Point and interval forecasting of short-term electricity price with machine learning: A theoretical and practical evaluation of benchmark accuracies for the Dutch intraday market

Thesis report

T.P. (Timo) Vijn

Submitted in partial fulfillment of
the requirements for the degree of
Master of Science (MSc)

Systems & Control
Delft University of Technology
Delft, The Netherlands
Version 4.0
October 15, 2021

Intentionally blank

Acknowledgements

Throughout the whole period of working on the topic, it has ignited numerous conversations with many people around me. Some of them may even have noticed something that can only be described as joy and excitement in my eyes or in my voice, which has not been observed there often, regarding such a topic. It means that the university, my supervisors, and the extensive field of research have given me the opportunity to learn and to experiment, to try and to fail, to try and to succeed, and most importantly to express some creativity within a technical boundary. I loved that.

Acknowledgement goes to Ties van der Heijden. Thank you for your openness, for your honesty, and for your enthusiasm. You were kind to help me with some personal struggles I needed to overcome, and having you as a knowledgeable buddy made it much more enjoyable for me. Acknowledgement goes to Edo Abraham. Thank you for the critical questions that you always delivered with a smile. I greatly appreciate that you were able to see beyond the confines of my work. It enabled me to do the same. Acknowledgement goes to Bart de Schutter. Thank you for your guidance, and for your swift replies no matter what. Only because of you did I discover electricity price forecasting and did I get the opportunity to work with Ties and Edo.

Ties, Edo, Bart—your work and work ethic, technical expertise, and attitude towards supervision have my admiration and respect; the way I felt before/during/after meetings is definitely something to remember.

Finally, acknowledgement goes to Jesus Lago. Thank you for sparking my interest in the topic when I left your office one morning with a thick bundle of reading material that I read with pleasure—be it with limited understanding at the time. This research contains your name and that of Mr de Schutter more than once, which implies that I had a great foundation to work from.

Several researches were published at the end of 2020 [103] and well into 2021 [54, 101] with scopes that partly overlapped my scope. Initially that felt as though the novelty of my work was in danger, but I can laugh about it now. I hope it implies that my work points towards where electricity price forecasting is heading. It is a good thing that this research does not nearly conclude the topic. If anything, it should demonstrate that there are many avenues that lie wide open. Deep down I wish it was me that continued where this research ends, although I am sure that fresh minds will be able to take it a step or two further than I ever could.

The past year was full of challenges that were sometimes hard to overcome, although I enjoyed the endless sculpting from humble beginnings nonetheless. I sincerely hope that I come across the topic again sometime in the future. Until that day, this report summarizes my efforts to contribute to the field of short-term electricity price forecasting. The "my" that would mean nothing if not standing on the shoulders of supervisors, of friends, and of family.

I am grateful for your generous support.

Timo

Delft, The Netherlands
October 15, 2021

This thesis report is submitted in partial fulfillment of the requirements for the degree of Master of Science (MSc) by student T.P. (Timo) Vijn, under supervision of Prof.dr.ir. B.H.K. (Bart) de Schutter, Dr.ir. E. (Edo) Abraham, and Ir. T.J.T. (Ties) van der Heijden.

Abstract

Background This research investigates price forecasting for the Dutch intraday electricity market. Trading volumes on this liberalized and close-to-delivery market have grown considerably as more variable renewable energy has entered the energy mix. That imposes uncertainties to market participants and challenges to the balance between supply and demand as long as large-scale storage is mostly unviable. Intraday trading is often incited by conditions that are unforeseen during day-ahead trading, which can be more or less drastic from hour to hour. The dynamic of day-ahead prices generally represents a reasonable indication for the dynamic of intraday prices, although they might abruptly diverge due to many nonlinear and unobservable factors. Complex and risk-introducing characteristics such as high volatility, negative values, and spikes are typically pronounced in intraday markets and necessitate reliable price forecasts.

Objective In the field of electricity price forecasting, there is limited attention for the intraday market. Research mostly examines point forecasts in single-step-ahead horizons, while multi-step-ahead horizons and interval forecasts are more applicable in practice as they offer a better foundation to base trading and/or dispatch schedules on. This research provides benchmark accuracies for forecasting of an aggregated price of the Dutch intraday market. While single-step-ahead point forecasts for that unresearched market provide novel insights already, the scope of this research also includes multi-step-ahead interval forecasts. Within that scope, this research evaluates accuracies attained by artificial neural network models that—despite their successful application for day-ahead price and other time-series forecasting—have not been studied extensively for intraday price forecasting. That includes a multi-output quantile neural network. Finally, this research addresses a nuance that is regularly overlooked; widely used scoring metrics usually represent merely an indication of overall accuracy that might not match with what is deemed as superior in practice. Therefore, this research applies forecasts in an operational context to provide new insights into the evaluation of superiority.

Methodology A forecasting procedure organizes several stages of in-sample and out-of-sample testing. It is systematic and similar to how forecasting is carried out in practice; the number of arbitrary choices is kept as low as possible. Initialization on the period from 2016 through 2017 provides insight into the size of the windows in rolling window estimation and into the selection of Dutch and German features. Calibration accounts for potentially evolving market conditions and updates the feature and hyperparameter sets during the out-of-sample rolling window estimation. Accuracies of point and interval (quantile) forecasts from naive, regression, and artificial neural network models are evaluated on the basis of scoring metrics (rMAE, MAE, MAPE, RMSE, PL, CRPS). Analysis of residual errors indicates whether dynamics of the intraday price are captured adequately. A simulation of a generic system comprising a battery and a wind turbine smartly dispatches stored energy according to a schedule optimized with model predictive control and future information on intraday price and generation. This research investigates whether the superior forecast leads to a higher profit, or to a different dispatch schedule altogether.

Conclusion The considered artificial neural network model based on a multilayer perceptron, which is capable of incorporating nonlinear relationships, provides a superior point forecast in the out-of-sample test. A similar architecture that is trained on the basis of several quantiles does not outperform a quantile regression averaging of the considered point forecasts, however. Relative and absolute accuracies of all forecasts are shown to vary significantly with the delivery hour and with the price regime. The year of 2018, with attained rMAEs and CRPSs as low as 0.81 and 3.53, is considerably more challenging than the year of 2019, with attained rMAEs and CRPSs as low as 0.77 and 2.24, due to high volatility and many extreme prices. Results of the simulation demonstrate that more accurate forecasts lead to slightly higher profits, although much of the essential information is captured by all forecasts. The schedule itself is more sensitive to the price forecast, however, and dispatch frequency and volume deviate more than 8% when based on different forecasts. Practitioners might take that into account and not gravitate towards point forecasts merely on the basis of accuracy.

Intentionally blank

Table of contents

Intentionally blank

Figures

Tables

Abbreviations

ANN	artificial neural network
ARIMA	autoregressive integrated moving average
CRPS	continuous ranked probability score
DM	Diebold-Mariano
EPF	electricity price forecasting
GRU	gated recurrent unit
GW	Giacomini-White
LASSO	least absolute shrinkage and selection operator
LSTM	long short-term memory
MCP	market clearing price
MLP	multilayer perceptron
RF	Random forest
RFE	recursive feature elimination
RNN	recurrent neural network
SVR	support vector regression
TPE	tree-structured parzen estimator

Intentionally blank

1 Introduction

Similarly to most major electricity markets in Europe, liberalisation of the Dutch electricity market started during the 1990s. From that point onward, the supply of electricity was no longer in hands of a single monopolistic supplier and other parties were able to partake in trading activities by investing in power plants and transmission lines, for instance. Consumers benefited from a transparent market with more competitive pricing, among other things, and responsibilities of authorities shifted from direct planning of supply to design and regulation. Market conditions have gradually turned more uncertain in such 'decentralised' environments, and strategic analyses as well as scenario and risk management have become crucial factors to market participants in considerations of scheduling. As long as large-scale energy storage is still mostly infeasible, there remains a fundamental requirement of constant balance between energy supply and demand to ensure grid stability. That amplifies the need of efficient matching of orders on day-ahead and intraday markets, where buyers and sellers of energy enter into contracts close to the time of delivery. Reliable forecasts of the price of those contracts, among other things, are indispensable for all parties involved in that process.

1.1 Research motivation

In the Netherlands, the ongoing growth of renewables in the energy mix has its origin around the same time as market liberalization. Initiatives that support the move to a more carbon-neutral energy mix, now heavily incentivized by political agendas and financial stimuli, flow from the the premise that it plays a vital role in mitigating dependence on importing energy, and in reducing the effects of global warming [38]. Generation of variable renewable energy sources, e.g. wind and solar, generally deviates more from generation schedules than that of traditional energy sources. The increased level of uncertainty that imposes to market participants leads to higher needs to balance portfolios just before delivery on the intraday market, which allows for continuous trading throughout the day, and up to 5 minutes before delivery. A steep rise of Dutch intraday trading volumes is the result, with repeated year-on-year growth figures between 30 and 70% during the last five years [23, 24, 25, 26, 27]. It is not surprising that the vast majority of literature investigates the German market; in 2020, German trading volumes were 216,221.3 and 63,627.0 GWh on day-ahead and intraday markets respectively, which represents approx. 45% of the total traded volume on all major European day-ahead and intraday markets [27]. To assess the added value of forecasting approaches it is vital to have access to benchmarks and to historical data. A body of research where studies investigate the same market, which is often the German market, is stimulating to both those ends. That does not imply that research on other markets is unnecessary; smaller and less-studied markets may be able to expose how different conditions such as market maturity, energy mix, and environmental circumstances affect forecasting. What is more, discussing results on an unexplored market can set the benchmarks for future work.

