EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering

Forecasting Grid System Imbalance: Case Study in Belgium

by

Tomás Urdiales

MSC THESIS

Assessment committee

Chair:	Dr. J.K. Kok
Member 1:	Dr. N. Paterakis
Member 2:	Dr. P.J. Hoes

Graduation

Program:	SET (SELECT)
Supervisor:	Dr. N. Paterakis
Date of defense:	September 04, 2023
Student ID:	1870823
Study load (ECTS):	60
Track:	Electrical Engineering

The research of this thesis has been carried out in collaboration with *Elia Group*.
This thesis is confidential for a period of 2 years until the publication date August 21, 2025.

This thesis has been realized in accordance with the regulations as stated in the TU/e Code of Scientific Conduct.

Disclaimer: the Department of Electrical Engineering of the Eindhoven University of Technology accepts no responsibility for the contents of MSc theses or practical training reports.

Forecasting Grid System Imbalance: Case Study in Belgium

Tomás Urdiales; TU Eindhoven, MSc Sustainable Energy Systems Engineering (SELECT, EIT InnoEnergy)

***Abstract*—European transmission system operators (TSOs) face the need to adapt to a rapidly evolving energy ecosystem as renewable energy and electrification gain traction, prompting the adoption of data-driven decision-making and advanced forecasting models. An essential aspect of this effort is understanding the nature of system imbalance, which directly reflects the net mismatch between electricity supply and demand, offering crucial insights into the network's condition. This study employs empirical analysis of the Belgian grid to ascertain and interpret system imbalance characteristics for optimal short-term predictions. Feature engineering, rigorous cross-validation, and custom linear and non-linear machine learning modeling techniques are combined to establish a comprehensive methodology. Results reveal rapid growth in Belgian system imbalance volumes over the past three years, with increasing extreme imbalance events. Relevant covariates are identified, including day-ahead and intraday cross-border nominations, key autoregressive features, and variables related to wind power, load, net regulation volumes, and ambient temperature. Collectively, the methodology developed reliably achieves a reduction in prediction error of 10-11% with respect to the Belgian TSO's current forecasting model, bringing the cross-validated prediction error down to just over 100MW on average. Emphasising interpretable linear models and non-sensitive, readily available data, this study provides a solid foundation for future expansion as the European grid continues to facilitate the energy transition in the coming years.**

I. Introduction

THE current energy landscape is characterised by growing concerns over human-caused climate change, security of national supply, and rising inflation rates driving much of the global economy towards economic uncertainty [1][2]. Such circumstances coincide with an era where renewable electricity generation has reached technological maturity and secured its place as a cost-competitive source of energy [3]. Consequently, the share of renewable energy production is growing rapidly across much of the developed world, introducing serious challenges for grid operators. With electrification remaining the likeliest path towards a decarbonised industrial society, these problems are only expected to increase in importance and severity [4].

Among the most pressing challenges of the on-going energy transition is the manner in which grid balancing needs to be addressed, i.e., how electricity consumption is set to match production at all times to ensure a constant electrical frequency and maintain stability. System imbalance (SI) is a direct representation of the net mismatch between the supply and demand of all electrical power injected and consumed through the grid. As renewables are integrated into the power generation mix and overall demand increases due to electrification, SI is expected to grow in volume and become more volatile in nature [5]. This accentuates the necessity for improved predictive models in grid balancing operations.

Increased interest in system imbalance forecasting is also being driven by European Union cooperation projects such as the Manually Activated Reserves Initiative (MARI). It is an international project developed by 29 European TSOs (transmission system operators) of the ENTSO-E group (European Network of Transmission System Operators for Electricity [6]) for the creation of a European platform for the exchange of manually activated balancing reserves (mFRR) [7]. The project aims to provide a framework for common technical, operational and market rules for a cross-border mFRR balancing market. Such a large market would in theory provide increased economic efficiency and security in the purchase and activation of balancing energy, while integrating most major European electricity markets and ensuring the financial neutrality of individual TSOs. Improved forecasting of system imbalance volumes is among the many technical requirements to enable such a large cooperation mechanism. At the moment, SI forecasts are only informative and not utilised directly in decision-making systems. However, an underlying intention of MARI is to use improved predictions to have market parties react beforehand in order to avoid high imbalance situations [7].

To address these issues, grid operators are developing advanced technologies to digitalise and automate much of the network's operation through data-based decision-making. A relevant tool among these is the use of advanced forecasting models that can provide accurate predictions. It is in this context that this study takes place.

A. Problem Statement

The main objective of this project is to research and implement an optimal short-term forecasting methodology for quarter-hourly system imbalance volumes in Belgium. The problem is thus one of generating numerical predictions for highly unpredictable phenomena. Balancing bidding markets in Belgium are organized quarter-hourly, so the main point of interest in SI forecasting is to predict, rather than the instantaneous signal, the 15-minute average of the SI. The present quarter-hour is labelled $qh{+}0$ and the upcoming $qh{+}1$. For example: if it is presently 12:06, SI_{qh+0} would be the average SI during 12:00 - 12:15; and SI_{qh+1} 12:15 - 12:30. The central goal is thus to predict both volumes (the mean) as accurately as possible. In terms of its physical dimensions, this quantity can be understood to represent measuring the total

Corresponding author e-mail address:
t.urdiales@student.tue.nl

imbalance energy [MWh] exchanged during a given quarter-hour and dividing by 0.25h to obtain an average power [MW].

According to the MARI initiative's technical specifications, dispatchers will need accurate SI predictions of qh+1 at minute 3 of qh+0 for the procurement of mFRR resources (i.e., they'll need an estimate of the average system imbalance volume during 12:15-12:30 by 12:03 as the latest). For model training and evaluation purposes, forecasts in this study are produced at every minute of the quarter-hour, so the minute-3 forecasts of interest to MARI are included.

B. Research Question

The principal research goal of this work is to describe the short-term forecastability of quarter-hourly grid system imbalance time-series signals, and to establish what are the most effective methodologies and covariates to do so. This is done by means of an empirical case study with the Belgian electricity grid. Additionally, initial stages of data analysis aim to ascertain trends and patterns in system imbalance, and analyse them in relation to the recent changes in the European energy ecosystem.

II. Preliminaries

To accurately frame the context of this study, a brief recapitulation of the fundamental physics of power systems is appropriate such that the importance of system imbalance can be appreciated in relation to the ongoing energy transition.

A. Physical Interpretation of System Imbalance

Synchrony is a property of alternating current (AC) networks such as the electricity grid [8]. In power systems, a synchronous rotating machine (generator) can be described by the ratio between the kinetic energy of its rotating mass [J] and its rated electrical power [VA], known as inertia constant:

$$H = \frac{\frac{1}{2}Jw^2}{S}$$

Where J is the moment of inertia [kg$\cdot m^2$] of the rotating mass, w its rotational velocity [rad/s] and S the electrical power [VA] of the machine. Employing this quantity and deriving from Newton's second law of motion, it is possible to summarise the system's behaviour most simply by the following dimensionless (per unit) expression:

$$2H\frac{dw}{dt} = P_m(t) - P_e(t) \qquad \text{[p.u.]} \quad (1)$$

Where $\frac{dw}{dt}$ is the derivative of the machine's rotational velocity with respect to time, $P_m(t)$ the active power generated from the mechanical torque impressed by a primer mover on the rotating mass, and $P_e(t)$ the active power demanded from the electrical torque produced by loads connected to the system.

