79 (2019), pp. 170–182. DOI: 10.1016/j.eneco.2018.06.009.

[20] Martin Hildmann, Andreas Ulbig, and Göran Andersson. "The role of market design when integrating variable renewable generation: a review of the literature". In: *Electric Power Systems Research* 119 (2015), pp. 230–238. DOI: 10.1016/j.epsr.2014.09.005.

[21] G. Thomassen and A. Fuhrmanek. *Future-Proofing the European Power Market: Redispatch and Congestion Management*. Florence School of Regulation, 2020.

[22] Philipp Staudt. "Transmission Congestion Management in Electricity Grids". PhD thesis. Karlsruhe Institute of Technology, 2019.

[23] Faddy Ardia. "Empirical Analysis of the Italian Electricity Market". PhD thesis. Università degli Studi di Milano, 2018.

[24] Terna S.p.A. *New Electricity Market Zones*. 2023. URL: https://lightbox.terna.it/en/insight/new-electricity-market-zones.

[25] Mahmood Hosseini Iman. "Empirical Analysis of Inter-Zonal Congestion in the Italian Electricity Market Using Multinomial Logistic Regression". In: *Energies* 14.9 (2021), p. 2524.

[26] Terna. *installed-renewables*. 2025. URL: https://dati.terna.it/en/generation#installed-renewables.

[27] Wikipedia. *Solar PV and the Conto Energia*. November 2025. URL: https://en.wikipedia.org/wiki/Renewable_energy_in_Italy#Solar_PV_and_the_Conto_Energia.

[28] Gestore dei Servizi Energetici (GSE). *Rapporto Statistico GSE – Energia da Fonti Rinnovabili in Italia – Anno 2023*. Tech. rep. GSE, 2024. URL: https://www.gse.it/documenti_site/Documenti%20GSE/Rapporti%20statistici/Rapporto%20Statistico%20GSE%20-%20Energia%20da%20FER%20in%20Italia%20-%20anno%202023.pdf.

[29] Terna S.p.A. *Statistical Data on the Italian Electricity System 2023*. Tech. rep. Available at https://www.terna.it/en/electric-system/statistical-data. Terna – Rete Elettrica Nazionale, 2024.

[30] Solargis. *Global Solar Atlas – Global Horizontal Irradiance (GHI) Map of Italy*. World Bank Group, Energy Sector Management Assistance Program (ESMAP). 2023. URL: https://globalsolaratlas.info/map.

[31] Ciaran O'Connor et al. "A Review of Electricity Price Forecasting Models in the Day-Ahead, Intra-Day, and Balancing Markets". In: *Energies* (2025).

[32] José R. Andrade et al. "Probabilistic Price Forecasting for Day-Ahead and Intraday Markets: Beyond the Statistical Model". In: *Sustainability* 9.11 (2017).

[33] Hazem Abdel-Khalek et al. "Forecasting Cross-Border Power Transmission Capacities in Central Western Europe Using Artificial Neural Networks". In: *Energy Informatics* (2019).

[34] Anders Løland, Egil Ferkingstad, and Magnus Wilhelmsen. "Forecasting Transmission Congestion in the Nordic Power Market". In: *The Journal of Energy Markets* 5.3 (2012), pp. 85–107.

[35] Hosseini Imani M. "Empirical Analysis of Inter-Zonal Congestion in the Italian Electricity Market Using Multinomial Logistic Regression". In: *Energies* 17 (2024), p. 5901.

[36] Katarzyna Maciejowska. "Fundamental and speculative shocks, what drives electricity prices?" In: *11th International Conference on the European Energy Market (EEM14)*. 2014, pp. 1–5.

[37] Rafał Weron. "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4 (2014), pp. 1030–1081. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2014.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0169207014001083.

[38] Petar Veličković et al. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML]. URL: https://arxiv.org/abs/1710.10903.

[39] Ding Lin, Han Guo, and Jianhui Wang. "Diffusion Model Based Probabilistic Day-ahead Load Forecasting". In: *arXiv* (2025).

[40] Flavio Corradini et al. "A systematic literature review of spatio-temporal graph neural network models for time series forecasting and classification". In: *Neural Networks* 195 (Mar. 2026), p. 108269. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2025.108269. URL: http://dx.doi.org/10.1016/j.neunet.2025.108269.

[41] Jelena Simeunovic et al. "Spatio-Temporal Graph Neural Networks for Multi-Site PV Power Forecasting". In: *IEEE Transactions on Sustainable Energy* 13.2 (Apr. 2022), pp. 1210–1220. ISSN: 1949-3037. DOI: 10.1109/tste.2021.3125200. URL: http://dx.doi.org/10.1109/TSTE.2021.3125200.

[42] Yuting Ji, Robert J. Thomas, and Lang Tong. *Probabilistic Forecast of Real-Time LMP and Network Congestion*. 2016. arXiv: 1503.06171 [stat.AP]. URL: https://arxiv.org/abs/1503.06171.

[43] Alejandro Hernandez-Matheus et al. "Congestion forecast framework based on probabilistic power flow and machine learning for smart distribution grids". In: *International Journal of Electrical Power & Energy Systems* 156 (2024), p. 109695.

[44] Mohan G.M.and Kumar T.A. and Srujana A. et al. "Real-time congestion control using cascaded LSTM deep neural networks for deregulated power markets". In: *Electric Power Systems Research* 15 (2025), p. 30581.

[45] Amir Bagheri et al. "Probabilistic optimal co-planning of distributed series reactor and dynamic line thermal rating for congestion management of highly-renewable-penetrated electric power systems". In: *Electric Power Systems Research* 248 (2025), p. 111924.

[46] Hosseini Imani M et al. "Impact of wind and solar generation on the Italian zonal electricity price". In: *Energies* 14 (2021), p. 5858.

[47] Mahmood Hosseini Imani. "Empirical analysis of inter-zonal congestion in the Italian electricity market using multinomial logistic regression". In: *Energies* 17 (2024), p. 5901.

[48] Bernardi Mauro and Lisi Federico. "Point and interval forecasting of zonal electricity prices and demand using heteroscedastic models: The IPEX case". In: *Energies* 13 (2020), p. 6191.

[49] Rafal Weron. "Electricity price forecasting: A review of the state-of-the-art with a focus on the Nordic power market". In: *International Journal of Forecasting* 30.4 (2014), pp. 1030–1081.

[50] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2008.

[51] Granville Tunnicliffe Wilson. "Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1". In: *Journal of Time Series Analysis* 37 (Mar. 2016), n/a–n/a. DOI: 10.1111/jtsa.12194.

[52] Chris Chatfield. *Time-Series Forecasting*. Chapman and Hall/CRC, 2000.

[53] Malte Lehna, Fabian Scheller, and Helmut Herwartz. "Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account". In: *Energy Economics* 106 (2022), p. 105742.

[54] Cu Y et al. "A Time Series Decomposition-Based Interpretable Electricity Price Forecasting Method". In: *Energies* 18 (2025), p. 664.

[55] Hochreiter Sepp and Schmidhuber Jürgen. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[56] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179/.

[57] A. Berizzi, M. Delfanti, and M. Merlo. "Congestion management in a zonal market by a neural network approach". In: *Electric Power Systems Research* 79.11 (2009), pp. 1490–1497.