The endeavour of forecasting the price of energy contracts has drawn broad academic and commercial interest. Gaining insight into possible future price realisations, and investigating how to improve models and modeling approaches to reduce forecast uncertainties has the potential to improve decision-making of market participants. That might lead to a more stable grid and increased profitability [43]. The price of short-term energy contracts is subject to many complex, nonlinear, and unobservable factors and exhibits challenging characteristics such as high volatility, negative prices, spikes, and jumps, that are generally much more pronounced than in other commodity or stock markets [15]. The field of time series forecasting offers many models and modeling approaches that are often tested extensively and challenged repeatedly over a long period of time. Autoregressive integrated moving average (ARIMA) models, for instance, are among the traditional models that have been demonstrated to generally offer dependable forecasts. On the other side of the spectrum are models that are less familiar, and that are put forward as 'challengers'. Nowadays, artificial neural networks (ANNs) are often found in that category. The potential of various types of ANNs have been shown by research, and especially the multilayer perceptron (MLP) and long short-term memory (LSTM) have been able to live up to the proposition that ANNs are particularly applicable to highly nonlinear problems in time series forecasting. That is not to say that they offer a 'holy-grail' solution to every problem; one should be aware of the specific challenges in terms of interpretability and optimisation that should be addressed thoughtfully. Besides the choice of models and optimisation, there are many more considerations regarding features, out-of-sample testing, and evaluation of forecasts that need to be addressed.

1.2 Problem description

The numerous models and modeling approaches that have been proposed to push performance beyond that of benchmarks, and the comprehensive frameworks that offer guidelines and best practices [7, 64], demonstrate that electricity price forecasting is an advanced field of research. Nevertheless, the attention for intraday markets is limited, despite their growing relevance. Many of the methodologies and recommendations published in the more extensive body of research on day-ahead markets can be utilized for the intraday market, but differences between the two require distinct approaches. For one, the continuous trading of most intraday contracts versus the auction-based trading of day-ahead contracts lead to fundamentally different circumstances regarding data availability.

Inward looking on intraday price forecasting shows that research scopes are often limited to (single-step-ahead) point forecasts, although recent recommendations note that interval forecasts are an obvious avenue for future work [78, 71]. By giving an indication of forecast uncertainty, interval forecasts offer a richer and truer representation of reality. In consideration of real-world applicability, broadening the research scope to include interval forecasts, and also assessing to what extent performance holds up when forecasting in multi-step-ahead horizons might offer profound perspectives that are novel to intraday price forecasting.

Ideally, gains from forecasts are evaluated in situations where they will actually be used [6]. In many cases, that is not or not all a possibility, although some reluctance to base conclusions upon accuracy is appropriate. Those measures might not fully coincide with financial measures that are used in operational contexts and thus might fail to address problems that are important to practitioners. That matter is identified in literature and alternative approaches that consider optimal participation of power plants are brought forward [85, 67]. Notwithstanding, a lot of research does not acknowledge any theoretical-practical mismatch at all, and utilizes traditional measures only.

1.3 Research objective

The research objectives rest on *forecasted series* of the ID3 price index, which contains a single value per hour in the case of single-step-ahead point forecasting. A forecasting model is trained on T observations of a target variable $\boldsymbol{y} \in \Re^{T}$ and T observations of F explanatory variables $\boldsymbol{X} \in \Re^{T \times F}$ to approximate the relationship between them. Subsequently, that model can be exploited (*alias* tested) with observations of explanatory variables only, to forecast the target variable.

Training and testing is systematically performed, and parameters are re-adjusted continually. Explanatory variables are historical (and future) price variables as well as exogenous variables. The *reduced candidate set* contains all explanatory variables that partake in the forecasting procedure, and is the result of converting raw data to a candidate set of features and assessing how it can be reduced in size at little or no cost to accuracy.

In consideration of the limited research and increasing intraday trading volumes on intraday markets in general, there is the opportunity to investigate components that constitute a systematic forecasting procedure. Furthermore, due to the lack of research on the Dutch intraday market, there are opportunities to establish benchmark accuracies. The research objective is to evaluate forecasts of an aggregated price for the Dutch intraday market from naive, regression, and ANN models. To push the boundaries of the current body of research, not only point forecasting but also interval forecasting is investigated, in single- and multi-step-ahead horizons. A systematic forecasting procedure ensures that there are a limited number of arbitrary choices and that "what is considered optimal" is reconsidered repeatedly. To assess whether theoretical superiority coincides with practical superiority, forecasts are evaluated on the basis of accuracy as well as on the basis of profit when deployed in an operational context. Research questions direct this research to measurable findings that contribute to the field of electricity price forecasting (EPF).

As there is no literature to directly compare results with, it is necessary that this research establishes accuracies for relatively simple, linear models.

Research question 1 *What accuracies do naive and regression models of varying complexity attain on the task of point and interval forecasting of an aggregated price in single- and multi-step-ahead horizons?*

Findings in literature demonstrate that ANNs models that are capable to learn nonlinear relationships attain high accuracies in other (similar) markets. They might thus be able to provide more accurate forecasts than the naive and regression models. This research investigates two such models, and addresses the complexities in training.

Research question 2 *What accuracies do ANN models, including those with MLP and GRU architectures, attain on the task of point and interval forecasting of an aggregated price in single- and multi-step-ahead horizons?—Are their forecasts superior to those from naive and regression models?*

Finally, forecasts are deployed in an operational context to assess whether theoretically-found superiority coincides with practically-found superiority.

Research question 3 *Does theoretical superiority of forecasts coincide with practical superiority, such that more accurate forecasts lead to more profitable dispatch schedules in a generic simulation of an energy plant with storage capacity?*

1.4 Report outline

Chapter 2 illustrates the relevance and complexity of intraday markets through recent developments and price characteristics. Chapter 3 explicitly states the procedures from raw data to a candidate feature set. Chapter 4 introduces the approach to point forecasting, and presents point forecasting models in three distinct categories. Chapter 5 has a similar structure, but for interval forecasting. Chapter 6 describes the procedure of forecasting, specifically of initialization, calibration, and exploitation, and explains how three intricacies are incorporated. Chapter 7 discusses results of the forecasting procedure with special attention to the performance across the various models and intricacies. Chapter 8 establishes a case study where the dispatch of energy from a power plant with storage capacity is optimized in a framework of model predictive control, and Chapter 9 presents results. Chapter 10 presents conclusions, and finally Chapter 11 offers a discussion, that includes several avenues for future research and limitations of this research.

2 Intraday markets: Growing relevance and complexity

This chapter demonstrates the growing importance of the intraday market, and of intraday price forecasting. It is concluded that in the Netherlands, there is a strong growth of variable renewable energy in consumption and production in an otherwise mostly constant energy mix. That growth is expected to persevere. Uncertain schedules of variable renewable energy sources introduce uncertainty on energy markets, and thus higher volumes are traded on the intraday market than before. Energy trading is thus shifting closer to delivery and the difference between volumes on day-ahead markets and intraday markets is shrinking every year. Dutch intraday price has complex characteristics such as seasonality, spikes, and negative values. A distribution with a high degree of skewness and kurtosis and that is inconsistent from year to year is the result.

2.1 Short-term trading and forecasting

The volatility and uncertainty that is inherent to variable renewable energy can have serious impact on the trading of energy. For one, their almost zero marginal cost and therefore relatively low market value can send day-ahead prices down considerably whenever their production becomes available. In addition to that, they generally increase the need to adjust day-ahead positions on intraday and imbalance markets [68]. Since the 1990s, variable renewable energy sources have become a more significant part in energy mixes around the world, and their share is only expected to increase further for the foreseeable future. The growth of variable renewable energy in the Netherlands becomes particularly evident from the severe growth of solar and wind energy during the twenty years prior to 2020; from a 0.8 to a 13.9% share in production and from a 0.2 to a 4.2% share in consumption. The shares in the Netherlands are currently about half of those in Germany, which has undergone an even more severe growth of solar and wind energy during this period; from a 1.0 to 28.5% share in production and from a 0.4 to an 11.8% share in consumption. From the energy mixes of production and consumption in the Netherlands, shown in Figure 1, it becomes apparent that solar and wind have been steadily growing within an otherwise mostly constant landscape.

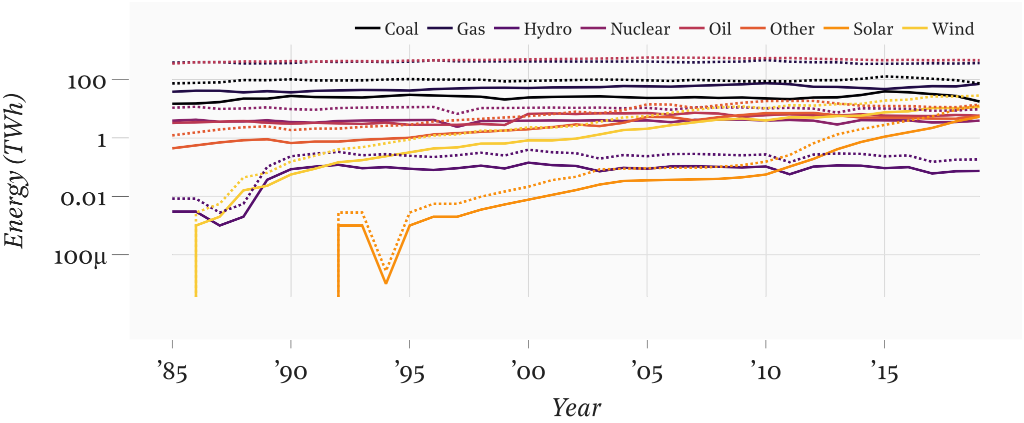


Figure 1: Energy production and consumption in the Netherlands by energy source. *Production (line) and consumption (dash). Data from BP [13].*

The growth of variable renewable energy in supply is a major factor that has urged the move towards trading closer to delivery and thus the move towards trading on the intraday market. Figure 2 shows the growth of intraday volumes for the Netherlands and for Germany. In addition, there is a move towards trading closer to delivery *within* the intraday market; in 2019 for instance, the German market saw 30% of the total hourly intraday volume traded within one hour before delivery while that was only 12% in 2012.