Because generators in an AC network are electromechanically coupled, the relationship given in (1) not only models the rotational velocity of a single synchronous machine, but it also provides a direct analogous description of how the entire electricity grid behaves. $P_m(t)$ may be taken to represent an aggregate of the total amount of electrical power produced and injected into the grid at any given moment, $P_e(t)$ the equivalent for demand (power consumed by all loads), and $\frac{dw}{dt}$ the rate at which grid frequency deviates from its nominal value (50 Hz in Europe). If new loads are connected to the system and $P_e(t)$ becomes larger than $P_m(t)$, the system frequency starts to decrease. Conversely, if generators produce more power than is consumed, frequency rises. The speed at which this effect takes place is modulated by the value of the inertia constant.

System imbalance (SI), measured typically in megawatts [MW], is the most direct metric to reflect the dynamic relationship between electricity supply and demand, and thus of system stability. It can be defined conceptually in a succinct manner: the net mismatch between the supply and demand of all electrical power on the grid, $P_m(t) - P_e(t)$.

Normal operational practice within European power systems is to keep frequency deviations under ±1% of the nominal value. However, the standard range for Continental Europe is in fact 50 ± 0.05 Hz, deviations up to 50 mHz [9]. Frequency deviations of 100 - 200 mHz for prolonged periods of time can damage equipment, cause unforeseen malfunctions and disrupt industrial processes. Larger deviations trigger blackouts. Thus, a delicate balance between supply and demand (low system imbalance) must be enforced to maintain grid stability.

B. Renewable Power Generation

Traditionally, electricity grids have assumed the principles described above through large synchronous generators (high-inertia) with dispatchable production, i.e., controllable energy input. The paradigm is quickly changing as carbon-intensive coal power plants and nuclear power plants are disconnected and decommissioned, and a large amount of renewable power is installed [2].

Renewable energy sources, such as wind and solar power, draw energy from sources that are subject to weather conditions, and are therefore non-dispatchable producers that fluctuate significantly in output [4]. These fluctuations make it more difficult for grid operators to predict and manage the overall power balance on the grid.

In terms of inertia, renewables are also problematic for an AC grid. Photovoltaic panels are connected to the grid through an inverter, with no physical rotating mass. Even the largest wind turbines can't provide inertia on their own, as they are not connected to the grid directly and instead go through a frequency converter first [4]. While hydropower plants are dispatchable and capable of providing inertia, most generation potential in Europe is already being exploited, and very few new sites are under consideration [2].

Additionally, as the share of renewable energy sources in the electricity mix increases, the grid is becoming more decentralized. While in the past production was controlled by a few large power plants, these are rapidly being replaced by numerous independent solar and wind farms with much lower individual output [4], making it more complex to predict and coordinate the collective output of all the generating parties in the grid.

All in all, the energy transition is expected to bring about higher overall system imbalance volumes, as well as an increasingly volatile behaviour.

C. Quantifying System Imbalance

While system imbalance can be simply defined, in practice it is somewhat complex to quantify, especially so because countries adopt different definitions and legislative procedures in order to regulate grid balancing. In Europe, the task of grid balancing rests entirely upon TSOs, who manage the high voltage transmission lines and monitor power quality. The exact definitions used in this project are those of Belgium's TSO Elia [10], which closely resemble those in the rest of Europe. Needless to say, the underlying principles are common to all electromechanical power systems.

In its least granular mathematical definition:

$$SI = -NRV + ACE \qquad [\mathrm{MW}] \quad (2)$$

Where NRV stands for instantaneous Net Regulation Volume and ACE stands for Area Control Error.

In Belgium, ACE is equivalent to the 'frequency restoration control error' (FRCE), which essentially represents the mismatch between how much 'instantaneous' frequency support (exact definition is provided in the next subsection) is expected from balancing mechanisms and how much is actually delivered. It is measured for a given 'load-frequency control' (LFC) area over its alternating current interconnectors and modelled as:

$$ACE = \Delta P + K\Delta f$$

Where ΔP is the sum of power control error (the real-time difference between the measured actual real time power interchange P and the expected control program P_0) and $K\Delta f$ is the frequency control error: the product of the K-factor [MW/Hz] (representing the frequency response capacity in a given LFC block) and the frequency deviation [9].

It is clear from its definition that when system operation is running smoothly the ACE volume is fairly small. Nonetheless, one the largest threats to the otherwise remarkably stable synchronous grid of Continental Europe are inaccurate ACE measurements, as illustrated in events such as [11]. Under normal operation, however, the calculation of system imbalance is dominated by the negative NRV term in (2).

NRV is calculated as the difference, for each moment, between the sum of volumes of all upwards balancing regulations and the sum of all volumes of downward regulations. Mathematically:

$$NRV = GUV + GDV + SRV \qquad [\mathrm{MW}] \quad (3)$$

GUV stands for gross upward volume, GDV gross downward volume and SRV for the volume of strategic reserves activated. The latter is a contingency resource that is yet to actually be used. The former are defined as:

$$GUV = IMP_{IGCC} + \sum_{k=activated\,bids} \int aFRR_{upward}dt + \sum_{activated\,bids} \int mFRR_{upward}dt$$

$$GDV = EXP_{IGCC} + \sum_{k=activated\,bids} \int aFRR_{downward}dt + \sum_{activated\,bids} \int mFRR_{downward}dt$$

Where IMP_{IGCC} and EXP_{IGCC} stand respectively for the balancing volumes imported and exported trans-nationally. The Automatic Frequency Restoration Reserve (aFRR) term, or secondary balancing reserves, are systems that respond automatically to activation signals sent by the TSO. They are scheduled to enter the grid within 30 seconds of an imbalance. mFRR stands for Manual Frequency Restoration Reserve, the tertiary reserves. If frequency deviations are not sufficiently dampened by primary reserves and aFRR, mFRR is used (5 minutes after the start of an anomaly). These reserves are activated manually by the asset owner in response to signals sent out by the TSO.

D. Market Interpretation of System Imbalance

Imbalance may be interpreted as the result of all other forecast errors in the energy ecosystem. TSOs and market parties use a number of forecasts to provision supply. These range in timescale between years, days, and hours before that energy needs to be injected in the grid. The essential quantities to predict with a high degree of accuracy are load (for which climate conditions and human events need to be accounted) and renewable generation potential (primarily wind and solar power). If forecasts are accurate, the market can adapt beforehand and system imbalance is kept low (a smaller amount of balancing is needed because supply has been provisioned in advance to match supply). System imbalance therefore arises as a result of market failures in coordinating the matching of supply and demand, which is for the most part due to prediction errors.

III. State of the Art

A. In Published Literature

Not as much has been published for imbalance volume forecasting as for imbalance price forecasting, but there are nonetheless notable studies. Interestingly, a majority of these are applied to the Belgian grid, possibly due to the availability of grid data through Elia's platform.

An elaborate model for probabilistic forecasting of Area Control Error (ACE) was studied in [12], producing interpretable results and insights into the relevance of input features over time. Their model uses a custom neural network architecture that combines encoder-decoder, LSTM, and attention layers. Other studies show the potential in using RNNs and deep learning for SI forecasting [13]. In trying to forecast Belgian imbalance prices, a methodology was developed in [14] based on computing NRV state transition probabilities, effectively discretising and forecasting imbalance. Group Method of Data Handling neural networks were used in [15] to perform point forecasts on many of Elia's grid variables and reported improved forecast accuracy (compared to Elia's own in 2019), though it must be said that the published article does not present comparative results for system imbalance nor specify what forecast horizons are being used. One study compared different methodologies for forecasting net regulation volumes in the United Kingdom and found that an LSTM-based predictor outperformed gradient boosted trees and ARIMA models [16]. Lastly, a different approach altogether is to perform scenario generation and reduction instead of attempting to forecast the signal, as is the case in [17] for imbalance prices.