[58] Johan Mathe et al. "PVNet: A LRCN Architecture for Spatio-Temporal Photovoltaic Power-Forecasting from Numerical Weather Prediction". In: (2024). arXiv: 1902.01453 [cs.LG]. URL: https://arxiv.org/abs/1902.01453.

[59] Weicong Kong et al. “Short-term residential load forecasting based on LSTM recurrent neural network”. In: *IEEE Transactions on Smart Grid* 10.1 (2019), pp. 841–851.

[60] Daniel L. Marino, Kasun Amarasinghe, and Milos Manic. “Building energy load forecasting using Deep Neural Networks”. In: *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*. 2016, pp. 7046–7051.

[61] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. “Sequence to sequence learning with neural networks”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, 2014, pp. 3104–3112.

[62] David Salinas et al. “DeepAR: Probabilistic forecasting with autoregressive recurrent networks”. In: *International Journal of Forecasting* 36.3 (2020), pp. 1181–1191. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301888.

[63] Bryan Lim and Stefan Zohren. “Temporal Fusion Transformers for interpretable multi-horizon time series forecasting”. In: *International Journal of Forecasting* 37.4 (2021), pp. 1748–1764.

[64] Zonghan Wu et al. “A Comprehensive Survey on Graph Neural Networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2021), pp. 4–24.

[65] Jie Zhou et al. “Graph Neural Networks: A Review of Methods and Applications”. In: *AI Open* 1 (2020), pp. 57–81.

[66] Wenlong Liao et al. *A Review of Graph Neural Networks and Their Applications in Power Systems*. 2021. arXiv: 2101.10025 [cs.LG]. URL: https://arxiv.org/abs/2101.10025.

[67] Minsoo Kim, Vladimir Dvorkin, and Jip Kim. *Probabilistic Dynamic Line Rating Forecasting with Line Graph Convolutional LSTM*. 2025. arXiv: 2411.12963 [eess.SY]. URL: https://arxiv.org/abs/2411.12963.

[68] Dhruv Suri and Mohak Mangal. *PowerGNN: A Topology-Aware Graph Neural Network for Electricity Grids*. 2025. arXiv: 2503.22721 [cs.LG]. URL: https://arxiv.org/abs/2503.22721.

[69] Francesco Fusco et al. *Knowledge- and Data-driven Services for Energy Systems using Graph Neural Networks*. 2021. arXiv: 2103.07248 [cs.LG]. URL: https://arxiv.org/abs/2103.07248.

[70] Dominik Beinert et al. “Power Flow Forecasts at Transmission Grid Nodes Using Graph Neural Networks”. In: *Energy and AI* 14 (2023), p. 100262.

[71] Seyedamirhossein Talebi and Kaixiong Zhou. *Graph Neural Networks for Efficient AC Power Flow Prediction in Power Grids*. Feb. 2025. DOI: 10.48550/arXiv.2502.05702.

[72] Feiyan Sun et al. “A survey on spatio-temporal series prediction with deep learning: taxonomy, applications, and future directions”. In: *Neural Comput. Appl.* 36.17 (Apr. 2024), pp. 9919–9943. ISSN: 0941-0643. DOI: 10.1007/s00521-024-09659-1. URL: https://doi.org/10.1007/s00521-024-09659-1.

[73] Majdi Mansouri, Khadija Attouri, and Shady Refaat. “Multimodal Learning Techniques for Time Series Forecasting in Renewable Energy Systems: A Comprehensive Survey”. In: *IEEE Access* PP (Jan. 2025), pp. 1–1. DOI: 10.1109/ACCESS.2025.3602914.

[74] Xwégnon Ghislain Agoua, Robin Girard, and Georges Kariniotakis. “Photovoltaic Power Forecasting: Assessment of the Impact of Multiple Sources of Spatio-Temporal Data on Forecast Accuracy”. In: *Energies* 14 (2021), p. 1432.

[75] Bowoo Kim and Dongjun Suh. “A Hybrid Spatio-Temporal Prediction Model for Solar Photovoltaic Generation Using Numerical Weather Data and Satellite Images”. In: *Remote Sensing* 12 (Nov. 2020), p. 3706. DOI: 10.3390/rs12223706.

[76] Clara M. St. Martin, Julie K. Lundquist, and Mark A. Handschy. “Variability of interconnected wind plants: correlation length and its dependence on variability time scale”. In: *Environmental Research Letters* 10.4 (2015), p. 044004.

[77] Jelena Simeunovic et al. “Spatio-temporal graph neural networks for multi-site PV power forecasting”. In: *CoRR* abs/2107.13875 (2021). arXiv: 2107.13875. URL: https://arxiv.org/abs/2107.13875.

[78] Akin Tascikaraoglu et al. “Compressive spatio-temporal forecasting of meteorological quantities and photovoltaic power”. In: *2017 IEEE Manchester PowerTech*. 2017, pp. 1–1. DOI: 10.1109/PTC.2017.7981257.

[79] Amir Miraki, Pekka Parviainen, and Reza Arghandeh. “Probabilistic forecasting of renewable energy and electricity demand using Graph-based Denoising Diffusion Probabilistic Model”. In: *Energy and AI* 19 (2025), p. 100459. ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2024.100459. URL: https://www.sciencedirect.com/science/article/pii/S2666546824001253.

[80] Yue Jiang et al. *SAGDFN: A Scalable Adaptive Graph Diffusion Forecasting Network for Multivariate Time Series Forecasting*. 2024. arXiv: 2406.12282 [cs.LG]. URL: https://arxiv.org/abs/2406.12282.

[81] Flavio Corradini et al. *A Systematic Literature Review of Spatio-Temporal Graph Neural Network Models for Time Series Forecasting and Classification*. 2025. arXiv: 2410.22377 [cs.LG]. URL: https://arxiv.org/abs/2410.22377.

[82] Seong Ho Pahng and Sahand Hormoz. *Improving Graph Neural Networks by Learning Continuous Edge Directions*. 2025. arXiv: 2410.14109 [cs.LG]. URL: https://arxiv.org/abs/2410.14109.

[83] Dominik Fuchsgruber et al. “Graph Neural Networks for Edge Signals: Orientation Equivariance and Invariance”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=XWBE90OYlH.

[84] Muberra Ozmen, Florence Regol, and Thomas Markovich. “Benchmarking Edge Regression on Temporal Networks”. In: *Journal of Data-centric Machine Learning Research* (2024). Dataset Certification, Reproducibility Certification. ISSN: XXXX-XXXX. URL: https://openreview.net/forum?id=4k4cocpuSw.

[85] Asim Javed and Reza R. Derakhshani. “Machine Learning Ensembles for Grid Congestion Price Forecasting”. In: *2022 North American Power Symposium (NAPS)*. 2022, pp. 1–6. DOI: 10.1109/NAPS56150.2022.10012217.

[86] Kemp J.M., Millstein D., and Gorman W. et al. “Electric transmission value and its drivers in United States power markets.” In: *Nature* 16 (2025), p. 8055.

[87] Haobo Fu et al. “Data-Driven Proactive Early Warning of Grid Congestion Probability Based on Multiple Time Scales”. In: *Energies* 18 (May 2025), p. 2530. DOI: 10.3390/en18102530.

[88] Arturo Berizzi, Mario Delfanti, and Mauro Merlo. “Congestion Management in a Zonal Market by a Neural Network Approach”. In: *IEEE PowerTech Conference*. 2009.