For successful integration of variable renewable energy sources into the energy system, it is crucial that market players have access to forecasts of the primary elements of energy markets. The price of energy on the short-term markets, as well as load and generation, are elements that literature is concerned with most. The possible future realisations of these elements can then be used in a form of

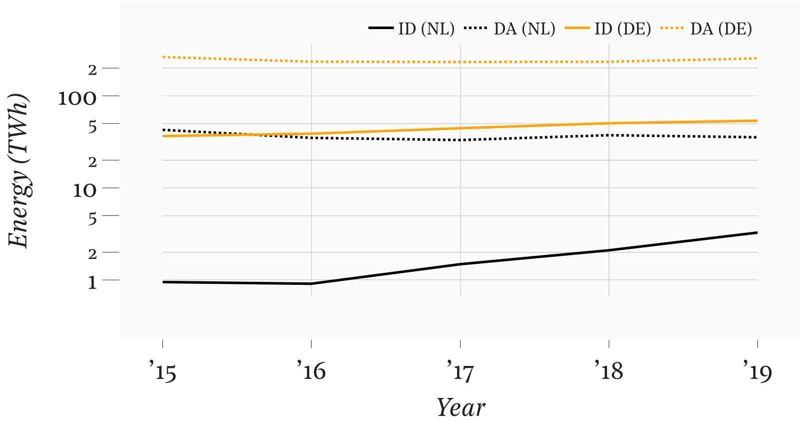


Figure 2: Energy volumes traded on the Dutch and German intraday and day-ahead markets. *Data from EEX [22, 23, 24, 25, 26].*

optimization. Typical examples of such "predict-then-optimize" frameworks are found in the design and optimization of dispatch strategies for renewable generators [61].

2.2 Intraday market dynamics

Because trading volumes on the intraday market have risen considerably, it has become a more suitable place for parties to mitigate deviations from earlier accepted contracts. Still, actual supply might deviate from planned schedules even after intraday adjustments and thus real-time balancing by system operators has not become redundant. However, there is a shrinking amount of deviations settled on the balancing market as market players become better at adjusting their schedules on the intraday market, in spite of the growing uncertainty due to variable renewable energy [93]. Overall, the purpose of the intraday market to limit severe shortfalls or surpluses seems to be fulfilled. What is more, the interplay of the three major energy markets, and how it has shifted over time, demonstrates the relevance of this market as well as of adequate price forecasts.

The intraday and day-ahead markets differ in terms of product opening and closing times. What is more, the pricing mechanism of these two markets are fundamentally different. On the intraday market, bid and ask orders placed for a certain product enter that product's limit order book. A limit order enters with a specified price and can match fully or partially with an opposing order, or remains in the order book without a match. A market order gets matched instantly with the best opposing order. Due to the "pay-as-bid" pricing mechanism of the *continuous* intraday market, orders for the same product, that potentially enter the order book at almost the same point in time, can lead to contracts of very different prices [73]. There is a stark contrast with the pricing mechanism of the *auction-based* day-ahead market, that makes use of a market clearing price (MCP); the price of the last accepted contract for a product becomes the price for all contracts on that product.

A number of observations regarding trading behaviour can be done upon examination of the contracts that were accepted for all products on an arbitrarily chosen day in the past, shown in Figure A.1. In this particular subset, 80, 50, and 4% of all contracts were accepted no earlier than 4 hours, 2 hours, and 1 hour before delivery. There were slightly less contracts that were accepted extremely close to delivery than on an average day, as evaluated on the whole period these figures become 76, 47, and 6%. What generally holds true is that most high-volume contracts are accepted many hours before delivery. Something that also regularly occurs are multiple contracts for different products, of equal or varying volume, that are accepted at exactly the same time. An example of this occurring in Figure A.1 are three transactions just before 06:00 for the products with delivery at 12:00, 13:00, and 14:00. A possible explanation is that parties determine their position for multiple products at a certain time before algorithmically placing bids or asks on all products at once. Extreme price variation between contracts can be observed; on this day, there were a total of 7 occurrences where the price of two contracts, that were for the same product and were accepted within 10 seconds of each other, differed more than 25%.

Unlike how the day-ahead market has an MCP for all products, the continuous intraday market has no naturally occurring and uniquely defined price for each product. Therefore, EPEX has introduced several intraday price indices: the ID1; the ID3; and the IDFull. These indices represent the volume-weighted average price of all continuous trades for a product within respectively one hour, three hours, or all hours before delivery of the product, respectively. Given a set $\mathbb{T}_t = {}^{-3.25}_{-3}\mathbb{T}_t(t_d)$, that contains all transactions for a product with time of delivery t_d in the time window from 3 hours to 15 minutes before t_d. The ID3 index represents the volume-weighted average price of all transactions $k \in \mathbb{T}_t$, and is calculated by

$$\text{ID3}_{d,h} = \left(\sum_{k \in \mathbb{T}_t} v_k \right)^{-1} \sum_{k \in \mathbb{T}_t} v_k p_k \tag{1}$$

where v_k is the volume and p_k is the price of transaction k. This research is concerned with forecasting the price of the ID3 index (*alias* ID3 price), although it is noted that the price of individual transactions may vary considerably [52]. That holds true even within the 4-hour window before delivery, as demonstrated by the trades shown in Figure A.1. Figure 3 shows the deviation between the price of individual transactions and the ID3 price, i.e. $p_k - \text{ID3}_{d,h}$ for $k \in \mathbb{T}_{d,h}$. Color represents the number of observations within a bin. Deviation is especially obvious for products with delivery from 15:00 through 23:00, and prices of transactions are distributed rather uniformly around the ID3 price. Deviations of 5 to 10 €/MWh are not uncommon. Prices of transactions for products with delivery at night are relatively more concentrated around the ID3 price.

Although the Dutch ID3 index does not represent the price of *all* intraday transactions, Figure 3 shows that it at least represents the price of *most* transactions such that the ID3 price is considered as an adequate indication of individual transactions. Forecasting the price of individual transactions adds many complexities that lie outside the scope of this research. Similarly to earlier research [99, 71], this research employs a time of forecasting that is four hours earlier than the time of delivery, i.e. $t_f = t_d - 4$. The forecast is thus provided well before the start of the ID3 window, so that the practitioner has enough time to exploit the forecast and also has a high probability to settle on a satisfactory price that could be the ID3 price.

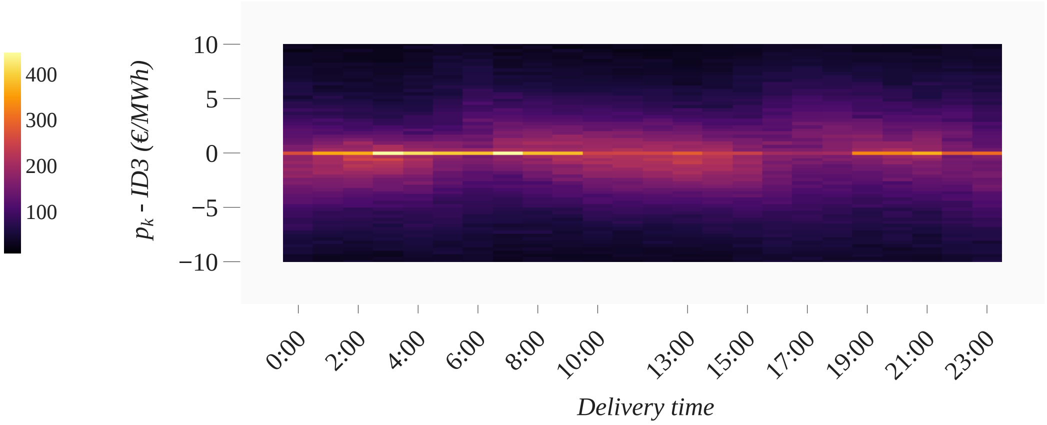


Figure 3: Deviation between the price of individual transactions and the ID3 price. *Color represents number of observations. 2015–2020. Data from EEX [28].*

Empirical distribution Time series analysis is concerned with many distributional assumptions. Figure 4 shows the empirical distribution of the ID3 series and Table B.1 shows the main statistical properties. The distribution is clearly asymmetric (skewness much greater than 0), and it has heavier tails than the fitted normal distribution (kurtosis much greater than 3). In practice, that is reflected by an unequal spread of observations around the empirical mean and by more and/or more extreme outliers than when observations would be drawn from a Gaussian distribution with the same mean and standard deviation. Figure B.1 shows the distribution for the German ID3 series. The German price distribution has lower skewness and kurtosis, and is thus more accurately approximated by a normal distribution than the Dutch price distribution.

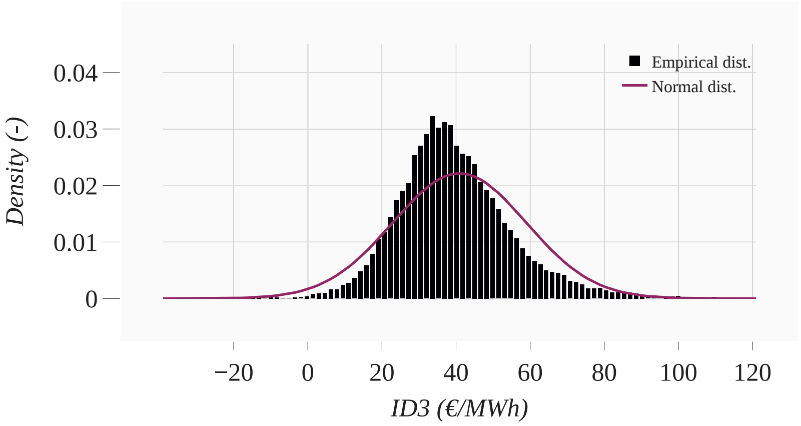


Figure 4: Empirical distribution of the Dutch ID3 index. *Color represents density. 2015–2020. Data from EEX [28].*

The empirical distribution that is found for the whole period might not be representative for all periods that it comprises. For instance, periods of tightness between demand and supply, or of severe errors in load and generation forecasts, might lead to considerably deviant distributions [114]. What is more, the distribution might change as the market evolves over the years. To confirm whether the aggregated distribution shown in Figure 4 is representative, the empirical distributions for individual years are examined. Figure 5 shows that the distribution varies considerably from year to year. Particularly remarkable is that 2019 is more symmetrical than other years, that 2018 has far more occurrences of high prices, and that 2016 and 2020 have far more occurrences of low prices. In 2015, 2018, and 2020 prices are less concentrated than in other years. Figure B.2 shows the distributions for the German ID3 series. For all considered years, the German price distributions are similar to the Dutch price distributions. This result suggests that systematically lower or higher prices observed on a yearly scale are due to effects that transcend national markets.