All in all, some literature has been published on the topic of system imbalance forecasting but for the most part it focuses on implementations of complex algorithmic approaches (predominantly neural network-based) rather than on forecasting performance. Most studies do not specify which covariates and autoregressive features they have found useful, nor present how their cross-validated forecasting results compare with other methods. Thus, it is difficult to establish what the state-of-the-art in SI forecasting performance is at the moment. Nonetheless, these articles provide a breadth of ideas on how to approach the problem, and highlight the importance of novel machine learning methods in time-series forecasting.

B. Elia's Forecasts

Elia Group's AI CoE developed a system imbalance forecasting model in 2021, and their results are updated and published in real-time [18]. They tested a variety of models but their presently deployed system is based on an autoregressive linear model using as covariates intra-day load forecasts and net regulation volume measurements. It incorporates both quarter-hourly and minute-wise values, and produces a forecast at every minute of $qh + 0$ for the quarter-hourly average of $qh + 0$ and $qh + 1$ (separately). Elia's model has a cross-validated $qh{+}1$ forecast error (methodology explained in detail in the next section) of: ∼112MW mean-absolute-error and ∼149MW root-mean-square-error. This currently deployed system is used in this work as the main reference to benchmark results against.

IV. Methodology

A. Statistical Analysis & Testing

Initial stages of exploratory data analysis make use of standard time-series statistical methods. The primary goal is to extract trend and seasonality patterns in the target time-series, measure autocorrelation and cross-correlations in the data, and identify possible data quality issues. Statistical testing also plays an important role during the training and evaluation stages to guide the modelling approach and check the significance of results.

The autocorrelation function is used extensively to measure the linear relationship between lagged values of a time-series. It serves primarily to identify initial signs of trend and seasonality in the data, as well as indicate which type of statistical modelling approach to use [19]. Crucially, it is also employed to test residuals (error on training/fitted data) in order to identify dynamics in the data that have not been properly captured by a model. If required, autocorrelations are formally tested using a portmanteau test to check that they are significantly different from what would be expected from a white noise process [19]. The test used in this study is the Ljung-Box test [20]. The partial autocorrelation function is used to measure the partial correlation of a time-series with respect to its lagged values, i.e., the relationship between two lags after removing the effects of all previous lags in-between [19]. It serves a critical purpose in identifying relevant autoregressive features for all models tested here.

Time-series decomposition was employed sparsely throughout the project due to the highly stochastic nature of the target signal. The algorithms tested all failed to extract any seasonality or trend. Nonetheless, modern versions of the STL algorithm [21] were used to analyse some covariates with more pronounced patterns.

The Augmented Dickey-Fuller test [22] is the unit root test of choice in this work to determine stationarity in a time-series. Lastly, the Diebold-Mariano test [23] is used to check that the difference between two forecasts is statistically significant and determine which model produces superior results, in conjunction with the relevant error metrics.

Other standard visual and statistical techniques are manually adapted to the data when off-the-shelf algorithms fail. Notably, the use of custom moving average smoothing and differencing operations.

B. Time Reference & Forecasting Target

As previously mentioned, the main forecast of interest is that of SI_{qh+1} during $qh{+}0$. Since SI_{qh+0}, the 15-minute average of the SI during the present quarter-hour, is unknown at the time of prediction, the problem is thus of forecasting with a time horizon $h = 2$. This is not strictly true because some minute-wise telemetry data is available (about $qh + 0$) during $qh + 0$, and it is integrated into the forecasting pipeline. The resulting data matrix therefore has minute-wise granularity, but the forecasting target is quarter-hourly. To summarise: the forecasting models produced in this work combine quarter-hourly and minute-wise data to forecast SI_{qh+1} during all minutes of $qh + 0$: as new telemetry data comes in every minute (0-14) a new forecast is produced for SI_{qh+1}. The latest forecast is produced at minute 14 of $qh + 0$. Naturally, the forecast error is expected to decrease as the quarter-hour progresses.

C. Feature Engineering

All models developed in this study are autoregressive and most include a large number of covariates, which are also often lagged. Some rolling average features are also produced from some signals when useful. Dummy variables are used to encode some datetime features, most notably the minute of quarter-hour. The data pipeline therefore integrates many different sources with different time-granularity, to produce data matrices that often have more than 50 features in total.

Feature selection is done either manually, using the Boruta algorithm [24], or through Automatic Feature Selection (AFS) using different decision tree ensemble methods. Ultimately, the criteria to determine whether a lag or feature is utilised is if it provides any additional predictive power (decreases test-set error metrics on average).

All time-series are datetime indexed, and localized to Central European Time (CET). This is done for data consistency reasons (to deal with problematic daylight saving time entries), and because system imbalance dynamics are assumed to depend more on human time conventions (the hour of day) than on natural phenomena (UTC time).

Feature scaling is an important element in the forecasting pipeline and the scaler used here is the zero-mean-unit-variance transformation:

$$x' = \frac{x - \overline{x}}{\sigma}$$

Where $\overline{x}$ is the in-sample mean, and σ is the in-sample variance. Other scalers are sometimes used for robustness to outliers, mainly the quantile and robust transformers.

D. Modelling

This study explores a variety of modelling approaches based on both traditional statistical methods and more modern machine learning regression methods.

Since the project builds on previous work by Elia engineers, many classical statistical models like ETS (exponential smoothing) or ARIMA (autoregressive integrated moving average) have not been given much emphasis. The Elia team experimented with these methods when they built their currently deployed system, and found lesser success than with custom linear models. Additionally, these off-the-shelf algorithms are often difficult to customise to the particular nature of this problem (both quarter-hourly and minute-wise data, many covariates, etc), whereas linear regression models offer much more flexibility. It is important to note that complex non-linear models also failed to match the performance of the linear model when Elia first built the current system. This is likely due to the more generalisable nature of linear regression models when compared with more complex models, which tend to overfit (memorise the data) in highly stochastic, challenging regression problems of this kind [19].

Following their original findings, most research and development starts from an autoregressive linear model like Elia's. The majority of the work looks to enhance this model with extensive experimentation in covariate research and feature engineering; to form a consistent basis with which to then explore more complex modelling approaches.

Other linear models are used here for varying purposes. Lasso regression is sometimes used to analyse feature importance, and Ridge regression to prevent large coefficients that can lead to numerical instability in the face of data quality issues. Other notable algorithms used are Huber regression and linear SVR (non-kernelised support vector regression) [25][26].

A variety of non-linear methods also play an important role in this work. Most notably: decision tree ensemble methods such as Random Forests [27], Gradient Boosting (XGBoost) [28] and Histogram-based Gradient Boosting (LightGBM) [29]; kernelised Support Vector Regression [26]; and artificial neural network models. Only the classical feed-forward Multilayer Perceptron architecture [30] is used here, for reasons explained in-depth in the Discussion section.

Probabilistic (distributional) forecasts are explored in this project; however, their primary relevance lies in business operations, such as dimensioning balancing reserves, rather than in addressing the central research question of this study. Nonetheless, the primary objective of identifying the optimal methodology to minimise errors in forecasting the mean of SI also translates to optimal distributional forecasts in nearly all instances. Therefore, the results presented here also pertain to the production of probabilistic forecasts.

Lastly, ensembles are built combining multiple forecasts of different origin with the aim of reducing overall variance and improving predictive performance [31][32]. This is done either according to manual model selection and arithmetic means, or by simple stacking methods to produce a weighted sum of forecasts.