[89] Federica Davò et al. “Forecasting Italian electricity market prices using a Neural Network and a Support Vector Regression”. In: *2016 AEIT International Annual Conference (AEIT)*. 2016, pp. 1–6. DOI: 10.23919/AEIT.2016.7892764.

[90] Alessandro Franco and Cecilia Pagliantini. “Forecasting Electricity Demand in Renewable-Integrated Systems: A Case Study from Italy Using Recurrent Neural Networks”. In: *Electricity* 6 (2025), p. 30.

[91] Copernicus Climate Change Service. *ERA5 hourly data on single levels from 1940 to present*. https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels. Accessed: 2025-12-15. 2026.

[92] Terna. *Electricity production by province and energy source*. https://dati.terna.it/en/download-center#/production/electricity-source. 2026.

[93] Oliver Neumann et al. “Using weather data in energy time series forecasting: the benefit of input data transformations”. In: *Energy Informatics* 6 (Nov. 2023). DOI: 10.1186/s42162-023-00299-8.

A

Additional experiments

A.1. Experiment A: How does forecast accuracy degrade with horizon?

At which forecast horizons does the model provide the most value, and how quickly does accuracy degrade?

A market participant would use this model at a specific gate closure horizon. Understanding how accuracy changes with lead time informs the practical deployment window and the design of the training objective.

A.1.1. Results

Tables A.1–A.3 report NMAE, DA, and CF1 for all models across horizons h=1 to h=6. Detailed standard deviations are reported in Appendix B.2, and discussed here only when relevant.

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	69.8	69.8	69.8	69.8	69.8	69.8
Long persistence	134.9	134.9	135.0	135.0	135.0	135.0
Scheduled flow*	204.6	204.6	204.6	204.6	204.6	204.6
ARIMA	**28.1**	46.4	59.5	69.7	78.2	85.2
LSTM	60.5	70.8	77.8	82.2	85.9	87.6
XGBoost	65.3	75.5	81.7	85.9	88.8	91.1
Linear regression	62.3	72.5	78.2	81.7	84.0	85.6
GNN	73.7	75.3	76.5	77.7	78.6	79.6
Proposed model	32.5	**45.1**	**52.7**	**57.6**	**61.0**	**63.6**

Table A.1: NMAE ↓ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

A.1.2. Analysis

Overall degradation pattern. All models degrade with horizon, but at very different rates and with different asymptotic behaviours. The proposed model degrades smoothly and sub-linearly: NMAE grows from 32.5% at h=1 to 63.6% at h=6, a factor of roughly 2×, while ARIMA grows from 28.1% to 85.2%, a factor of 3×. This means ARIMA leads at h=1 but is overtaken by the proposed model already at h=2, and the gap widens consistently with horizon. While the proposed model does not systematically beat the baselines for all horizons on DA and CF1 metrics (as these were not the primary optimization targets), it is always ranking either first or second. The per-horizon results are thus consistent with the superior position of the proposed model overall, as noticed in experiment 5.2.

The h=1 anomaly. At h=1, ARIMA achieves its best result (NMAE 28.1%, DA 94.9%, CF1 76.3%), outperforming the proposed model on all three metrics. This is expected: at a one-step-ahead horizon,

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	86.1	86.1	86.1	86.1	86.1	86.1
Long persistence	76.5	76.5	76.5	76.5	76.5	76.5
Scheduled flow*	86.4	86.4	86.4	86.4	86.4	86.4
ARIMA	**94.9**	**91.7**	89.0	86.7	84.7	83.0
LSTM	91.4	89.5	88.3	87.3	86.9	86.4
XGBoost	92.1	90.1	89.1	88.4	87.8	87.4
Linear regression	90.5	88.7	87.5	86.8	86.3	85.9
GNN	84.2	84.0	83.9	83.7	83.5	83.4
Proposed model	93.3	91.1	**89.7**	88.8	88.3	87.8

Table A.2: DA ↑ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	56.9	56.9	56.9	**56.9**	**56.9**	**56.9**
Long persistence	16.2	16.2	16.2	16.2	16.2	16.2
Scheduled flow*	37.6	37.6	37.6	37.6	37.6	37.6
ARIMA	**76.3**	**66.5**	**59.6**	54.9	51.3	48.6
LSTM	63.9	55.6	50.4	47.2	45.1	44.4
XGBoost	70.8	57.9	50.5	45.5	42.7	41.4
Linear regression	62.6	52.2	46.1	42.6	39.8	38.2
GNN	40.9	40.5	40.1	39.8	39.5	39.4
Proposed model	73.4	64.2	58.6	56.3	55.1	53.4

Table A.3: CF1 ↑ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

the autoregressive structure of the flow series alone is a very strong predictor, and ARIMA is specifically designed to exploit it. The proposed model (NMAE 32.5%) is already competitive but pays a small cost for the additional complexity of its encoder. Beyond h=1, the autoregressive advantage of ARIMA erodes as prediction errors compound, while the proposed model's predictions remain more stable.

Practical deployment window. Italy's intraday (MI) markets have gate closures ranging from roughly 2 hours (MI-A2) to 8 hours (MI-A1) before delivery. The proposed model is most relevant for MI-A2, corresponding approximately to h=3 to h=6, where it holds a significant NMAE advantage over other models. At h=6, the model still provides a CF1 of 53.4% compared to 48.6% for ARIMA and 44.4% for LSTM, retaining practical value for early intraday positioning. The model is less critical at h=1 given ARIMA's strong performance there, but ARIMA requires per-edge and per-horizon fitting, making it indeed a considerable opponent.

Congestion detection is more resilient than magnitude accuracy. CF1 degrades more slowly than NMAE for the proposed model: from 73.4% at h=1 to 53.4% at h=6 (a drop of 20 pp), whereas NMAE roughly doubles. This asymmetry is operationally valuable: even at longer horizons where the precise flow magnitude becomes uncertain, the model retains meaningful ability to flag congested versus uncongested states. Notably, from h=4 onward, the proposed model's CF1 (56.3–53.4%) exceeds ARIMA (54.9–48.6%) by a small margin while also beating all other models except for short persistence.

A.1.3. Comparison with industry benchmarks and long-term stability

To assess performance at horizons beyond the primary evaluation window ($h > 6$) and to benchmark against an industry reference, the model was extended to predict up to 24 hours ahead and compared against forecasts provided by Electricity Maps for a common subset of 10 major Italian interconnections. These edges cover the primary transmission lines of the Italian grid: the five main domestic edges (Calabria → Sicilia; Central-southern → Sardegna; Central-southern → Southern-Italy; Central-northern → Central-southern; Central-northern → North) and the five main international borders (North → {AT, CH, FR, SI}; Southern-Italy → GR).

Training at extended horizons: the role of curriculum learning. Extending the model to $h = 24$ required a non-trivial adjustment to the training procedure. An initial attempt with a standard uniform loss over all 24 horizon steps produced a model that performed worse than the short-persistence baseline even at short horizons ($h \leq 6$).
The cause is a gradient competition effect: with uniform weighting, the model sacrifices near-horizon accuracy to marginally reduce error at distant steps, where any model has limited predictive power.
To address this, a staged curriculum learning approach was adopted. Training is divided into N discrete stages, each lasting a fixed number of epochs. At stage 0, the per-horizon loss weights decay exponentially with step index, concentrating the gradient signal on the first few horizons.
Across successive stages the weight profile is gradually flattened until it is uniform at the final stage. Between stages the early-stopping patience counter is reset, so the model can re-adapt to the new loss landscape without being terminated prematurely. This curriculum allowed the model to first learn the strong short-range temporal structure and then progressively extend its predictive horizon, recovering competitive accuracy at $h \leq 6$ while achieving meaningful performance out to $h = 24$.