Particularly 2018 and 2019 are important for later stages of this research, as forecasts are evaluated from an out-of-sample test of that period. While intraday price seems to be rather temperate for 2019, it is far more extreme for 2018. That year saw an increase in electricity prices across Europe that was especially severe for the Netherlands, which is explained mostly due to increases of carbon emission allowances [25]. In 2020, the social distancing rules that were introduced in reaction to the COVID-19 pandemic led to a deeply rooted imbalance between supply and demand due to an abrupt global decline in power consumption [27].

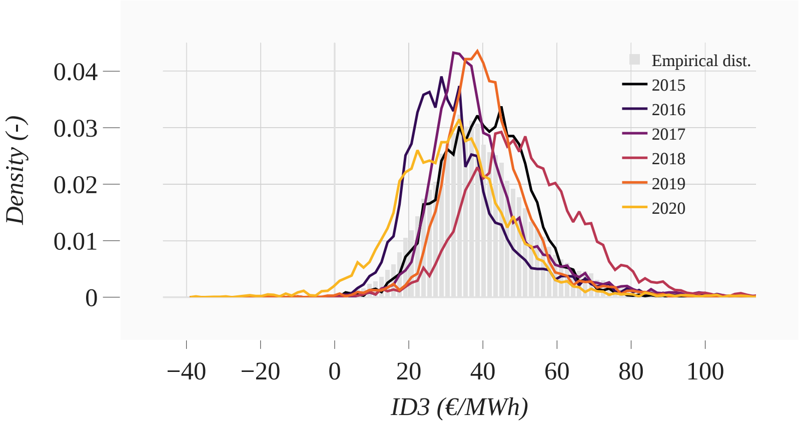


Figure 5: Empirical distribution of the Dutch ID3 index as function of year. *Data from EEX [28].*

To further explore temporal variations in statistical properties, Figure 6 shows the kernel density estimates when individually calculated for the hours of a day. Color indicates density. It can be seen that values are dispersed for all hours, although certainly to a higher extent during the day (from 06:00 through 18:00). Not only lower values, but also a more concentrated distribution is generally observed during the night. Figure B.3 shows the kernel density estimates for the German ID3 series. German prices are more concentrated than Dutch prices, and the difference between higher and lower prices is more pronounced.

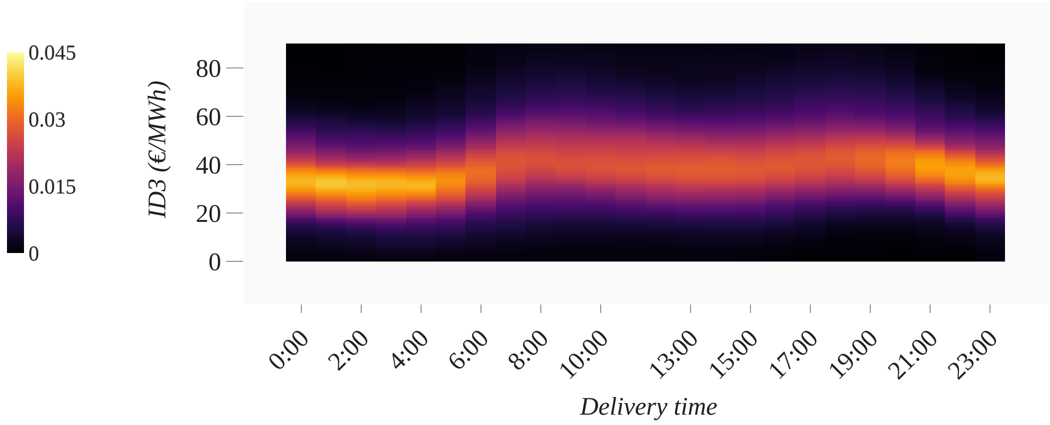


Figure 6: Kernel density estimates of the Dutch ID3 index as function of delivery time. *2015–2020. Data from EEX [28].*

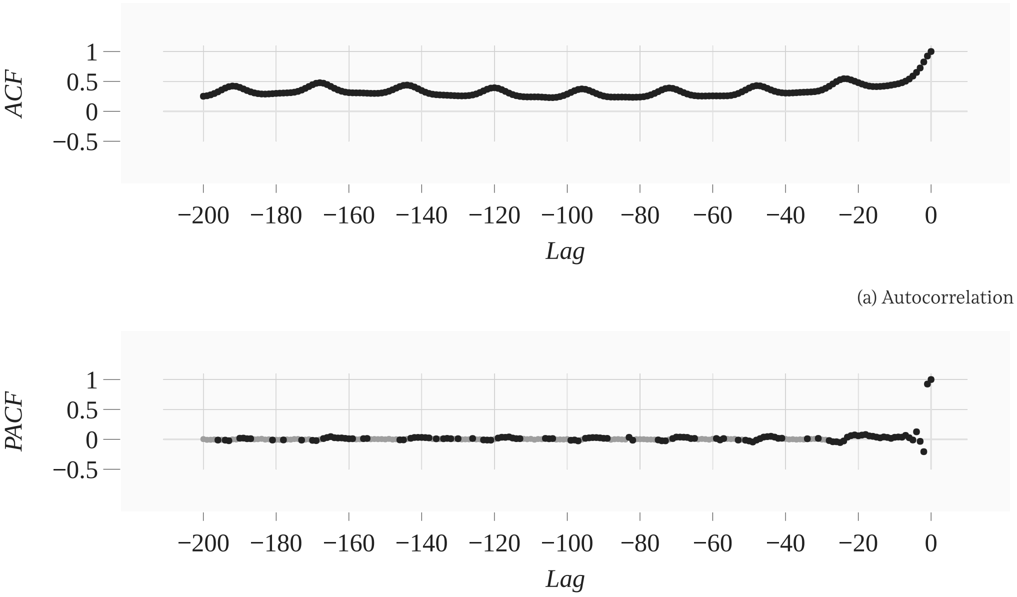
2.3 Price characteristics

It is important that a number of challenging characteristics that spot prices have are dealt with. That includes mean reversion, high volatility, seasonality, spikes, and negative values. What follows is a short evaluation of these characteristics for the Dutch energy market.

Mean reversion While the prices of many other commodities evolve mostly unconstrained, the price of energy gravitates around production costs. Although short-term price movements may diverge, prices generally tend to revert to production cost levels [15]. Moreover, demand is highly influenced by economic activity and by environmental conditions that are subject to mean reversion themselves [87]. It is not surprising therefore, that the mean reversion process is often used as basis for energy price simulation [15, 75, 55].

Seasonality A prominent characteristic of intraday price is the existence of seasonal effects. Repeating patterns can therefore be observed on daily and weekly timescales, which arise from calendar-dependent patterns that are naturally present in supply and demand. Autocorrelation finds the strength of the linear relationship between a series and its lagged series, and can therefore give insight into such repeating patterns. The autocorrelation shown in Figure 7 shows local maxima at every 24-hour multiple, which suggests a diurnal cycle. What is more, the local maxima at every 168-hour multiple are especially obvious, which suggests weekly seasonality. When there is a significant correlation with lag i, and lags i and j are also correlated, then autocorrelation might also find a significant correlation with lag j although there might not be a significant correlation. For this reason, Figure 7 also shows partial autocorrelation, which demonstrates that correlations with the 1-, 2-, and 4-hour lags are strong, while all others—including those with the 24- and 168-hour lags—are much weaker. Figure B.4 shows the autocorrelation and partial autocorrelation for the German ID3 series. The correlations are very similar, although the variation between relatively strong and weak correlations are slightly more obvious for the German prices. The lags that are 24-hour multiples have similar correlations for Dutch and German prices, but the lags that lie in between have stronger correlations in the Dutch price series than in the German price series.

To further demonstrate seasonal effects, Figure B.8 shows the mean of the ID3 series evaluated on different timescales. On an hourly scale, Figure B.8a shows that prices reach high values in the morning at 08:00 and in the evening at 18:00, and low values at night at 02:00. This coincides with the peak and base loads during the day. On a daily scale, Figure B.8b shows that prices reach high values at the start of the work-week, especially on Tuesdays, and continuously decrease towards and during the



(a) Autocorrelation

(b) Partial autocorrelation

Figure 7: Autocorrelation and partial autocorrelation of the Dutch ID3 index. *Coefficients that are within the 95% confidence interval are a lighter shade.*

weekend. Finally, on a monthly scale, Figure B.8d shows that intraday prices reach higher values in Winter months, especially from November till January, and reach lower values during the months of spring. These seasonal effects seem to be comparable to what has been found in literature for similar markets [66]. The whiskers in Figure B.8 represent the 95% and 50% quantile range. They are wide, which indicates that price variability is high.

Extreme prices Figure 8 shows a normality test for the period from 2015 through 2020. A normally distributed series is represented by a straight line. It is evident from the tails of the empirical distribution that extreme prices have a much higher probability of occurring than when assuming a Gaussian distribution. Especially at the side of high prices, the empirical distribution starts to considerably deviate from a Gaussian distribution at about 100 €/MWh.