E. Time-Series Cross-Validation: Backtesting Criteria

Time-series cross-validation is a key element in evaluating system imbalance forecasting systems due to the highly time-dependent nature of SI. For the same model, certain times of the year show much reduced forecast errors compared to others. The data is sometimes easier to predict due to lower occurrences of unpredictable events such as outages and climate anomalies. This may result in non-representative error metrics when the testing window is not long enough.

This study addresses this issue with a custom backtesting procedure that trains many identical models at different points in time and stitches together their resulting forecasts to generate predictions for the entire year of 2022. To do this, data usage for training is restricted to 2021 and 2022 historical records. This time window is chosen somewhat arbitrarily according to data availability/quality and to avoid any residual effects on system imbalance dynamics from the 2020 COVID-19 pandemic. This approach does allow for complex non-linear modelling, although it significantly increases computational costs.

All models are therefore compared following this methodology, where, importantly, error metrics are computed on the entire 2022 predicted time-series rather than on short individual forecasts. This is a key issue with common non-linear metrics such as root mean square error (RMSE), where looking at the individual error of each split leads to distorted evaluation (more optimistic error metrics).

Figure 1 illustrates how train-test time-splits are set up. In practice, more splits and longer training windows are used than shown on the figure. The number of splits varies from around 90 to a minimum of 13, depending on the computational cost of training. Some models benefit from much longer training windows, so in some configurations data is used as far back as the beginning of 2021. In any case, the configuration is always carefully set up to produce predictions for all of 2022 without any data leakage.

F. Evaluation of Forecast Errors

The evaluation metrics chosen to assess the quality of a model's forecasts are:

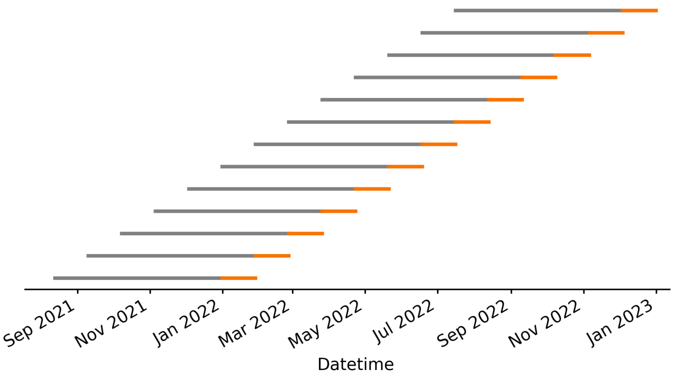

Fig. 1. Train-test time-split configuration for cross-validation. Each horizontal line represents a model, where the gray part shows the length of the training window and the orange the predictions produced.

- Mean absolute error (MAE). Represents the mean forecast error across all samples. Optimising for minimum MAE leads to forecasts of the median of the distribution.

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

- Root mean squared error (RMSE). Represents the mean of squares of the error, accentuating the effect of individual high error points on the forecast evaluation. Minimising RMSE yields forecasts of the mean, and most optimization algorithms in this study (for point forecasts) take this approach. It is used alongside MAE to emphasize the effect of extreme imbalance scenarios on forecast quality. If these events are filtered out from the dataset, the RMSE approaches the MAE, with overall lower error metrics for both.

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

- Mean absolute scaled error (MASE). Represents the relative MAE of a forecast with respect to the *in-sample* (training data) error of the naïve method. The naïve predictor, or persistence model, simply takes the last observation as the forecast for future values, assuming no change between the present and the future.
  This scaled error metric has a meaningful (and invariant) scale: comparing how well a forecast does with how well the simplest method would; is symmetric (penalises positive and negative errors equally); and is immune to the numerical problems faced by metrics such as MAPE (mean absolute percentage error) when any step in the time-series has a value of zero [33].

$$MASE = \frac{MAE}{MAE_{naive,\, in-sample}}$$

- For distributional (interval) forecasts, probabilistic evaluation methods are needed. This study uses the Quantile Loss and the Winkler Score [34]. Mathematically, for a prediction interval of $100(1-\alpha)\%$:

$$L_{p,t} = max[\, p(y_t - q_{p,t}), (p-1)(y_t - q_{p,t})\,]$$

$$W_{\alpha,t} = \begin{cases} (u_{\alpha,t} - l_{\alpha,t}) + \frac{2}{\alpha}(l_{\alpha,t} - y_t) & y_t < l_{\alpha,t} \\ (u_{\alpha,t} - l_{\alpha,t}) & l_{\alpha,t} \le y_t \le u_{\alpha,t} \\ (u_{\alpha,t} - l_{\alpha,t}) + \frac{2}{\alpha}(y_t - u_{\alpha,t}) & y_t > u_{\alpha,t} \end{cases}$$

$$W_{\alpha,t} = (L_{\alpha/2,t} + L_{1-\alpha/2,t})/\alpha$$

Where p is the probability of the quantile $q_{p,t}$ being predicted; and $l_{\alpha,t}$ and $u_{\alpha,t}$ are the lower and upper bounds of the prediction interval (interval width).

G. Data & Code Availability

The majority of the data utilised throughout this project pertains either to Belgium or to neighbouring countries with which cross-border exchanges occur. Almost all of the data is time-series-based and limited manually to the period of 2021-2022, providing a comprehensive two-year dataset for data analytics and forecasting.

Though not all yield predictive power, some relevant quarter-hourly variables used are: net regulation volumes, total load and generation, wind and solar generation, cross-border nominations, upward and downward regulation volumes, imbalance prices and ambient temperature. Additionally, external forecasts are used as covariate predictors of system imbalance: day-ahead load forecasts, day-ahead and intra-day cross-border nominations, and day-ahead renewable generation. The interval width and error of these are also tested as covariates.

And the minute-wise variables used: net regulation volumes, marginal incremental and decremental prices, and imbalance prices.

Quarter-hourly historical records (metering values) are validated by the TSO and used for clearing balancing market bids, so it is this data that is always used for analytics.

Minute-wise data usually comes from telemetry (substation SCADA systems) and is available much closer to real-time. To follow real-life practical constraints, the target time-series to predict in this study is in fact the minute-wise reading of SI_{qh+1} at minute 14, i.e., the last measurement made by automatic systems of the average system imbalance in that quarter-hour. Minute-14 readings are very close to the official validated quarter-hourly records, but in some circumstances differ due to system error or manual adjustments. Henceforth, this target time-series is referred to as 15-minute cumulative system imbalance.

Additionally, the latest minute-wise observations assumed to be accessible to a model when predicting in real-time are those two minutes before the present minute. In practice, some data is obtained as quickly as minute-minus-one, but building a model dependent on this availability comes at a great cost of reliability and complicated data engineering.

Besides rare occurrences of missing data, no significant data quality issues that could affect the quality of results have been identified.

All data used for this study is non-personal and non-sensitive under the General Data Protection Regulation (GDPR) [35]. It can be publicly accessed through Elia's

OpenData portal or through their public APIs [36]. All code developed for this study is made free and open-source [37].

V. Results

A. Analytics

This section presents relevant descriptive analytics of the system imbalance signal. All results are calculated on the entire dataset of validated quarter-hourly historical records going from September 2019 to January 2023. The numbers are nearly identical when using the aforementioned minute-14 readings, and the resulting conclusions remain the same.

The evolution of system imbalance during 2022 is shown on Figure 2. It outlines the highly stochastic, non-trended and non-seasonal nature of imbalance. It primarily oscillates in a range between -300 and 300 MW, though outliers are common and often extreme, reaching values of more than 1000MW. The causes behind these phenomena can be speculated upon but are difficult to pinpoint, though many peaks can be traced back in the data to forced outages or unforeseen drops in generation. The signal's distribution is fairly close to a normal distribution, although the extreme values are over-represented when compared with a Gaussian in a quantile-quantile plot.