Performance on major edges. On the common subset of 10 edges (Table A.4), the proposed model achieves the best NMAE at all three reported horizons: 38.9% at $h = 6$, 42.0% at $h = 12$, and 41.4% at $h = 24$. Two aspects of this result deserve attention.

First, the model's degradation from $h = 6$ to $h = 24$ is remarkably small: only +2.5 pp. This near-flat profile is shared by Electricity Maps (45.3 → 48.0, +2.7 pp) and Linear Regression (40.5 → 44.6, +4.1 pp), and contrasts sharply with ARIMA, which degrades from 64.6% at $h = 6$ to 79.6% at $h = 12$ before partially recovering to 52.0% at $h = 24$ as its 24-step seasonal lag becomes useful. The flat profile of the GNN and the industry benchmark suggests that, for the major edges, the dominant drivers of flow are largely predictable from features available well in advance of delivery.

Second, the standard deviation of the proposed model across edges is consistently tight (±13–15 pp), comparable to Electricity Maps and well below ARIMA (±35–48 pp). This indicates that the GNN's advantage is not driven by a few easy edges but is distributed across the edge subset.

Linear Regression is a noteworthy second-best on this subset, finishing only 1.6 pp behind the GNN at $h = 6$ and 3.2 pp at $h = 24$, suggesting that the major edges are relatively well-explained by linear feature combinations and that the GNN's additional expressivity provides a consistent but modest gain over a strong linear baseline.

Model	h=6	h=12	h=24
Proposed model	**38.9 ± 14.2%**	**42.0 ± 15.2%**	**41.4 ± 13.6%**
Short persistence	46.1 ± 18.4%	46.1 ± 18.4%	46.1 ± 18.4%
Linear Regression	<u>40.5 ± 16.3%</u>	<u>43.4 ± 16.6%</u>	<u>44.6 ± 15.7%</u>
ARIMA	64.6 ± 35.7%	79.6 ± 47.8%	52.0 ± 23.1%
LSTM	42.6 ± 16.2%	45.5 ± 17.0%	45.6 ± 16.2%
Electricity Maps	45.3 ± 14.2%	45.7 ± 14.4%	48.0 ± 14.1%

Table A.4: NMAE ↓ (%) per horizon on the 2025 test set, only for subset of edges. **Bold** indicates best, <u>underline</u> second-best.

Degradation on the full edge set. The picture changes substantially when results are computed over all edges (Table A.5). The proposed model's NMAE at $h = 6$ rises from 38.9% on the major corridors to 64.5% on the full set, a gap of +25.6 pp driven entirely by the harder edges not covered by Electricity Maps. More critically, at $h = 12$ and $h = 24$ the model (72.1%, 73.0%) is overtaken by the short-persistence baseline (69.8%, horizon-invariant by construction), with the crossover occurring at around $h \approx 10$.

The explosion in standard deviation from ±14 pp on the common subset to ±45–55 pp on the full set, reveals that the additional edges are highly heterogeneous: some remain well-forecast, but others are dominated by stochastic or locally-driven dynamics that currently used features cannot capture. In contrast, Linear Regression (86%) and LSTM (89%) fare considerably worse on the full set, suggesting

these models overfit or extrapolate poorly on volatile edges, the GNN's graph inductive bias makes it more robust, even if ultimately insufficient for this subset. Electricity Maps achieves 45.3–48.0% on the full set, but since it only covers its own 10 edges, this figure is not directly comparable.

These results suggest that the proposed model is best understood as a specialist for structurally predictable, high-volume corridors. A practical deployment could route predictions for minor or high-variance edges to a simpler fallback (like short persistence), reserving the GNN for the edges where its structural inductive bias provides a genuine advantage.

Model	h=6	h=12	h=24
Proposed model	64.5 ± 45.8%	72.1 ± 52.5%	73.0 ± 54.8%
Short persistence	69.8 ± 45.0%	69.8 ± 45.0%	69.8 ± 45.1%
Linear Regression	86.2 ± 96.9%	89.5 ± 96.6%	91.1 ± 96.6%
ARIMA	84.9 ± 49.6%	103.1 ± 58.5%	76.5 ± 47.2%
LSTM	89.1 ± 84.4%	93.6 ± 86.5%	92.6 ± 84.2%
Electricity Maps*	**45.3 ± 14.2%**	**45.7 ± 14.4%**	**48.0 ± 14.1%**

Table A.5: NMAE ↓ (%) per horizon, on all edges. **Bold** indicates best, underline second-best. * indicates the model's results are only on a subset of edges.

A.2. Experiment B: Does the graph structure help?

Is spatial context from neighboring zones informative for flow forecasting, and does performance improve with graph depth?

The GNN is the central architectural novelty. If removing graph structure does not degrade performance, the model could be simplified to a per-edge LSTM at a fraction of the computational cost. This experiment isolates the contribution of the graph by varying the number of message-passing layers, including the degenerate case of zero layers (GNN disabled). Hyperparameters were fixed across all configurations to isolate the effect of message-passing depth.

A.2.1. Results

Table A.6 reports the three primary metrics averaged across all 16 directional edges and 6 forecast horizons.

Variant	NMAE ↓	DA ↑	CF1 ↑
No GNN (0 layers)	53.5	89.6	58.2
1 GNN layer	54.1	89.4	58.1
2 GNN layers	53.4	89.5	58.8
3 GNN layers	52.2	89.8	58.5
4 GNN layers	52.9	89.8	58.6
5 GNN layers	53.6	**90.0**	58.8
6 GNN layers	**52.1**	89.8	**60.2**

Table A.6: Effect of graph depth on test performance. Bold marks the best value per column.

Overall trend. Adding message-passing layers progressively reduces NMAE, with the best single-metric score reached at 6 layers (52.1%) and a local minimum at 3 layers (52.2%). Directional accuracy (DA) and congestion F1 (CF1) follow a similar pattern, with 6 layers achieving the highest CF1 (60.2) and 5 layers the highest DA (90.0%). The gain from 0 to 6 layers is moderate but consistent: approximately −1.4 pp in NMAE and +2.0 pp in CF1.

Statistical significance. Paired Wilcoxon signed-rank tests against the *No GNN* baseline (96 edge×horizon pairs) reveal a non-monotonic pattern (Table A.7).

Variant	ΔNMAE	W	p	sig
1 GNN layer	+0.56	1762.0	0.039	*
2 GNN layers	−0.15	2078.0	0.361	ns
3 GNN layers	−1.30	1673.0	0.017	*
4 GNN layers	−0.61	1692.0	0.020	*
5 GNN layers	+0.12	2072.0	0.350	ns
6 GNN layers	−1.41	1642.0	0.012	*

Table A.7: Paired Wilcoxon signed-rank tests for NMAE vs. *No GNN* baseline (n = 96 edge×horizon pairs). ΔNMAE = variant − No GNN, negative values indicate improvement. * p<0.05.