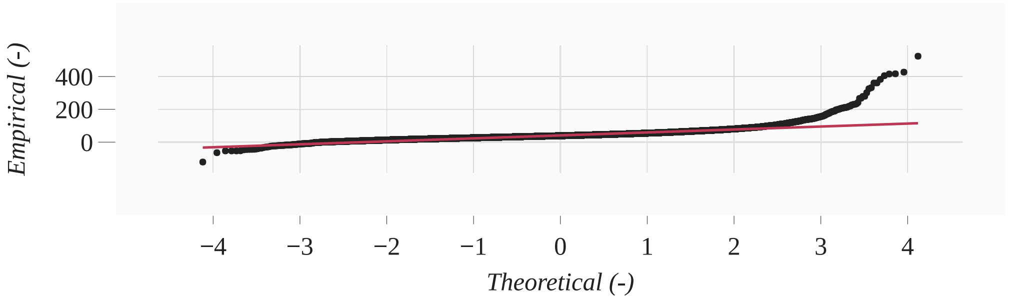


Figure 8: Normality test for the Dutch ID3 index. *2015–2020. Data from EEX [28].*

A characteristic of energy prices that contributes to this effect is the occurrence of many spikes. There is no absolute consensus about their definition, although they are a major factor contributing to the complexity of the series. Traditional mean-reverting jump-diffusion models intuitively model the smooth variation in price and jumps in price. Spikes further complicate such models. As the reversion rate and jump intensity, generally assumed to be constant, can be extremely high for spike periods and are generally much lower for non-spike periods, regime-switching approaches might be necessary [12, 55].

Many approaches for labeling these short-lived yet extreme observations are put forward in literature, ranging from simple thresholds such as *thresholds at fixed price* to more sophisticated thresholds such as *thresholds at matching kurtosis* [51]. Regardless of what approach is favourable all things considered, treating a price series with any validated threshold approach will generally lead to a price series that is significantly favorable to work with, compared to the original price series [51].

Figure 9 shows for the period from 2015 through 2020, the labeling of 414 occurrences of extreme price with *thresholds at fixed price*. Dots indicate prices with an absolute value above 100 and a vertical line indicates the first of consecutive occurrences. Figure B.6 shows the occurrence of extreme prices for the German ID3 series. During all years, extreme prices were slightly less pronounced for the German market, with a total of 193 occurrences. Occurrences where high or low prices are reached steadily—without any extreme jumps—might be more explainable than true price spikes characterized by more instantaneous jumps.

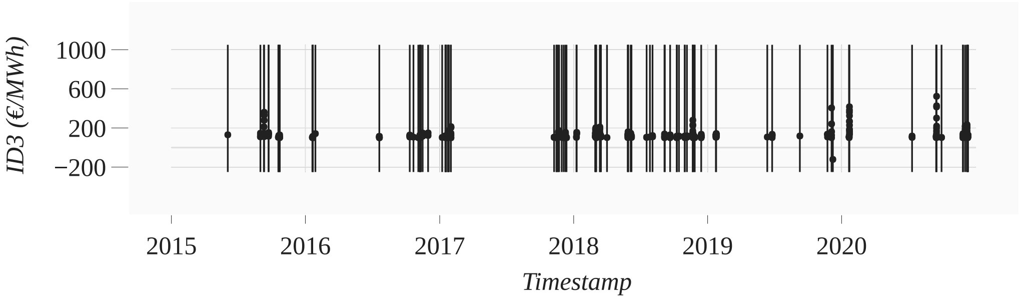


Figure 9: Occurrence of extreme prices in the Dutch ID3 index. *Data from EEX [28].*

Negative prices In markets that demand a considerable amount of rebalancing close to delivery, grid operators may be forced to occasionally offer prices that are negative to achieve market clearing. That might happen when, for instance, low demand goes alongside unforeseen surges in supply due to unplanned generation from variable renewable energy sources [75]. Contracts with negative prices are not completely uncommon on intraday markets, and are large enough in number and magnitude to push aggregated prices, including the ID3 index, to negative values as well.

Figure 10 shows for the period from 2015 through 2020 a total of 161 occurrences of negative prices in the ID3 index. Aggregating all consecutive negative occurrences results in a total of 31 periods, and the vertical lines indicate the start of each period. The vast majority of negative prices occurred in in 2020, with 133 observations in 21 periods. Similar surges of negative prices were observed across European markets at the time, induced by major imbalances due to COVID-19 restrictions [27]. Figure B.7 shows the occurrence of negative prices for the German ID3 series. During all years, negative prices occurred extremely more often in the German ID3 index, with a total of 1180 occurrences and 344 in 2020, which might suggest that more aggressive rebalancing is necessary for the German market. A factor that might contribute to that effect is that the German energy mix contains considerably more renewable energy. Because of the higher occurrence of negative prices, the German ID3 index is more symmetrically distributed, which is reflected in the normality test shown in Figure B.5.

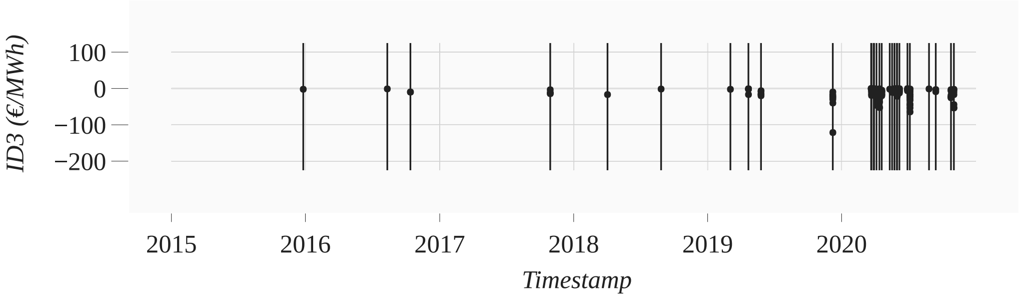


Figure 10: Occurrence of negative prices in in the Dutch ID3 index. *Data from EEX [28].*

3 Data and feature engineering

This chapter presents what types of raw data are available and explains how they are transformed into features in the categories price, calendar, generation, load, and weather. It is concluded that besides a number of base features, many other features should be considered that arise from lags/leads or from market integration. It is concluded that German features should be considered in the full feature set. It is concluded that features should be scaled so that they all have a similar scale. It is concluded that this research should investigate the effects of scaling by means of standardization or normalization, as EPF research provides no preference towards either one. It is concluded that the full feature set should be reduced in size. To that end, this research utilizes a RF regression within a framework of recursive feature elimination to eliminate redundant features from the full feature set.

3.1 Raw data

An increasing amount of high-resolution data regarding energy markets and environmental conditions is available to practitioners. The European Network of Transmission System Operators for Electricity (ENTSO-E), for instance, sustains the 'Transparency Platform', which is part of an ambition to advance the transparent sharing of data with market participants. The platform holds data on the energy market including actual and forecasted load and generation that is hourly or intra-hourly, and updated daily. In addition, there are parties that offer data commercially. Among those is EEX, that offers subscription access to price data of all the European energy markets that they (EPEX SPOT) operate. For this research, ENTSO-E and EEX are the main sources of raw data. Table 1 shows details. All data is retrieved for the period from 2015.07 through 2021.07.

Ind.	Type	Resolution	NL	DE	Source	Public
R_{p1}	Transactions ID	Continuous	•		EEX [28]	
R_{p2}	Price ID (ID3)	Hourly		•	Energy Charts [29]	•
R_{p3}	Price DA (MCP)	Hourly	•	•	EEX [28]	
R_{g1}	Generation* actual	Hourly	•	•	ENTSO-E [30]	•
R_{g2}	Generation forecast DA	Hourly	•	•	ENTSO-E [30]	•
R_{g3}	Generation forecast ID	Hourly		•	ENTSO-E [30]	•
R_{l1}	Load actual	Hourly	•	•	ENTSO-E [30]	•
R_{l2}	Load forecast DA	Hourly	•	•	ENTSO-E [30]	•

Table 1: Raw data

An abundance of data does not mean that forecasting becomes trivial; there remains the need to carefully engineer explanatory variables (*alias* features) to prevent over-fitting and to accurately capture the target value [79]. What is more, features added to an already well-functioning model might not increase—or might *decrease*—performance, and the computational burden of adding explanatory variables can be substantial. Assessing the relevance of features in the full feature set is thus crucial.

3.2 Full feature set

Motivated by literature, various features are included in the full feature set $\mathcal{F}_F$. The set, shown in Table 2, contains features that originate from 28 unique time-series and are assigned into one of five categories. Most features are the result of basic computations on the raw data of Table 1. Whereas raw data can come in any resolution and form, a feature must have a single value per pre-defined time-step, which is equal to one hour in this research.

Two concepts that this research employs to establish the full feature set are lags/leads and market integration. Utilizing lags/leads in time-series analysis amounts to creating many features from a single time-series, where instances are shifted forward/backward in time. Market integration in time-series analysis amounts to using features from a different market than the target market. This has been shown to improve forecasting performance when dynamics of the considered markets are related [63, 103]. Many features in the full feature set are lags/leads, and many have both Dutch (NL) and German (DE) variants.

Ind.	Time-series	Lags/Leads	Raw	NL	DE	No.
F_{p1}	Price ID (ID3)	$\{t_d - 168,\ t_d - 48\}$ & $[t_d - 24,\ t_f]$	R_{p1}, R_{p2}	•	•	23
...	...	$[t_d - 24,\ t_f]$	...	...	...	...
F_{p2}	Price ID (ID3) avg. (last week)		R_{p1}, R_{p2}	•	•	1
F_{p3}	Price ID (ID3) avg. $[t_f - 24,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p4}	Price ID (ID3) avg. $[t_f - 168,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p5}	Price ID (ID3) avg. $[t_f - 5040,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p2}	Price DA (MCP)	$\{t_d - 168,\ t_d - 48\}$ & $[t_d - 24,\ t_f + 12]$*	R_{p3}	•	•	36
...	...	$[t_d - 24,\ t_f + 12]$*	...	...	...	...
F_{p7}	Price DA (MCP) avg. (last week)		R_{p3}	•	•	1
F_{p8}	Price DA (MCP) avg. $[t_f - 24,\ t_f - 1]$		R_{p3}	•	•	1
F_{p9}	Price DA (MCP) avg. $[t_f - 168,\ t_f - 1]$		R_{p3}	•	•	1
F_{p10}	Price DA (MCP) avg. $[t_f - 5040,\ t_f - 1]$		R_{p3}	•	•	1
F_{c1}	Hour of day			•		2
F_{c2}	Day of week			•		2
F_{c3}	Month of year			•		2
F_{c4}	Day of year			•		2
F_{g1}	Generation actual	$[t_d - 24,\ t_f - 1]$	R_{g1}	•	•	21∗3
F_{g2}	Generation forecast DA	$[t_d - 24,\ t_f + 5]$	R_{g2}	•	•	37∗3
F_{g3}	Generation forecast ID	$[t_d - 24,\ t_f]$	R_{g3}		•	37∗3
F_{g4}	Generation error DA	$[t_d - 24,\ t_f - 1]$	R_{g1}, R_{g2}	•	•	21∗3
F_{g5}	Generation error ID	$[t_d - 24,\ t_f - 1]$	R_{g1}, R_{g3}		•	21∗3
F_{l1}	Load actual	$[t_d - 24,\ t_f]$	R_{l1}	•	•	21
F_{l2}	Load forecast DA	$[t_d - 24,\ t_f + 12]$	R_{l2}	•	•	37
F_{l3}	Load error DA	$[t_d - 24,\ t_f - 1]$	R_{l1}, R_{l2}	•	•	21

Table 2: Full feature set $\mathcal{F}_F$

What follows is a description of how the features that represent intraday price are calculated. Appendix D explains the conversions from raw data to all other features in detail.