Taking descriptive statistics on the entire signal, the mean is around -16.5MW, the median -10.6MW, and the interquartile range [-109.3, 79.7] MW. This indicates that most of the time system imbalance is negative, i.e., there is higher demand than supply, suggesting that the TSO is using primarily upward regulation to balance the grid. This result is consistent with frequency readings that show a historical mean below 50Hz. The minimum and maximum points in the series are -1419.7 and 1257.1 MW respectively. The overall standard deviation is 170.6MW.

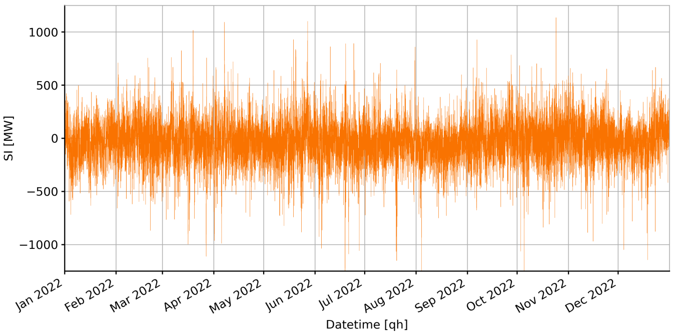

Fig. 2. System imbalance during 2022.

Running an Augmented Dickey-Fuller test [22] on this signal yields a negative ADF statistic (-25.9) with a p-value of zero, strongly rejecting the null hypothesis that the time-series is non-stationary. As Fig. 2 depicts, the signal closely resembles white noise, and it is highly stationary.

While statistical tests and STL decomposition algorithms fail to extract any trend or seasonality in the signal, it is possible to identify patterns manually. Figure 3 makes use of long moving average smoothing windows to display two trends in the evolution of imbalance over the past few years. The graph on the left shows significant growth in the net imbalance monthly volumes: the total imbalance energy, positive or negative, measured each month in GWh. A measurable increase of around 25% (+20GWh) has occurred over the last three years. The graph on the right plots the standard deviation of the SI signal over every week. It shows that the distribution is becoming more spread out, i.e., occurrences of extreme imbalance are becoming more common. It is important to note that neither y-axis in Fig. 3 is truncated; both axes start at zero. This accentuates how worrisome these trends are considering the relatively short time span over which they are measured.

These results are likely attributable to the the recent integration of solar and wind power generation capacity. This increase can be appreciated directly on the Elia datasets used in this study, and despite its high seasonality, the data for these two renewable sources clearly displays a sharp increase in peak power over the past few years (especially so for solar). While there are important electrification tendencies in urban transport and heating, load is not seen to increase from year to year (nor is its standard deviation), implying that the primary cause for higher imbalance is the presence of renewable generation.

A direct relationship between renewable generation and forecast error (in load, wind, solar, etc.) has been studied in [38]. In the case of wind power, for every additional GW generated the increase in forecast error is as sharp as $\sim$100MW. This phenomenon, combined with the results presented here, supports the interpretation that system imbalance is the result of market failure to predict and adapt to variable generation and demand.

The trends depicted in Fig. 3 span such a significant time period that they lack any utility for forecasting purposes. Nevertheless, the data does exhibit other discernible patterns that can be utilised in forecasting.

Figure 4 displays the average hourly profile (median and interquartile range) of system imbalance. There are some dynamics to be appreciated: during the ”quieter” hours of the night, between 2AM and 5AM, high imbalance scenarios are rare, possibly due to low demand and the absence of solar generation. The distribution spreads out during high demand hours, most notably at the 17PM demand peak where the median is heavily displaced towards the negative. Calculating net imbalance hourly volumes also shows higher imbalance during busy hours (10AM-12AM, 17PM) than during the period between 2AM and 5AM, with the difference being as large as 50% ($\sim$50GWh). These results are indicative of weak daily seasonality that can be exploited in modelling.

Figure 5 shows the autocorrelation function plot of quarter-hourly SI. Despite high stochasticity, the signal shows strong correlation with its immediately previous values, although it rapidly decreases after about five quarter-hours. It is important to note some correlation exists with respect to lag-96, the previous day. This is consistent with the results shown in Fig. 4. This weak correlation extends past one week of lags, hinting towards some level of additional weekly seasonality.

Figure 6 displays the partial autocorrelation of the SI signal, indicating which lags hold predictive power in autoregressive models. The first five lags again have the most significant partial autocorrelation, with those around lag-96 also showing

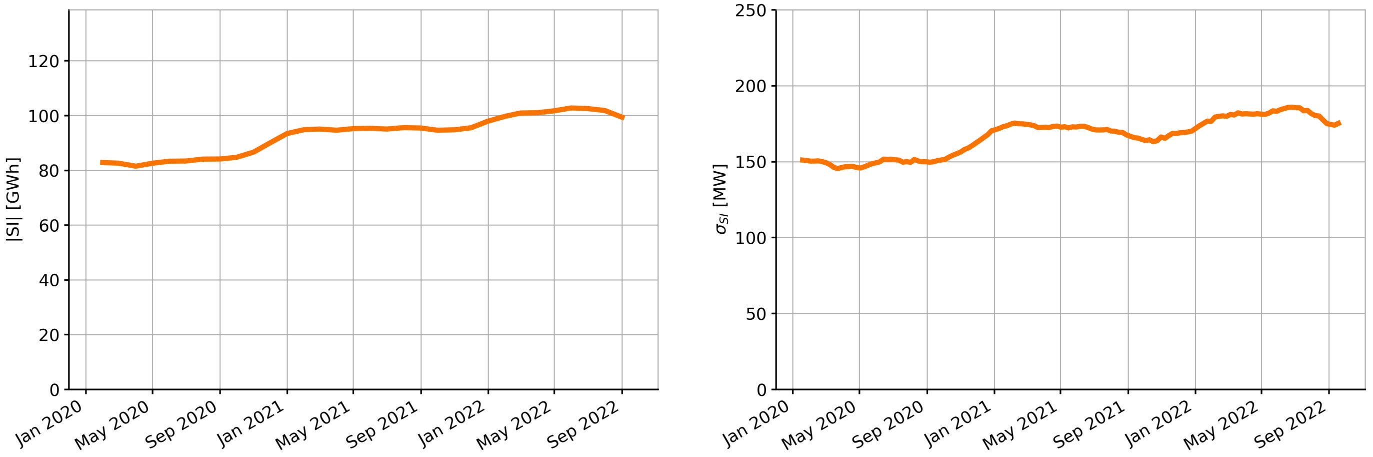

Fig. 3. Left: evolution of monthly SI volumes, 10-month centered moving average (units in GWh). Right: evolution of SI weekly standard deviation, 40-week centered moving average (units in MW). Note that y-axes are not truncated on either figure.

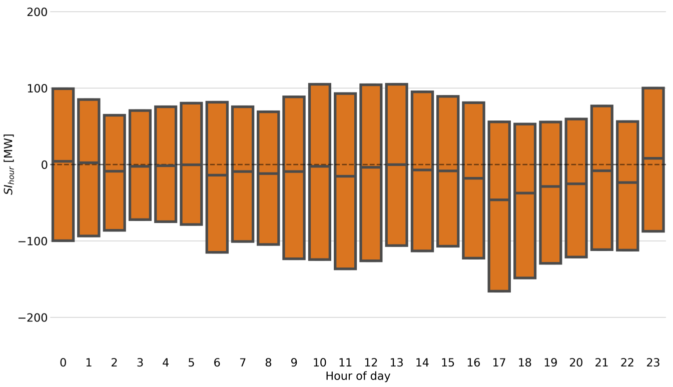

Fig. 4. Hourly profile of system imbalance. The extension of the vertical intervals corresponds to the interquartile range at every hour, and the continuous black line in the middle to the median. Units are in MW.