Three key observations emerge. First, a single message-passing layer is significantly worse than no graph at all (ΔNMAE = +0.56 pp, p = 0.039): with only one hop, each node aggregates its immediate neighbours but lacks the receptive field to resolve conflicting local signals, introducing noise rather than useful context. Second, improvements become statistically significant once the network reaches sufficient depth: 3, 4, and 6 layers all outperform the No-GNN baseline ($p < 0.05$), whereas 2 and 5 layers do not differ significantly. Third, the best overall model (6 layers, ΔNMAE = −1.41 pp) corresponds to the fully tuned configuration from Experiment 1, suggesting that the hyperparameter search naturally favored deeper graphs.

A.2.2. Analysis

Why does depth matter? The Italian transmission network has a chain-like north–south topology with several island interconnections. A single message-passing step reaches only direct neighbours, which for interior nodes (like Central-southern Italy) means only two or three zones. At 3–4 layers, the receptive field spans the full peninsula, allowing the model to propagate signals from the northern industrial load centers to the southern renewable-rich edges. The non-monotonic dip at 5 layers (ΔNMAE = +0.12 pp, ns) might reflect mild over-smoothing: beyond a certain depth, repeated aggregation homogenizes node representations and dilutes edge-specific information.

Practical implication. The graph structure is a meaningful component of the model: removing it entirely is statistically comparable to using 2 or 5 layers, but significantly worse than the best configurations (3, 4, or 6 layers). Given that the marginal computational cost of additional layers is small relative to the LSTM encoder, operating at 3–6 layers is recommended. A depth of 3 layers offers the best trade-off between receptive field coverage and over-smoothing risk, while 6 layers yields the largest absolute improvement when combined with full hyperparameter tuning.

A.3. Experiment C: Which architectural design choices matter?

Do the non-standard architectural choices individually contribute to performance, or are some redundant?

Several design choices needs to be motivated : Pre-LayerNorm residuals, learnable residual gates, antisymmetric edge encoding, edge updates, and max pooling for decoder initialization. A sixth ablation removes the dedicated country-level encoder/decoder LSTMs, replacing them with the same graph-level LSTM used for internal nodes. Each was motivated theoretically in Chapter 4. This experiment ablates each choice in isolation to verify that the motivation translates into empirical gains.

A.3.1. Results

Table A.8 reports the overall test-set performance (averaged across all 16 edges and all 6 horizons) for the proposed model and each single-component ablation. All differences were assessed with one-sided paired Wilcoxon signed-rank tests (unit of observation: one metric value per edge×horizon cell, N = 96 for NMAE/DA and N = 90 for CF1), with Benjamini–Hochberg FDR correction applied across all 18 tests (6 variants × 3 metrics). Significance levels refer to adjusted p-values.

Variant	**NMAE ↓**	**DA ↑**	**CF1 ↑**
Proposed model	**52.1**%	89.8%	**60.2**%
Post-LayerNorm	54.9%	89.5%	58.8%
No residual gates ($g_n = g_e = 1$)	54.3%	89.8%	58.2%
Symmetric edge encoding	57.1%	89.0%	56.7%
No edge updates	55.9%	89.4%	58.1%
Mean pooling (decoder init)	57.1%	89.4%	59.3%
No special LSTMs for countries	53.9%	**89.9**%	59.0%
No encoder-decoder LSTMs (GNN only)	76.9%	83.8%	40.0%

Table A.8: Ablation of architectural design choices averaged over all 16 edges and all 6 forecast horizons. Each row removes or replaces exactly one component of the proposed model. See Table A.9 for full statistics.

Variant	**Metric**	N	**Median** Δ	W	p_{raw}	p_{adj}	
Post-LayerNorm	NMAE	96	+1.35	3894.0	<0.001	<0.001	***
	DA	96	−0.09	1461.0	0.0018	0.0025	**
	CF1	90	−0.68	1341.0	0.0022	0.0029	**
No residual gates	NMAE	96	+0.59	3688.0	<0.001	<0.001	***
	DA	96	−0.04	1988.5	0.2252	0.2384	
	CF1	90	−1.00	1117.0	<0.001	<0.001	***
Symmetric edge encoding	NMAE	96	+1.21	4343.0	<0.001	<0.001	***
	DA	96	−0.23	774.0	<0.001	<0.001	***
	CF1	90	−2.22	799.0	<0.001	<0.001	***
No edge updates	NMAE	96	+1.30	3849.0	<0.001	<0.001	***
	DA	96	−0.22	1057.0	<0.001	<0.001	***
	CF1	90	−1.54	1056.0	<0.001	<0.001	***
Mean pooling	NMAE	96	+0.71	3981.0	<0.001	<0.001	***
	DA	96	−0.13	1264.0	<0.001	<0.001	***
	CF1	90	−0.77	1402.0	0.0047	0.0056	**
No country LSTMs	NMAE	96	+0.58	3613.0	<0.001	<0.001	***
	DA	96	+0.02	2390.5	0.7244	0.7244	
	CF1	90	−0.72	1553.0	0.0233	0.0262	*
GNN only	NMAE	96	+23.237	4656.0	0.0000	0.0000	***
	DA	96	-4.763	212.0	0.0000	0.0000	***
	CF1	90	-15.657	6.0	0.0000	0.0000	***

Table A.9: Paired Wilcoxon signed-rank tests (one-sided) comparing each ablation variant against the proposed model. Unit of observation: one metric value per edge×horizon pair ($N = 96$ for NMAE/DA, $N = 90$ for CF1, edges without any congestion events excluded). Δ = variant − baseline (positive Δ for NMAE means degradation, negative Δ for DA/CF1 means degradation). p_{adj}: Benjamini–Hochberg FDR correction across all 18 tests. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

A.3.2. Analysis

Every ablated component causes a statistically significant degradation in at least NMAE, confirming that none of the six non-standard design choices is redundant. The magnitude and pattern of degradation differs across components, however, and each result is interpretable in light of the theoretical motivation.

Pre-LayerNorm. Switching to Post-LN increases NMAE by a median of +1.35 percentage points (pp) and degrades both DA (−0.09 pp, $p_{\text{adj}} = 0.0025$) and CF1 (−0.68 pp, $p_{\text{adj}} = 0.0029$). The result is consistent with the theoretical concern raised in Chapter 4: with Post-LN, the residual stream is normalised *after* the skip connection, which can cause the LSTM's hidden state contribution to be rescaled relative to the residual, effectively overwriting temporal memory accumulated across time steps. Pre-LN places the normalization *before* the sublayer, keeping the residual path unnormalized and preserving the LSTM's influence over successive layers.

Residual gates. Removing the learnable scalar gates ($g_n = g_e = 1$) raises NMAE by +0.59 pp (***) and drops CF1 by −1.00 pp (***), while DA is not significantly affected ($p_{\text{adj}} = 0.24$). The gate's primary role is therefore congestion-sensitive: when a edge is near capacity, the model benefits from being able to suppress or amplify the residual update selectively. The absence of a significant DA effect suggests that directional sign prediction is a coarser task that does not require fine-grained control of information flow between layers.

Antisymmetric edge encoding. This is the most damaging single ablation for CF1: replacing the antisymmetric encoding with a symmetric one ($[e_{ij}] = [e_{ji}]$) degrades NMAE by +1.21 pp, DA by −0.23 pp, and CF1 by −2.22 pp, all at $p_{\text{adj}} < 0.001$. As anticipated, DA is more affected here than in any other ablation: a symmetric encoder assigns identical representations to both flow directions on a given edge, so the model cannot distinguish whether flow is running from node i to node j or vice versa from the edge embedding alone. A per-horizon breakdown showed that the gap widens monotonically with horizon, reaching +7.8 pp NMAE at $h = 6$, consistent with the accumulated ambiguity across longer horizons.