Because no historical data of the Dutch ID3 index is available, F_{p1} requires manual calculation of the volume-weighted average of individual transactions. Following Equation 1 and the rules given in the official definition [31], the ID3 index is calculated from the transactions in R_{p1}, taking into account

- that the value of the ID3 index is the volume-weighted average price of transactions in the period from 180 to 5 minutes before delivery;
- that cross-border transactions are included;
- that transactions where the bid and ask are from the same party (i.e. self-transactions) are not included;
- and that $\text{ID3}_{d,h} = \text{IDFull}_{d,h}$ in the case that the total volume of transactions within that period does not reach 10MW.

The German ID3 index is available in R_{p2} and thus requires no manual calculations. The features contained in F_{p1} are the lags of one week ago and of two days ago $\{t_d - 168,\ t_d - 48\}$ and all lags from one day ago through the time of forecasting $[t_d - 24,\ t_f]$.

3.2.1 Correlation

Figure E.1 shows the correlation coefficients of the features considered in the full feature set with lag -4. What follows are the most noteworthy correlations. There is a strong positive correlation of the Dutch intraday price with the non-lagged, and 4-hour lagged intraday price of the Dutch and German intraday and day-ahead markets, which demonstrates that price dynamics of these four markets are connected to a substantial degree. Moving on beyond features of price, there is a reasonably strong positive correlation of the Dutch intraday price with the forecasted load as well as with the actual load. Naturally, the correlation of (NL, ID3, 0) with (NL, LOAD_A, 0) is higher than with (NL, LOAD_A, -4), although it is still considerable. That holds true for all variables of which the 4 hour lagged version is the most up-to-date version that can be employed. There is a strong correlation between solar generation and the hour of the day, as well as a reasonably strong correlation between solar and wind generation and month of the year. There are weak correlations between intraday price and renewable generation. However, they might become stronger as renewable generation continues to grow in the energy mix. There are strong positive and negative correlations between exogenous variables, which are not discussed here. It demonstrates that when employing all features of the full feature set, a lot of the same information enters the system. A sophisticated selection of features might be able to capture close to all information with less features.

3.2.2 Transformation

Transformation of the full feature set amounts to scaling and spike reduction. Appendix C contains some background.

Scaling When features have very dissimilar orders of magnitude, models might give more importance to certain features for no reason other than magnitude. Therefore, the full feature set is scaled so that the orders of magnitude of all features are similar. Although neither the field of research nor the utilised features of [91] match that of this research, it demonstrates that the optimal scaling approach can vary considerably for different models. To the best knowledge of the author, EPF literature lacks a comparison of scaling approaches. This research considers the widely used standardization and normalization, and utilizes the scaling that leads to optimal results.

Spikes The technique of *thresholds at absolute value change* is employed to reduce spikes, where the threshold lies at a value of the interquartile range above and below the previous observation. It is not an option to simply eliminate these points as that would introduce discontinuities in the series. Instead, spikes may be replaced by a value that is based on the value of the threshold [90], or on the value(s) of neighboring points [106, 34, 108], for instance. Considerations to somewhat preserve abnormal conditions in training can justify reducing the magnitude of a spike, instead of denying its existence entirely by concealing it between ‘normal’ values [90]. Therefore, this research employs an approach where an identified spike is replaced by the threshold value.

3.3 Candidate feature set

Feature selection is an integral part of price forecasting, and the importance of properly selecting and limiting the number of input features is undisputed in recent literature. Appendix C contains some background. Feature selection is often unsystematic, however, or relies heavily on expert knowledge, or is wrongfully of a linear nature [103, 94]. As a result, research is often based on feature sets where many lags have been removed with little or no justification. Evaluations of such extensive feature sets such as that shown in Table 3 are thus rarely found, although they do exist [59, 2].

This research employs a recursive feature elimination, which is a wrapper-based approach that amounts to repeatedly selecting a feature subset, fitting observations on a model, and reducing the subset based on an evaluation of accuracy. Wrapper-based approaches are capable of evaluating features jointly to infer relative usefulness and are thus equipped to take feature interaction into consideration. In addition, the thoroughness by which wrapper-based approaches consider many subsets makes them capable of finding an optimal candidate feature set. For that initial stage of feature selection, where there is no need for repeated execution and it is the aim to find an optimal set of features, such an approach is particularly suitable and the high computational cost is justified.

4 Point forecasting

This chapter formalizes point forecasts as single-value estimations of future values. Accuracies of point forecasts are evaluated by means of scoring functions. The rMAE is employed so that forecasts can be evaluated across problems. The sMAPE as well as the MAE and RMSE, that differently penalize outliers, accompany the rMAE, as they are often used in other EPF research and provide slightly different notions of accuracy. It is concluded that the (conditional) GW test should be employed to assess statistical significance of forecast superiority, as it indicates which model is estimated to attain higher accuracies in a real time forecasting application more reliably than the (unconditional) DM test. Naive models, regression models, and artificial neural network models for point forecasting are proposed, that vary in terms of complexity and in terms of capability to learn nonlinear relationships.

As a representation of future outcomes that are inherently uncertain, single-value (*alias* point) forecasts are always a simplified representation of reality. Nevertheless, point forecasts underlie decision-making in many practices and within many fields, and are not simply put aside by more information-laden representations, whenever they are available [39].

Assumed is a relationship between the target variable $\boldsymbol{y}$ and the explanatory variables $\boldsymbol{X}$, represented by $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$. Approximation of that relationship is a problem addressed in literature, where fundamental models are proposed that, for instance, simulate the market clearing [72]. However, just like the majority of literature, the research is not troubled by naivety in that respect. The intention is rather that given explanatory variables $\boldsymbol{X}$, a forecasted series $\hat{\boldsymbol{y}}$ is calculated that is as close as possible to the corresponding realized series $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{X}) \approx \boldsymbol{y}$.

4.1 Evaluation

4.1.1 Scoring metrics

A scoring function S is a formalization of what an accurate forecast is. Although many scoring functions are proposed in literature, the squared error $S(a,\ b) = (a-b)^2$, absolute error $S(a,\ b) = |(a-b)|$, absolute percentage error $S(a,\ b) = |(a-b)/b|$, and relative error $S(a,\ b) = |(a-b)/a|$ are employed most often for evaluation in in the field of EPF.

Given a series of T forecasted values of a target variable $\hat{\boldsymbol{y}} = \{\hat{y}_1,\ \ldots,\ \hat{y}_T\}$ and the corresponding series of T realized values of that target variable $\boldsymbol{y} = \{y_1,\ \ldots,\ y_T\}$. It is common practice to represent the 'accuracy' of the point-forecasted series as a single value by means of the average score $\bar{S}$ over all observations, i.e.

$$\bar{S}(\boldsymbol{y},\ \hat{\boldsymbol{y}}) = \frac{1}{T}\sum_{t=1}^{T} S(y_t,\ \hat{y}_t) \tag{2}$$

Optimization in most supervised learning is achieved by tuning parameters based on the minimization of that score. Properties that vary for different scoring functions might encourage practitioners to employ certain scoring functions, some of which are addressed in Appendix F. It is not so much that for every problem, there exists a scoring function that is clearly more suitable than all others; practitioners should instead be aware of the biases that scoring functions might impose on the solution of a problem, and should motivate choices accordingly. For the field of EPF, where published results and drawn conclusions are often based on absolute scoring functions, [64] suggests that it should become the norm to employ the relative mean absolute error (rMAE) (*alias* mean relative absolute error), which normalizes the MAE score to what a reference forecast $\hat{\boldsymbol{y}}^*$ achieves, i.e.

$$\mathrm{rMAE}(\boldsymbol{y},\ \hat{\boldsymbol{y}},\ \hat{\boldsymbol{y}}^*) \coloneqq \frac{1}{T}\sum_{i=t}^{T} \frac{|y_t - \hat{y}_t|}{|y_t - \hat{y}^*_t|} \tag{3}$$

When the rMAE is at least included in evaluation, accuracy can be more easily compared across problems. To the best knowledge of the author, intraday price forecasting literature has not (yet) employed the rMAE scoring function. Therefore, the research employs the mean absolute error (MAE), root mean square error (RMSE), and symmetric mean absolute percentage error (sMAPE) to accompany the rMAE. The MAE and RMSE are used widely and often exclusively in EPF and intraday price forecasting [78, 52, 71]. Like the rMAE, the sMAPE normalizes the score, such that evaluation of accuracy can be more easily compared across problems..

Designing models to perform particularly well in the tails of the distribution is not within the scope of the research. Literature demonstrates that models intended for extreme events should be specifically trained for that task, and are often merely a companion to more generally trained models [47, 104]. For that reason, this research optimizes based on minimization of the MAE, as it scores proportionally, in contrast to the RMSE.