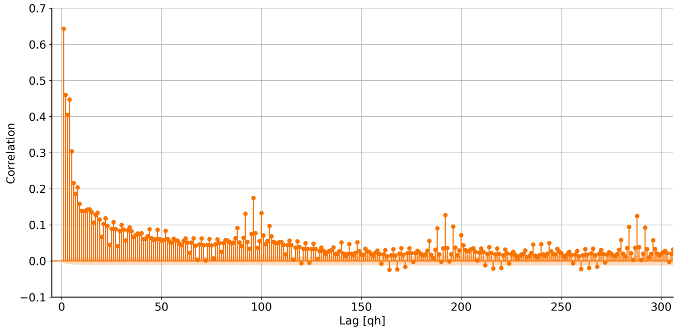

Fig. 5. Autocorrelation function plot of system imbalance, with lags reaching up to three days in the past. Lag-0 is omitted.

noticeable levels. This weak daily partial autocorrelation also extends past one week of lags, although it decreases more rapidly than in the ACF. After the first day (lag-96), correlation values are still above the confidence interval but below 0.1.

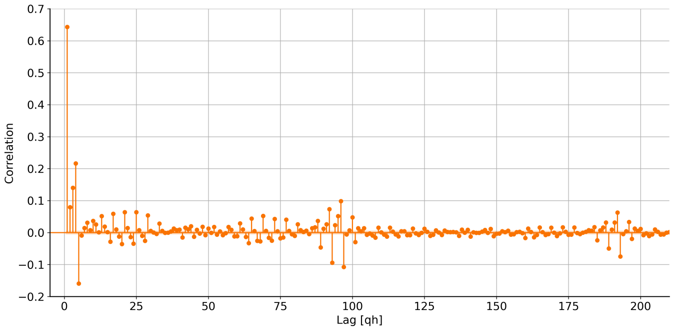

Fig. 6. Partial autocorrelation function plot of system imbalance, with lags reaching up to two days in the past. Lag-0 is omitted.

B. Covariate Relevance

1) Autoregressive Features: A strictly autoregressive linear model with no exogenous variables can be created to match and slightly outperform the present Elia system (by up to about 4%). This model offers the benefits of simplicity, minimal computational cost, and resilience, as it relies solely on the SI time-series as input.

Judging primarily by performance improvement and their weight in linear regression models, the most important AR lags appear to be: the last available minute-wise reading (two minutes prior to prediction time), the quarter-hourly value for the previous hour, the quarter-hourly value for the previous day, the quarter-hourly value for the previous week, and the values from one quarter hour prior to the 1-day and 1-week lags. These seasonal past lags provide more predictive power than the quarter-hourly lags immediately prior to forecasting time (qh-1, qh-2, etc.), indicating a lack of predictable intra-hour dynamics in the SI signal.

Automatic feature selection can be used to detect additional lags that yield marginal improvement (up to an apparent ceiling of 4% over Elia's model). These often lack interpretability but do on average improve predictions, a phenomenon also observed for some exogenous variables.

The following pseudo-code notation is used to show the exact covariate configuration, where lag +1 of quarter-hourly 15-minute cumulative SI is the target (qh+1), qh+0 is the present quarter hour at the time of prediction, and negative lags are past measurements relative to qh+0 (i.e., lag-95 corresponds to the value of the previous day's qh+1).

```
qh_parameters = {
   "system_imbalance_cum15": {
      "lags": [-1, -2, -3, -4, -5,
      -95, -94, -93, -96,
      -191, -190, -189, -192,
      -287, -286, -285, -288,
      -383, -382, -381, -384,
      -479, -478, -477, -480,
      -575, -574, -573, -576,
      -671, -670, -669, -672,
      -17, -21, -23]},
}

minute_parameters = {
   "system_imbalance": {
      "lags": [-2, -18, -33, -63, -64, -181]},
}
```

2) Exogenous variables:
A comprehensive set of external variables and their lags has been tested, rejecting the majority but finding significant predictive power in some. Most notably, day-ahead and intra-day cross-border nominations can be combined into a grand-total cross-border nominations signal to provide a powerful covariate ("xb_grand_total"). The current granularity for the cross-border market is hourly, and allowing models access to the nominations from both the next and previous hour yields improved predictions in all scenarios. Interestingly, the greatest improvement (especially for linear models) is attributable to having both qh+0 and qh+1 cross-border nominations as input features. Since the data is hourly, these two features differ only when the hour changes, which seemingly provides significant predictive information. The coefficients for these two features in linear models are of opposite sign and high absolute values, sometimes warranting the need for L2 regularisation to avoid numerical instability.

The aforementioned autoregressive and cross-border nomination features provide the greatest prediction improvements, and the following are added for the best overall set:

- Recent measurements of net wind power generation. Interestingly, intra-day forecasts do no provide any predictive value, nor does any solar generation data.
- Recent measurements of total load, as well as intra-day forecasts for the next hour.
- Net regulation volume (NRV) measurements of the last hour, as well as the latest available minute-wise reading.
- Ambient temperature changes with respect to the coming hour. This covariate does add marginal improvement, but depends on the assumption that hourly changes in temperature can be well predicted. The Elia datasets available at the time only included measurements, so perfect forecasts needed to be assumed in order to test for any relationship with SI.
- Time features encoding the minute of the quarter-hour (0-14) through dummy variables improve larger non-linear models' performance. No additional temporal dependencies have been identified (including human-caused events like weekends and holidays).

The complete best overall feature set appears to be:

```
qh_parameters = {
   "system_imbalance_cum15": {
      "lags": [-1, -2, -3, -4, -5,
      -95, -94, -93, -96,
      -191, -190, -189, -192,
      -287, -286, -285, -288,
      -383, -382, -381, -384,
      -479, -478, -477, -480,
      -575, -574, -573, -576,
      -671, -670, -669, -672,
      -17, -21, -23]},
   "xb_grand_total": {
      "lags": [4, 1, 0, -4]},
   "wind_rt_mw": {
      "lags": [-1, -3]},
   "nrv_rt": {
      "lags": [-1, -2, -3, -4]},
   "load_rt_mw": {
      "lags": [-3]},
   "total_load_last_mw": {
      "lags": [3, -3]},
   "temperature_diff": {
   "lags": [4, 0]},
}

minute_parameters = {
   "system_imbalance": {
      "lags": [-2, -18, -33, -63, -64, -181]},
   "net_regulation_volume": {
      "lags": [-2]},
}
```

It is important to note that largely improved predictions can be obtained using only a subset of around 25 out of the 49 features of this complete set. While many of these features can be considered redundant, they allow for a final 1-2% enhancement over Elia's predictions. Similarly, introducing additional lags of these covariates is likely to marginally improve forecasts further, at the cost of increased computational cost.

C. Forecasting Performance & Benchmarks

Table I presents the best cross-validated error metrics obtained with the highest performing models tested. Percentages show improvement with respect to Elia's model. The naïve estimator used here for comparison is the best performing naïve method found: taking the last available minute-wise reading as the prediction of SI_{qh+1}. Unless mentioned explicitly, all of the following model results make use of the complete covariate set shown previously. Non-linear models are also given access to additional dummy variables encoding the minute of the quarter hour.