Edge updates. Removing iterative edge embedding updates (hence reducing the GNN to node-only message passing with static edge conditioning) raises NMAE by +1.30 pp, DA by −0.22 pp, and CF1 by −1.54 pp, all at $p_{\text{adj}} < 0.001$. Because the prediction target lies directly on the edges, the ability to refine edge representations across GNN layers is particularly valuable: each pass allows the edge embedding to integrate neighborhood context that was not available at the previous layer.

Max pooling for decoder initialization. Replacing max pooling with mean pooling for the aggregation that seeds the decoder LSTM increases NMAE by +0.71 pp (***), reduces DA by −0.13 pp (***), and reduces CF1 by −0.77 pp (**). Max pooling is a natural choice when congestion is spatially concentrated: it selects the most activated node representation in the neighborhood, preserving the signal from the bottleneck node rather than diluting it in a mean. The congestion metric (CF1) being significantly affected supports this interpretation.

Country-level LSTMs. Removing the dedicated encoder/decoder LSTMs for country-aggregated nodes and replacing them with the same graph-level LSTM used for internal nodes degrades NMAE by +0.58 pp (***) and CF1 by −0.72 pp (*), but leaves DA unaffected ($p_{\text{adj}} = 0.72$). Country nodes aggregate flows across multiple internal edges, so a dedicated recurrent module allows the model to learn a separate temporal dynamics for this coarser spatial level. The limited impact on direction accuracy relative to magnitude and congestion metrics suggests that the country-level context is more important for quantitative accuracy than for directional classification.

GNN-only architecture (no encoder–decoder LSTMs). Removing all temporal encoder–decoder LSTMs and relying solely on a per-step GNN constitutes a qualitative change in the model architecture rather than a marginal ablation. As shown in Tables A.8 and A.9, this variant exhibits by far the largest degradation across all metrics (e.g. +23.2 pp NMAE, −4.8 pp DA, −15.7 pp CF1; all $p_{\text{adj}} < 0.001$), indicating that purely spatial message passing is insufficient for the task. Without recurrent components,

the model cannot accumulate or maintain temporal context across forecast steps, and each prediction depends only on instantaneous graph snapshots. The collapse in performance therefore highlights the central role of LSTM-based temporal processing in the proposed architecture: flow forecasting and congestion events are fundamentally temporal phenomena that cannot be captured by a static GNN alone.

Summary. All six core design choices contribute positively and significantly to at least one metric. Considering the encoder-decoder GNN, the two ablations with the broadest impact across all three metrics are antisymmetric edge encoding and edge updates, both yielding highly significant degradation (***) on NMAE, DA, and CF1 simultaneously. Pre-LayerNorm and max pooling show moderate but consistent improvements across all metrics. Residual gates and country LSTMs provide more targeted gains, primarily in magnitude accuracy (NMAE) and congestion detection (CF1), with no significant effect on directional accuracy.

A.4. Experiment D: Can the congestion signal be used as an intraday trading signal?

Does the predicted congestion ratio translate into actionable value for an intraday market participant?

The practical relevance of the model extends beyond forecasting accuracy. Formally, let $\hat{\mathbf{f}}_{ij}^t \in \mathbb{R}$ denote the net forecast flow on interconnection (i, j) at time t, and let C_{ij}^t be the corresponding day-ahead published capacity. We define the congestion indicator as:

$$r_{ij}^t = \frac{\hat{f}_{ij}^t}{C_{ij}^t} \tag{A.1}$$

An intraday trader who observes a predicted congestion ratio above a threshold θ could take a position on the price spread between the two zones: if congestion is predicted on edge $(i \to j)$, zone j is expected to trade at a premium relative to zone i (the exporting zone becomes the cheaper one once the line is saturated). This experiment evaluates the model's output as a trading signal, providing a domain-specific assessment of its practical value.

How the trade works. European intraday electricity markets allow trading the *same* delivery hour at multiple points in time before delivery. In Italy, the MI-A2 intraday auction closes at approximately 22:00 the day before delivery, while the XBID continuous platform remains open until ≈1 h before delivery. Both markets quote a price per zone for the *same* delivery product, the difference is purely temporal. MI-A2 is the earlier, less informed price, and XBID is the later, more informed price that acts as the closest proxy to physical realization.

Consider a concrete example.
At 12:00, the model predicts congestion on edge $(A \to B)$ for the 18:00 delivery slot, 6 h ahead. The market has not yet priced this in: XBID prices for 18:00 are nearly identical across zones ($P_A^{\text{MI-A2}} = 100\,€/\text{MWh}$, $P_B^{\text{MI-A2}} = 102\,€/\text{MWh}$, spread = +2 €/MWh).
Acting on the signal, the trader simultaneously:

1. **buys** 1 MWh in zone A for delivery at 18:00 at 100 €/MWh (the cheap, exporting zone), and
2. **sells** 1 MWh in zone B for delivery at 18:00 at 102 €/MWh (the expensive, importing zone).

The initial net cost of opening this spread position is $102 - 100 = 2\,€/\text{MWh}$. The trader then holds the position.

At 17:30, physical congestion materializes on the line as predicted.
Zone A has excess supply and its XBID price falls to 80 €/MWh, zone B is starved of power and its XBID price spikes to 150 €/MWh. The spread has widened from 2 €/MWh to 70 €/MWh. The trader closes both legs against the XBID last-traded price: the gross profit is $(150 - 80) - (102 - 100) = 70 - 2 = 68\,€$ per MW traded.

In general, the gross Profit and Loss (PnL) of one signal is the *change* in the zonal spread between the entry price and the XBID settlement, in the direction predicted by the model:

$$\mathrm{PnL}^{t}_{\mathrm{gross}} = \mathrm{sign}(\hat{f}^{t}_{ij}) \cdot \left[\left(P^{\mathrm{XBID}}_{j,t} - P^{\mathrm{XBID}}_{i,t} \right) - \left(P^{\mathrm{entry}}_{j,t} - P^{\mathrm{entry}}_{i,t} \right) \right]. \tag{A.2}$$

If the spread moves in the predicted direction ($\mathrm{PnL}_{\mathrm{gross}} > 0$), the trade is profitable, whereas if the market had already priced in the congestion at entry time, the spread does not widen further and the trade breaks even or loses. The XBID last-traded price thus serves as the *ground truth* for what the market ultimately believed about physical conditions at delivery. The sign of $\hat{f}^{t}_{ij}$ encodes the predicted flow direction, which determines the trade direction: the trader buys 1 MWh in the source zone (expected to be cheap) and sells 1 MWh in the target zone (expected to be expensive).

Entry price proxy. A key modelling assumption concerns the entry price P^{entry}. When the signal fires at 12:00 for an 18:00 delivery slot (6 h ahead), MI-A2 has already closed (it closes around 22:00 the day before delivery). However, XBID continuous trading is already open at 12:00, and MI-A2 auction prices are available from the *previous* day's session at the same delivery hour. In this simulation, the MI-A2 auction price is used as a proxy for the XBID continuous price at the time of signal firing. This is an optimistic assumption: it implicitly supposes that the XBID market has not moved significantly from the MI-A2 reference level at the moment of entry, so that the trader can be filled near the MI-A2 price in the continuous market. In practice, XBID prices evolve continuously and the actual fill price could differ, this assumption therefore represents a theoretical upper bound on performance. The bid-ask spread, not modelled here while non-zero in practice, is not expected to materially affect the conclusions for the purpose of this analysis.