4.1.2 Statistical tests

Superiority Practitioners that have access to two or many forecasts from multiple forecast models are necessitated to assess whether and when to exploit available models. For evaluation of inter-model accuracy, conclusions regarding superiority that are drawn based on testing scores alone lack a proper foundation whether there is any statistical significance, which makes them incomplete and thus unsatisfying [18]. Hence, a number of statistical significance tests have emerged that can complement the scores to ascertain whether superiority can be considered as anything other than a lucky performance.

The Diebold-Mariano (DM) test, introduced in [19], is a statistical test that is often used to that end as it represents the last-presented methodology of most EPF and intraday EPF research [52, 99, 78]. The main intention put forward in [19] includes the evaluation of forecasts, but the succeeding twenty years of literature shows that DM-type tests are often “unfortunately” employed for comparing models in pseudo-out-of-sample environments as well—while more effective methods exist [18].

DM tests are asymptotic Z-tests of a null hypothesis that the mean of the loss differential series is zero, i.e.

$$H_0 \; : \; E\Big(\Delta_t^{A,B}\Big) = E\Big(S(y_t, \hat{y}_t^A) - S(y_t, \hat{y}_t^B)\Big) = 0 \tag{4}$$

where $S(\cdot)$ is a scoring function such as the squared error or absolute error. Thus, DM tests test a hypothesis by means of the parameters of the whole group, β^A and β^B. Hence, DM tests are unconditional; there is consideration for the forecast accuracy within the period of evaluation, but there is no consideration for the expected forecast accuracy in the future. A DM test might therefore favor a forecast from a model with many coefficients above a forecast from a model with only some, even in the case that most coefficients are small and their estimates have enormous sampling variation in finite samples [16]. As practitioners may very well favor the simpler model in such a case, DM tests might point to impractical and thus undesirable conclusions.

The Giacomini-White (GW) test, introduced in [37], addresses that unconditional attribute, among other things, as it tests a hypothesis by means of estimates of the parameters of the whole group based on information up to time t, $\hat{\beta}_t^A$ and $\hat{\beta}_t^B$ [84]. As the accuracies of testing and future forecasting depend on the same parameters, that is a better indicator of future performance. For that reason, the research employs a GW test to assess superiority. A GW test comprises regressions of the difference in the absolute forecast error from competing models over its lag and a constant, i.e.

$$\Delta_i^{A,B} = \alpha + \beta \cdot \Delta_{i-1}^{A,B} + \epsilon_i \tag{5}$$

It tests the null hypothesis that the competing forecasts have equal conditional predictive ability. Simply put, forecast A is favored above forecast B if the forecasted value of the error gap between the two forecasts is non-zero or exceeds a given threshold [32], i.e.

$$H_0 \; : \; \alpha + \beta \cdot \Delta_{i-1}^{A,B} = 0 \tag{6}$$

Residuals The error between the forecasted and realized values (*alias* residuals) $\boldsymbol{\epsilon} = \hat{\boldsymbol{y}} - \boldsymbol{y}$ provides insight in the adequacy of a forecast to capture the characteristics of the target variable. An indication that the vector of forecasted values $\hat{\boldsymbol{y}} = \{\hat{y}_1, \ldots, \hat{y}_T\}$ captures the information contained in the realized values $\boldsymbol{y}$ adequately, is when the residuals vector $\boldsymbol{\epsilon}$ resembles white noise. It can thus be validated whether $\boldsymbol{\epsilon}$ has an expected value of zero; whether its variance is constant over time; whether its values are uncorrelated in time; and whether its values are normally distributed.

The characteristics of residuals are often overlooked, despite the fact that they can indicate when forecasts are unable to capture the characteristics of the original time series fully, and thus inherit characteristics of the original time series [5]. Analysis of residuals is mostly ignored in EPF, although there are examples where it is utilized to evaluate price forecasts of the day-ahead market [53, 109]. This research provides the results of four concise statistical tests to assess whether the four desired characteristics are attained by point forecasting.

4.2 Models

This section presents the considered naive, regression, and ANN point forecasting models, and shortly describes how they obtain forecasts. What values are considered for the hyperparameters is included in Appendix H. The aim of this research is not to find the forecast or model that attains the absolute highest accuracy of the Dutch intraday price. Therefore, this research does not provide an exhaustive overview of forecasting models, but rather discusses why a select number are evaluated. [62] contains a comprehensive evaluation of accuracy for day-ahead price forecasting that includes many of the point forecasting models proposed in recent years.

4.2.1 Naive models

The first type of considered models base their forecast on raw historical observations directly, without learning any causalities, and are thus referred to as 'naive' models. In essence, they represent a minimum-effort solution, and set the bar that forecasts from more sophisticated models should be able to *at least* outperform.

Day-ahead price Considering the high correlation between prices of the Dutch day-ahead and intraday markets, it stands to reason that the market clearing prices (MCP) of the day-ahead market, that are published for hours $h \in \{0, \ldots, 23\}$ of day d after 12:00 on day $d - 1$, can be exploited directly.

NVE.DA The naive model referred to as NVE.DA obtains a forecast of the ID3 price by assuming that it matches the corresponding MCP value. The forecast for delivery on day d and hour h is formally defined as $\hat{y}_t = \text{MCP}_t$. That naive model is also employed in literature [71, 78].

Intraday price A naive approach to obtain a forecasted ID3 value from intraday prices is to simply utilize the most recent ID3 price that is available at the time of forecasting.

NVE.ID The naive model referred to as NVE.ID obtains a forecast of the ID3 price by assuming that it matches the volume-weighted average transaction price in the three hours before the time of forecasting. The forecast is formally defined as $\hat{y}_t = {}_{-4}^{-7}\text{ID}_t$.

Other naive approaches are to utilize different lags, e.g. the 24-hour or 168-hour lagged ID3 price, or to calculate a slightly different aggregated price, e.g. the volume-weighted average transaction price in the most recent 15 minutes of trading. The main issue with the second approach is that the most recent 15 minutes of trading starts 4 hours and 15 minutes before the time of delivery, and might thus not be very representative of especially the transactions that are close to the time of delivery. Besides, for a market that does not always have a plethora of transactions, such a small window might result in sensitivity to the price of abnormal transactions.

Complexity Complexity and computational cost of the considered naive models is low. There is no training involved and there are no hyperparameters. Therefore, the naive models are insensitive to the forecasting procedure.

As most intraday trading is incited by conditions that are unforeseen during day-ahead trading, the day-ahead market clearing price largely defines the intraday ID3 price, which is reflected in high correlations. In the search of forecasts that are able to capture the dynamics of intraday price, forecasts must thus be capable to learn to some extent what sets them apart. The forecast from NVE.DA thus represents the forecast to at least outperform, and this research utilizes the forecasted series of NVE.DA to calculate the rMAE score of other forecasts, i.e. NVE.DA attains an rMAE of 1.00.

4.2.2 Regression models

The second type of considered models base their forecasts on explanatory variables $x_{f,t}$, and estimate how a set of independent variables relates to the target variable of intraday price. The simplest 'linear' regression assumes a linear relationship β between one explanatory variable x_t and the independent variable y_t, i.e. $y_t = \beta \cdot x_t$ and estimates what coefficient $\hat{\beta}$ minimizes the residual error $y_t - \hat{y}_t$ over $t = \{1, \ldots, T\}$ observations. Multiple linear regression extends that to any amount of explanatory variables F, i.e. $y_t = \sum_{f=1}^{F} \beta_f \cdot x_{f,t}$. More sophisticated regression models exist, however.

Least absolute shrinkage and selection operator (LASSO) regression Given a target variable $\boldsymbol{y} = \{y_1, \ldots, y_T\}$ and explanatory variables $\boldsymbol{X} = \{\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_F^\top\}$ where $\boldsymbol{x}_f = \{x_{f,1}, \ldots, x_{f,T}\}$. A least absolute shrinkage and selection operator (LASSO) regression introduces a regularisation term to the cost function of an ordinary least squares regression, i.e.

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 = \sum_{t=1}^{T} \left(y_t - \sum_{f=1}^{F} \beta_f \cdot x_{f,t} \right)^2 + \lambda \sum_{f=1}^{F} |\beta_f| \tag{7}$$

Minimizing the cost function for $\boldsymbol{\beta}$ will find the estimated coefficients, as a function of the regularisation parameter λ, i.e.

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^F}{\arg\min} \left\{ \sum_{t=1}^{T} \left(y_t - \sum_{f=1}^{F} \beta_f \cdot x_{f,t} \right)^2 + \lambda \sum_{f=1}^{F} |\beta_f| \right\} \tag{8}$$

Because of the added 'penalty' term, the problem is now penalized on the sum of the magnitudes of the coefficients. Effectively, this constrains the magnitude of the coefficients to a certain range of values, and a larger regularisation parameter λ leads to tighter constraints on the coefficients. As a consequence of this, some coefficients may become zero, and the number of coefficients that are zero increases as λ increases. In the extreme case that $\lambda = \infty$, all coefficients will be equal to zero, while for $\lambda = 0$, the coefficients will be equal to what would be found when using an ordinary least squares regression.

That effect can be illustrated for a two-dimensional problem. The ordinary least squares term of Equation 7 gives rise to elliptical-shaped contours that are centered at the maximum likelihood estimator. On a single contour lie all sets of coefficients that lead to the same residual sum of squares. The higher the features are correlated, the flatter these ellipses are. At the same time there is the penalty term of Equation 7 that gives rise to a diamond-shaped region that is centered at the origin and whose area shrinks with increasing λ. If we set that region fixed for a certain value of λ, the solution to Equation 8 is the first point where the elliptical contours, when expanding outward, hit the diamond-shaped region. Because that region has sharp corners, the elliptical region has a high chance of hitting such a point first, and because these corners lie on the axes, the coefficient of one of the features is zero whenever that happens.

REG.LASSO The model based on a LASSO regression is referred to as REG.LASSO. That model employs the features in the input feature set $\mathcal{F}_I$ as explanatory variables, i.e. $\boldsymbol{X} \equiv \mathcal{F}_I$, and thus has as many regression coefficients as the number of features in $\mathcal{F}_I$. The forecast is formally defined as

$$\hat{y}_t = \sum_{f=1}^{F} \hat{\beta}_f \cdot x_{f,t} \tag{9}$$

The hyperparameters considered for REG.LASSO are the regularization parameter λ only.