The train-test configuration varies depending on the model. Linear regression models are trained on 30 weeks of data, and cross-validated in 91 different splits with a testing window of 4 days each, which amounts to a total 364 days of test predictions (multiple of 7 makes some calculations easier) for 2022.

Non-linear models are trained on 50-70 weeks of data, and cross-validated in 13 or 26 splits, with 28 or 14 test days each. The training windows are chosen according to optimal performance. The ratio of splits to testing days is chosen primarily based on computational cost (more splits implies more models need to be trained). It has not been observed that the length of each testing window affects forecast accuracy significantly.

Given the precision of input data, all results shown here are calculated in 32-bit floating-point arithmetic (single-precision), with numerical tolerances of either 10^{-3} or 10^{-4} for training the models.

TABLE I
CROSS-VALIDATED ERROR METRICS OBTAINED WITH THE BEST PERFORMING MODELS FOUND. PERCENTAGES SHOW IMPROVEMENT WITH RESPECT TO ELIA'S MODEL

Model	**MAE [MW]**	**RMSE [MW]**	**MASE**
Elia's linear regression model, using their original covariate set	111.98	148.97	0.769
Naïve estimator	145.55 (-30.0%)	191.93 (-28.8%)	1.000
AR linear regression model, using no exogenous variables	107.34 (4.1%)	143.25 (3.8%)	0.737
Ridge linear regression model with L2 regularisation ($\lambda = 0.5$)	102.50 (8.5%)	135.92 (8.8%)	0.704
Multi-layer Perceptron neural network regressor of 2 layers of 4 neurons each, trained through the ADAM optimiser, with light L2 regularisation ($\lambda = 0.1$), ReLU activation function, and adaptive learning rate starting at 0.001	101.80 (9.1%)	135.19 (9.3%)	0.699
Histogram-based gradient boosting regression tree model with 125 maximum trees/iterations, squared-error loss function, learning rate of 0.1, and unconstrained tree depth	105.17 (6.1%)	139.71 (6.2%)	0.723
15, minute-specific, Ridge linear regression models; with L2 regularisation ($\lambda = 0.5$)	101.43 (9.4%)	134.50 (9.7%)	0.697
Ensemble of forecasts by arithmetic mean	100.68 (10.1%)	133.64 (10.3%)	0.692
Ensemble of forecasts by stacking through linear regression	100.27 (10.5%)	133.02 (10.7%)	0.689

Comparison with the naïve estimator indicates that all models produced here do possess predictive power, outperforming naïve predictions by 20-30%; a positive outcome considering the highly stochastic nature of system imbalance.

A significant improvement with respect to Elia's current model is found with simple linear regression models due to the improved covariate set obtained through feature engineering. More complex non-linear models have only been able to match or very slightly outperform the linear ones. The best "single" model is built by training and combining 15 linear regression models, one for each minute of the quarter hour. This allows for the calibration of the predictive process depending on forecast horizon, in a similar manner as non-linear models would be able to do with access to dummy variables encoding the minute in the quarter hour. Looking at error per minute-of-quarter-hour curves shows that errors decrease in a quasi-exponential manner as the quarter hour progresses and the forecast horizon decreases, as would be expected. Creating ensembles of the best forecasts obtained by single models generally reduces overall variance [32]. Additionally, the combination of different training windows is theoretically advantageous in allowing for emphasis on somewhat different temporal dynamics. An ensemble of the two best performing models produces the best results overall (what is shown on Table I), achieving a reduction of 10-11% in error. A snapshot of the predictions produced with this ensemble is shown in Fig. 7 next to actual observations and the predictions from Elia's model. It can be visually appreciated that the new system is sometimes able to better track the real signal at times when imbalance fluctuates rapidly.

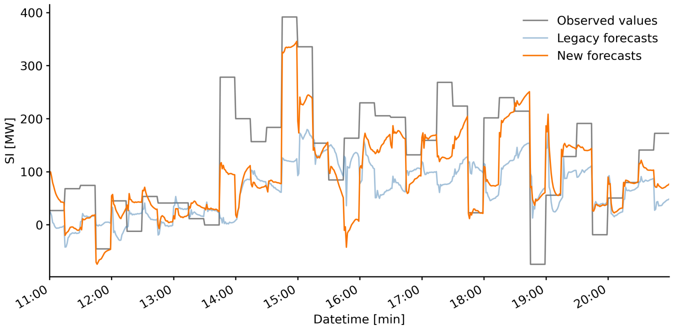

Fig. 7. Cumulative-15 system imbalance predictions alongside real observations and Elia's predictions. Time chosen randomly, corresponding to 2022/12/31.

In terms of probabilistic (distributional forecasts), the best results are obtained using a quantile regressor with the extended covariate set developed here. Over all of 2022, a 90% prediction interval (quantiles 0.95 and 0.05) can be forecasted with a reliability of 89.1% and an average Winkler score of 309.62. This represents an improvement of 6.5% with respect to what would be obtained using Elia's current configuration.

Forecast improvements are tested for statistical significance using the Diebold-Mariano test [23]. The tests are conclusive when checking any of the predictions obtained from the models presented in Table I against Elia's model: DM statistics are large and positive and p-values are zero across the board, strongly rejecting the null hypothesis that these results are caused by the specific choice of data values in the sample. Checking across the new models' predictions yields p-values of zero between linear and non-linear models, but the null hypothesis cannot be rejected when comparing different linear methods like standard linear regression and support vector regression ($p>0.05$).

Lastly, statistical analyses of forecast residuals show a complete absence of autocorrelation, suggesting that all temporal dynamics have been successfully exploited by the models fitted.

VI. CONCLUSION

This study presents an empirical description of grid system imbalance as a highly stationary and stochastic signal, which nonetheless exhibits certain periodicity patterns that may be exploited in forecasting. As observed earlier, it is possible to directly measure that Belgian system imbalance volumes have grown rapidly during the past three years, and that extreme imbalance events are becoming more common. This worrisome trend is at the forefront of the challenges facing the on-going energy transition, and the results shown here are likely to be apparent in neighbouring countries and the entire synchronous grid of Continental Europe.

Through a comprehensive process of feature engineering, a list of relevant covariates has been formulated for optimal short-term forecasting of quarter-hourly system imbalance. Notably, data on day-ahead and intra-day cross-border nominations has been found to be a powerful predictor, which in conjunction with autoregressive features and a number of other exogenous variables (wind power, load, net regulation volume, ambient temperature) can be exploited to build improved predictive models of system imbalance. Importantly, a wide number of covariates that would be expected to hold predictive value have been discarded. Signals relating to solar generation and energy market prices stand out in this regard. Additionally, wind power and load variables have provided less value that was initially expected. On the whole, this study finds evidence to support the interpretation that system imbalance is largely the result of market failure to predict and adapt to variable generation and demand, rather than a deterministic phenomenon that can be described in terms of patterns in electricity demand and weather conditions.

Initial expectations that non-linear modelling would bring about significant improvement have not been met. A strong tendency to overfit has proven to be an engineering challenge requiring careful hyper-parameter fine-tuning. Results suggest that the best performing non-linear models are shallow feed-forward neural networks, of no more than two layers and 6 to 20 neurons in total. Decision tree-based methods and kernelised support vector regression have not been found to produce any positive results, with decision tree ensembles appearing to be especially prone to overfitting. These findings partly justify the absence of more-advanced deep learning methods in this study, as no significant level of exploitable non-linear patterns has been identified. At best, non-linear models have only been seen to match or slightly outperform simpler autoregressive linear models. This comes at the cost of reduced interpretability, larger training data requirements, and much increased computational cost. Nonetheless, the predictions obtained from non-linear models differ significantly from the linear models, which makes for powerful forecast ensembles that produce the best overall results.