This setup is a simplified simulation: it assumes a price-taking trader with no position size limit. Only Italian-zone edges are evaluated as XBID data for cross-border zones (Austria, Switzerland, France, Slovenia, Greece) are unavailable.

Signal construction. A binary trading signal $s^{t}_{ij} = 1$ is issued when the predicted flow magnitude relative to capacity exceeds the threshold:

$$s^{t}_{ij} = \mathbf{1}\left[\frac{|\hat{f}^{t}_{ij}|}{C^{t}_{ij}} > \theta \right]. \tag{A.3}$$

Metrics. Four metrics are reported for each model and threshold:

- **Hit rate (%).** Among all hours where the model issued a signal, what fraction actually experienced physical congestion (i.e. $|f|/C > \theta$)?
  A high hit rate means that when the model says the line will be congested, it usually is. However, a high hit rate alone does not guarantee profit: the market may have already priced in that congestion at MI-A2.
- **Direction accuracy (%).** Among all signal hours, what fraction did the spread move in the direction the model predicted (i.e. $\mathrm{PnL}_{\mathrm{gross}} > 0$)?
  A value above 50% means the model is right more often than a random coin flip. This is the most direct measure of whether the signal contains information not yet reflected in MI-A2 prices.
- **Average PnL (€/MWh).** The mean gross profit per trade, in euros per megawatt-hour, assuming a 1 MW position on every signal.
  A positive average PnL means the strategy is profitable on average. Because electricity spreads can be very large on some hours and near-zero on others, the average also captures the magnitude of the opportunity, not just its frequency.
- **Sharpe ratio.** The Sharpe ratio measures whether the average profit is large relative to the variability of individual trade outcomes.
  Intuitively, two strategies might both earn +1 €/MWh on average: one with very consistent small gains (high Sharpe), and one that alternates between large wins and large losses (low Sharpe).

A higher Sharpe ratio indicates a more reliable signal, one that generates steady returns rather than occasional lucky spikes. Values above 1.0 are generally considered acceptable and values above 3.0 indicate a very consistent edge over the evaluation period. Here, it is annualized using 365×24 hourly periods per year.

Horizon sensitivity. Because each model produces predictions at six different lead times ($h \in \{1, \ldots, 6\}$ h), but MI-A2 and XBID prices are defined per delivery hour, only one horizon is selected per delivery hour to avoid counting the same trade multiple times. Two horizons are reported:

- $h = 2$ h: the signal fires approximately 2 h before delivery. The market has already partially priced the congestion, so less residual information is available.
- $h = 6$ h: the signal fires 6 h before delivery, when the congestion might not have been yet reflected in intraday prices. More residual information may be available, particularly for models that can predict physical conditions at longer lead times.

A.4.1. Results

Tables A.10 and A.11 report results for both horizons at $\theta = 0.9$. Results at $\theta = 0.95$ are consistent and are omitted for brevity.

Model	Hit rate (%)	Dir. acc. (%)	Avg. PnL (€/MWh)	Sharpe
Short persistence	41.6	52.5	0.82	3.28
Long persistence	24.6	50.8	-0.35	-2.18
Scheduled flow	6.0	49.7	0.03	0.21
ARIMA	47.0	52.4	0.48	3.02
LSTM	61.3	51.3	-0.40	-2.45
XGBoost	66.5	53.4	0.26	1.56
Linear regression	65.4	54.4	0.66	3.92
Proposed model	74.0	52.6	0.59	4.11

Table A.10: Two-market arbitrage simulation at $\theta = 0.9$, horizon $h = 2$ h. A signal is issued when $|\hat{y}_{ij}|/C_{ij} > \theta$, the position is opened in MI-A2 (intraday auction, ≈2 h before delivery) and settled against the XBID last-traded price (≈1 h before delivery).

Model	Hit rate (%)	Dir. acc. (%)	Avg. PnL (€/MWh)	Sharpe
Short persistence	41.6	52.6	0.83	3.33
Long persistence	24.6	50.8	-0.36	-2.25
Scheduled flow	6.0	49.7	0.02	0.11
ARIMA	23.6	50.6	0.35	2.62
LSTM	46.5	50.4	-0.03	-0.15
XGBoost	49.4	52.4	0.32	1.91
Linear regression	48.5	52.5	0.27	1.57
Proposed model	59.0	52.6	0.84	5.52

Table A.11: Two-market arbitrage simulation at $\theta = 0.9$, horizon $h = 6$ h. A signal is issued when $|\hat{y}_{ij}|/C_{ij} > \theta$, the position is opened in MI-A2 (intraday auction, ≈2 h before delivery) and settled against the XBID last-traded price (≈1 h before delivery).

A.4.2. Analysis

Positive PnL. At $h = 2$ (Table A.10), the proposed model achieves a positive average PnL of +0.59 €/MWh and a Sharpe ratio of 4.11, which is competitive with the best baselines (Short persistence: +0.82, Sharpe 3.28, Linear regression: +0.66, Sharpe 3.92).
The hit rate of 74% is the highest of all models, confirming that the proposed model identifies the most reliably congested hours.

Improvement at longer horizon. Switching to $h = 6$ (Table A.11) reveals a clear advantage for the proposed model: its average PnL rises to +0.84 €/MWh and its Sharpe ratio increases to 5.52, making it the best-performing model at this horizon. In contrast, most baselines do not improve : ARIMA drops

from +0.48 to +0.35 and Linear regression from +0.66 to +0.27.
Short persistence is essentially flat (+0.82 → +0.83), which is expected because it relies on recent flow observations rather than on multi-hour-ahead physical forecasts.
This pattern is consistent with the interpretation that the proposed model captures structural physical information that is not yet reflected in intraday prices 6 h before delivery, whereas simpler baselines do not add predictive value beyond what the market already knows.

Limitations and outlook. This experiment makes several simplifying assumptions: price-taking behavior, no position limits, a single 1 MWh trade per signal, and entry at MI-A2 prices as a proxy for the XBID price at signal time. This last assumption is the most consequential: MI-A2 had already settled at the end of the previous day when the 6 h-ahead signal fires, and the XBID continuous price at 12:00 for an 18:00 delivery slot may differ from the MI-A2 reference. The results therefore represent an optimistic upper bound, and a more realistic evaluation would require tick-level XBID data at the exact time of signal generation.
A further simplification is that no limit-order logic is modelled: in practice, a trader would define a target entry price and wait for the XBID continuous market to reach that level before executing, rather than assuming immediate fill at the reference price.
Nevertheless, the consistent directional result, the proposed model improves with lead time while baselines do not, suggests that extending the model to longer forecast horizons (12 h, or even 24 h) could further increase the exploitable information advantage before the congestion is reflected in intraday prices.

B

Variability

B.1. Variability of baseline performance

Model	NMAE ↓	DA ↑	CF1 ↑
Short persistence	69.8 ± 43.8%	86.1 ± 7.2%	56.9 ± 23.1%
Long persistence	135.0 ± 120.1%	76.5 ± 18.9%	16.2 ± 19.5%
Scheduled flow*	204.6 ± 133.8%	86.4 ± 10.8%	37.6 ± 29.9%
ARIMA	61.2 ± 42.8%	88.3 ± 9.8%	59.5 ± 26.2%
LSTM	77.5 ± 75.9%	88.3 ± 8.4%	51.1 ± 26.9%
XGBoost	81.4 ± 90.4%	89.1 ± 8.2%	51.4 ± 26.8%
Linear regression	77.4 ± 96.8%	87.6 ± 10.2%	46.9 ± 28.2%
GNN	76.9 ± 49.8%	83.8 ± 12.3%	40.0 ± 33.6%
Proposed model	52.1 ± 41.7%	89.8 ± 8.0%	60.2 ± 25.2%

Table B.1: Comparison with baselines on the test set (2025). NMAE is expressed per edge as a percentage of the edge training mean flow, then averaged across all edges. DA and CF1 are expressed as percentages. **Bold**: best, underline: second-best. *Scheduled flow NMAE is high because the day-ahead schedule is not designed to minimize MAE but to ensure feasibility so it is included as an operational reference.

B.2. Variability of per-horizon performance

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	69.8 ± 45.0%	69.8 ± 45.0%	69.8 ± 45.0%	69.8 ± 45.0%	69.8 ± 45.0%	69.8 ± 45.0%
Long persistence	134.9 ± 123.3%	134.9 ± 123.3%	135.0 ± 123.4%	135.0 ± 123.4%	135.0 ± 123.4%	135.0 ± 123.4%
Scheduled flow*	204.6 ± 138.2%	204.6 ± 138.3%	204.6 ± 138.3%	204.6 ± 138.4%	204.6 ± 138.4%	204.6 ± 138.4%
ARIMA	28.1 ± 18.5%	46.4 ± 30.8%	59.5 ± 38.0%	69.7 ± 42.8%	78.2 ± 46.7%	85.2 ± 49.6%
LSTM	60.5 ± 68.5%	70.8 ± 73.5%	77.8 ± 77.1%	82.2 ± 79.4%	85.9 ± 82.5%	87.6 ± 82.5%
XGBoost	65.3 ± 89.2%	75.5 ± 90.6%	81.7 ± 91.8%	85.9 ± 93.1%	88.8 ± 94.1%	91.1 ± 95.4%
Linear regression	62.3 ± 102.7%	72.5 ± 100.4%	78.2 ± 99.0%	81.7 ± 98.1%	84.0 ± 97.4%	85.6 ± 97.0%
GNN	73.7 ± 48.8%	75.3 ± 50.0%	76.5 ± 50.8%	77.7 ± 51.8%	78.6 ± 52.3%	79.6 ± 53.1%
Proposed model	32.5 ± 27.9%	45.1 ± 35.9%	52.7 ± 40.6%	57.6 ± 44.1%	61.0 ± 47.0%	63.6 ± 49.1%

Table B.2: NMAE ↓ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	86.1 ± 7.4%	86.1 ± 7.4%	86.1 ± 7.4%	86.1 ± 7.4%	86.1 ± 7.4%	86.1 ± 7.4%
Long persistence	76.5 ± 19.4%	76.5 ± 19.4%	76.5 ± 19.4%	76.5 ± 19.4%	76.5 ± 19.4%	76.5 ± 19.4%
Scheduled flow*	86.4 ± 11.2%	86.4 ± 11.2%	86.4 ± 11.2%	86.4 ± 11.2%	86.4 ± 11.2%	86.4 ± 11.2%
ARIMA	94.9 ± 3.0%	91.7 ± 5.5%	89.0 ± 7.7%	86.7 ± 9.8%	84.7 ± 11.5%	83.0 ± 13.0%
LSTM	91.4 ± 7.0%	89.5 ± 7.8%	88.3 ± 8.2%	87.3 ± 8.9%	86.9 ± 9.0%	86.4 ± 9.3%
XGBoost	92.1 ± 8.5%	90.1 ± 8.4%	89.1 ± 8.0%	88.4 ± 7.9%	87.8 ± 8.1%	87.4 ± 8.5%
Linear regression	90.5 ± 10.6%	88.7 ± 10.4%	87.5 ± 10.3%	86.8 ± 10.2%	86.3 ± 10.3%	85.9 ± 10.4%
GNN	84.2 ± 12.5%	84.0 ± 12.5%	83.9 ± 12.5%	83.7 ± 12.6%	83.5 ± 12.7%	83.4 ± 12.7%
Proposed model	93.3 ± 5.8%	91.1 ± 7.2%	89.7 ± 8.1%	88.8 ± 8.5%	88.3 ± 8.7%	87.8 ± 8.9%

Table B.3: DA ↑ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

Model	h=1	h=2	h=3	h=4	h=5	h=6
Short persistence	56.9 ± 23.8%	56.9 ± 23.8%	56.9 ± 23.8%	56.9 ± 23.8%	56.9 ± 23.8%	56.9 ± 23.8%
Long persistence	16.2 ± 20.1%	16.2 ± 20.1%	16.2 ± 20.1%	16.2 ± 20.1%	16.2 ± 20.1%	16.2 ± 20.1%
Scheduled flow*	37.6 ± 31.1%	37.6 ± 31.1%	37.6 ± 31.1%	37.6 ± 31.1%	37.6 ± 31.1%	37.6 ± 31.1%
ARIMA	76.3 ± 15.4%	66.5 ± 20.6%	59.6 ± 24.5%	54.9 ± 27.2%	51.3 ± 29.1%	48.6 ± 30.8%
LSTM	63.9 ± 22.9%	55.6 ± 26.2%	50.4 ± 27.3%	47.2 ± 28.1%	45.1 ± 28.0%	44.4 ± 27.6%
XGBoost	70.8 ± 17.8%	57.9 ± 23.4%	50.5 ± 26.0%	45.5 ± 27.8%	42.7 ± 28.2%	41.4 ± 28.0%
Linear regression	62.6 ± 28.3%	52.2 ± 27.4%	46.1 ± 27.6%	42.6 ± 27.5%	39.8 ± 27.8%	38.2 ± 27.8%
GNN	40.9 ± 34.9%	40.5 ± 34.8%	40.1 ± 34.7%	39.8 ± 34.6%	39.5 ± 34.4%	39.4 ± 34.3%
Proposed model	73.4 ± 17.1%	64.2 ± 22.6%	58.6 ± 26.2%	56.3 ± 27.0%	55.1 ± 27.2%	53.4 ± 28.0%

Table B.4: CF1 ↑ (%) per horizon on the 2025 test set. **Bold** indicates best, underline second-best.

C

Use of Generative AI Tools

In accordance with TU Delft's guidelines on the appropriate use of generative AI tools in end projects, the following discloses the use of such tools in this thesis.

Tools used. Copilot was used throughout this project.

Writing assistance. Copilot was used to improve the clarity, grammar, and academic style of drafted text. All ideas, arguments, interpretations, and technical content are the author's own, AI assistance was limited to rephrasing and readability improvements on text already written by the author.

Code assistance. Copilot was used for debugging support and improving code readability. All core architectural decisions, model design, and experimental logic were conceived and implemented by the author.

Responsibility and verification. The author retains full intellectual responsibility for all content in this thesis. All AI-generated suggestions were critically reviewed, verified, and edited before inclusion. No sensitive, confidential, or proprietary data was entered into any generative AI tool.