All in all, the forecasting methodology developed in this work reliably achieves a reduction in prediction error of 10-11% with respect to the system currently in deployment by Elia, bringing the cross-validated prediction error down to just over 100MW on average. Moreover, the study's approach favours simple, interpretable linear models, and readily available non-sensitive data; two features that facilitate expansion upon this work as the European grid continues to accommodate the integration of renewables in the coming years.

VII. Discussion

Possibly the largest limitation of this study is the absence of tests with advanced deep learning methodologies, and one of the main areas where this work could be expanded upon in the future. Initially it was planned to build LSTM and transformer-based neural networks, but, with limited time, work in feature engineering proved early on to be much more fruitful than attempts at non-linear modelling. Although none of the results obtained here indicate an advantage of larger more complex models over simpler linear regression, they do not disprove the potential for a well-calibrated neural network algorithm to outperform them. It would be particularly interesting to study whether a transformer-based architecture using time-dependent attention mechanisms to automatically select covariates would provide further improvement.

An additional aspect to consider is that the results shown here ought to be contextualised within the time frame and conditions they pertain to: the Belgian grid as of 2019-2023. This study's analysis of system imbalance reveals a highly dynamic and time-dependent nature. While these results hold relevance for other electricity grids, particularly within Europe, it is important to note that an optimal forecasting approach for other countries may differ significantly from the methodology applied for Belgium.

Furthermore, the prediction system developed in this study is heavily dependent on the data currently available; renewable power generation covariates have only been tested using Elia's external forecast providers, for example. The current provider's forecasts of wind power have an average absolute error of ∼261MW throughout 2022, much larger than the average prediction error obtained here for system imbalance. Load forecasts are wrong by ∼156MW on average. This possibly limits the extent to which system imbalance can be modelled as a function of load patterns and weather conditions.

Forecast quality is only going to grow in importance as the energy transition continues, and it is to be expected that improved covariate forecasts would aid in the development of system imbalance forecasting models.

Acknowledgment

This project takes place as a collaboration between the university student (author) and Elia Group, in the form of a full-time research internship at the Artificial Intelligence Center of Excellence. The author would like to thank Jose González, Thomas Craessaerts and Rachel Berryman of Elia Group for their continued trust and support throughout the development of this study.

References

[1] IPCC, *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2014.

[2] IEA, *Renewables 2022: Analysis and forecast to 2027*. International Energy Agency, Paris, 2022.

[3] IRENA, *Renewable Power Generation Costs in 2021*. International Renewable Energy Agency, Abu Dhabi, 2022.

[4] IEA, *Energy Technology Perspectivees*. International Energy Agency, Paris, 2020.

[5] S. Goodarzi, H. N. Perera, and D. Bunn, “The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices,” *Energy Policy*, vol. 134, 2019.

[6] L. Hirth, J. Mühlenpfordt, and M. Bulkeley, “The ENTSO-E Transparency Platform - A review of Europe’s most ambitious electricity data platform,” *Applied energy*, 2018.

[7] ENTSO-E, “Manually Activated Reserves Initiative (MARI),” https://www.entsoe.eu/network_codes/eb/mari/, 2023, [Online; accessed 30/01/2023].

[8] N. Tesla, “A New System of Alternating Current Motors and Transformers,” *American Institute of Electrical Engineers*, 1888.

[9] E. Commission, “Commission Regulation (EU) 2017/1485 of 2 August 2017, establishing a guideline on electricity transmission system operation,” 2017.

[10] E. T. Belgium, “Rules for the Compensation of Quarter-hourly Imbalances,” 2020.

[11] ENTSO-E, “Continental Europe significant frequency deviations—January 2019,” 2019.

[12] J.-F. Toubeau, J. Bottieau, Y. Wang, and F. Vallée, “Interpretable probabilistic forecasting of imbalances in renewable-dominated electricity systems,” *IEEE Transactions on Sustainable Energy*, 2021.

[13] C. Contreras, “System imbalance forecasting and short-term bidding strategy to minimize imbalance costs of transacting in the spanish electricity market,” *Repositorio Comillas*, 2016.

[14] J. Dumas, I. Boukas, M. M. de Villena, S. Mathieu, and B. Cornélusse, “Probabilistic forecasting of imbalance prices in the belgian context,” in *2019 16th International Conference on the European Energy Market (EEM)*. IEEE, 2019.

[15] N. Kayedpour, A. E. Samani, J. D. De Kooning, L. Vandevelde, and G. Crevecoeur, “A data-driven approach using deep learning time series prediction for forecasting power system variables,” in *2019 IEEE 2nd International Conference on Renewable Energy and Power Engineering (REPE)*. IEEE, 2019.

[16] E. Makri, I. Koskinas, A. C. Tsolakis, D. Ioannidis, and D. Tzovaras, “Short Term Net Imbalance Volume Forecasting Through Machine and Deep Learning: A UK Case Study,” *Artificial Intelligence Applications and Innovations*, 2021.

[17] B. Stappers, N. G. Paterakis, K. Kok, and M. Gibescu, “A class-driven approach based on long short-term memory networks for electricity price scenario generation and reduction,” *IEEE Transactions on Power Systems*, 2020.

[18] E. Group, “System imbalance forecast next quarter hour (near real-time). OpenDataElia.” https://opendata.elia.be/explore/dataset/ods147/custom/, 2023, [Online; accessed 03/07/2023].

[19] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice, 3rd edition*. OTexts: Melbourne, Australia, 2021.

[20] G. E. Box and D. A. Pierce, “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *Journal of the American statistical Association*, vol. 65, no. 332, 1970.

[21] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, “Stl: A seasonal-trend decomposition,” *J. Off. Stat*, vol. 6, no. 1, 1990.

[22] R. Mushtaq, “Augmented dickey fuller test,” *Social Science Research Network*, 2011.

[23] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *Journal of Business & economic statistics*, vol. 20, no. 1, 2002.

[24] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, “Boruta–a system for feature selection,” *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[25] P. J. Huber, “Robust estimation of a location parameter,” in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[26] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, 1995.

[27] L. Breiman, “Random forests,” *Machine learning*, vol. 45, 2001.

[28] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” *Advances in neural information processing systems*, vol. 30, 2017.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, 1986.

[31] J. M. Bates and C. W. Granger, “The combination of forecasts,” *Journal of the operational research society*, vol. 20, no. 4, 1969.

[32] A. F. Atiya, “Why does forecast combination work so well?” *International Journal of Forecasting*, vol. 36, no. 1, pp. 197–200, 2020.

[33] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *International journal of forecasting*, vol. 22, no. 4, 2006.

[34] R. L. Winkler, J. Munoz, J. L. Cervera, J. M. Bernardo, G. Blattenberger, J. B. Kadane, D. V. Lindley, A. H. Murphy, R. M. Oliver, and D. Ríos-Insua, “Scoring rules and the evaluation of probabilities,” *Test*, vol. 5, 1996.

[35] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, 2017.

[36] Elia, “Open Data,” https://opendata.elia.be/pages/home/, 2023, [Online; accessed 30/01/2023].

[37] T. Urdiales, “Code Repository,” https://github.com/tomasurdiales/System_Imbalance_Forecasting, 2023.

[38] H. Kazmi and Z. Tao, “How good are tso load and renewable generation forecasts: Learning curves, challenges, and the road ahead,” *Applied Energy*, vol. 323, p. 119565, 2022.