Support vector regression (SVR) An SVR (*alias* support vector machine regression) utilizes the ϵ-insensitive loss function, which does not tolerate errors that exceed ϵ. A transformation function $\phi(\cdot)$ is employed that transforms data points from the original 'input space' to a higher dimensional 'feature space'. It is attempted to construct a linear model in the feature space, that translates to a non-linear model in the input space. If it is not able to achieve this task within the constraints, the inputs are mapped into an ever higher dimensional feature space. Given a target variable $\boldsymbol{y} = \{y_1, \ldots, y_T\}$ and explanatory variables $\boldsymbol{X} = \{\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_T^\top\}$ where $\boldsymbol{x}_t = \{x_{t,1}, \ldots, x_{t,F}\}$, for T observations and F features. The linear function approximation can be denoted as

$$\boldsymbol{y} = f(\boldsymbol{X}) = \sum_{t=1}^{T} w_t \phi_t(\boldsymbol{x}_t) + \gamma = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{X}) + \gamma \tag{10}$$

where $\boldsymbol{\phi}(\cdot)$ is the nonlinear mapping function, $\boldsymbol{w} = \{w_1, \ldots, w_T\}$ is the weight vector, and γ is a bias term. The goal is to find a function, so that for all observations t, the absolute error between the value of the target variable and value of the estimated variable does not exceed ϵ. At the same time, the function should be as flat as possible. Therefore, the weight vector $\boldsymbol{w}$ should be minimized, e.g. by minimizing the norm $\|\boldsymbol{w}\|^2$, given the constraint on the error [45]. As that problem might be infeasible, two positive slack variables, ζ_t and ζ_t^*, are introduced to the problem to allow for relaxed error levels

$$\begin{aligned} \text{min} \quad & R = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{T}\sum_{t=1}^{T}(\zeta_t + \zeta_t^*) \\ \text{s.t.} \quad & y_t - \boldsymbol{w}^\top \boldsymbol{x}_t \le \epsilon + \zeta_t^* \\ & \boldsymbol{w}^\top \boldsymbol{x}_t - y_t \le \epsilon + \zeta_t \\ & \zeta_t, \zeta_t^* \ge 0 \end{aligned} \tag{11}$$

In this problem, the regularization parameter C regulates the trade-off between generalization ability and accuracy, and the error parameter ϵ regulates the tolerance to errors. When solving the problem of Equation 11 by Lagrangian multipliers and dual optimization techniques that are explicitly stated in [8], the solution is found to be

$$f(\boldsymbol{X}) = (\alpha_t - \alpha_t^*)K(\boldsymbol{x}_t, \boldsymbol{x}_s) + \gamma \tag{12}$$

where α_t and α_t^* are Lagrange multipliers. The solution to the problem can be found without considering explicitly the transformation $\boldsymbol{\phi}(\cdot)$ applied to the data, provided that there is a function that returns the scalar product (*alias* kernel function), i.e. $K(\boldsymbol{x}_t, \boldsymbol{x}_s) = \boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\phi}(\boldsymbol{x}_s)$, that computes the similarity of two points. This research employs the widely used Gaussian radial basis function kernel

$$K(\boldsymbol{x}_t, \boldsymbol{x}_s) = \exp \frac{-\|\boldsymbol{x}_t, \boldsymbol{x}_s\|^2}{2\sigma^2} \tag{13}$$

as it can produce complex decision boundaries and adds only a single hyperparameter, the standard deviation σ, to the SVR [65].

REG.SVR The model based on an SVR is referred to as REG.SVR. The hyperparameters considered for REG.SVR are the regularization parameters C and the standard deviation of the Gaussian kernel σ.

Complexity Complexity and computational cost of the considered regression models is moderate. There is training involved that does not require a sophisticated approach and there are a limited number of hyperparameters. Those factors result in the regression models being moderately sensitive to the forecasting procedure. Therefore, this research ensures that stages of training and hyperparameter optimization are systematic.

Many other and more complex regression models exist, such as those based on decision trees or based on principal components. This research considers specifically LASSO regression and SVR as their computational cost is relatively low, and they are among the most investigated models in EPF. Besides that, both REG.LASSO and REG.SVR have a relatively small number of important hyperparameters, thus results are more easily reproducible than with most other models.

4.2.3 Artificial neural network models

A popular group of models in the subspace of computational intelligence is that of artificial neural networks (ANNs). Conceptually based on the learning process of neurons in a human brain, ANNs have been introduced to many fields of research and a plethora of slightly and substantially different networks with unique properties have emerged.

Feedforward neural networks Multilayer perceptron (MLP) neural networks, that are often regarded as prototypical ANNs, have been demonstrated to perform well for EPF [41, 92, 62]. In contrast to recurrent neural networks, these feedforward neural networks have no cycles. In between input and output layers are one or many hidden layers that contain a given number of identical units. Each unit is connected to all units in the next layer and as such constitute a 'fully connected' network. Figure I.1 shows the architecture of an MLP.

The units of an MLP neural network (*alias* neurons) are linear threshold units that take an input vector $\boldsymbol{x} \in \Re^X$ and weight matrix $\boldsymbol{w} \in \Re^{N \times X}$, and utilize the weighted sum and an activation function $\phi(\cdot)$ to calculate their state. For layer $l \in \{0, \ldots, L\}$ the states of all units are given as

$$\boldsymbol{h}^{(l)} = \phi^{(l)}\big(\boldsymbol{w}^{(l)}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}\big) \tag{14}$$

where the first state vector is the input vector, i.e. $\boldsymbol{h}^{(0)} \Leftrightarrow \boldsymbol{x}$, and the last state vector is the output vector, i.e. $\boldsymbol{h}^{(L)} \Leftrightarrow \hat{\boldsymbol{y}}$.

Important considerations in the design of MLPs, as well as for other ANNs, are the amount of hidden layers and the amount of neurons in those layers. What generally holds true is that increasing the number of hidden layers beyond what is optimal for the problem at hand easily induces over-fitting, which goes at the cost of generalizability. Results might improve for the training data, but deteriorate for the testing data. A similar seesaw effect is generally true for the number of neurons; the complexity of the network and therefore the capability to account for more detailed relationships might improve, over-fitting might become a problem. Something that can benefit generalizability is randomly dropping out a certain percentage of neurons during training, i.e. temporarily setting their weight to zero. The weights of other neurons then adjust to account for that which is suddenly lacking. Generally, the network thus becomes less sensitive to individual weights of neurons.

ANN.MLP The model based on an MLP neural network is referred to as ANN.MLP. The hyperparameters considered for ANN.MLP are the number of hidden layers, the number of neurons in each layer, the activation function, and the dropout rate.

Recurrent neural networks Feedforward neural networks can be limited by the fact that the incorporation of temporal dependencies requires additional features. As the number of input features increases, the size of the network will increase rapidly because of the fully connected nature, which might lead to a considerable increase of computional cost. Recurrent neural networks (RNNs) address that limitation as they are of a chain-like nature. Essential to the successes of RNNs are networks with two particular kinds of gated units, namely the long short-term memory (LSTM) unit and the gated recurrent unit (GRU). These gated units ease optimization and reduce learning degeneracies that more traditional RNNs suffer from. The feature that traditional recurrent units lack, and that is shared by both these units, is that instead of replacing the value of activation with a new value for every new input, the new value is added to the previous value with a certain weight. This procedure not only allows the units to maintain important features for longer, but also effectively creates 'shortcut' paths to bypass multiple temporal steps. In an RNN, an index t indicates the position in a sequence and the state of the system $\boldsymbol{h}^{(t)}$ is a nonlinear mapping, i.e. $\boldsymbol{h}^{(t)} = f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)})$.

What follows are the recurrent stages of what is often regarded as a 'typical' LSTM unit, that is, the original unit as proposed in [48] with the addition of a forget gate as proposed in [35]. Peephole connections as proposed in [36] have been neglected as they complicate the learning process [113].

An LSTM unit employs three gates that control the flow of information. Generally, the activation of a gate is approximated by a linear combination of the previous hidden state $h^{(t-1)}$ and the current input $x^{(t)}$, and a nonlinear activation function $\phi(\cdot)$, i.e.

$$\boldsymbol{g}^{(t)} = \phi\big(\boldsymbol{W} \cdot \big[\boldsymbol{h}^{(t-1)},\ \boldsymbol{x}^{(t)}\big] + \boldsymbol{b}\big) \tag{15}$$

The forget gate is designed to decide what portion of the previous cell state $c^{(t-1)}$ is remembered in the current cell state $c^{(t)}$. The previous hidden state $\boldsymbol{h}^{(t-1)}$ and the current input $\boldsymbol{x}^{(t)}$, as well as the accompanying weight vectors that are combined into a matrix $\boldsymbol{W}_f$, pass through the gate. By means of a sigmoid activation function, values between 0 (forget) and 1 (keep) are obtained for the activation of the forget gate. In Equation 15, $\boldsymbol{g}^{(t)} \Leftrightarrow \boldsymbol{f}^{(t)}$, $\boldsymbol{W} \Leftrightarrow \boldsymbol{W}_f$, and $\boldsymbol{b} \Leftrightarrow \boldsymbol{b}_f$. The input gate is designed to decide what portion of the current candidate cell state is added to the current cell state. By means of a sigmoid activation layer, values are obtained for the activation of the input gate. In Equation 15, $\boldsymbol{g}^{(t)} \Leftrightarrow \boldsymbol{i}^{(t)}$, $\boldsymbol{W} \Leftrightarrow \boldsymbol{W}_i$, and $\boldsymbol{b} \Leftrightarrow \boldsymbol{b}_i$. By means of a hyperbolic tangent activation function, values between -1 and 1 are obtained for the current candidate cell state, i.e.

$$\tilde{\boldsymbol{c}}^{(t)} = \tanh\big(\boldsymbol{W}_c \cdot \big[\boldsymbol{h}^{(t-1)},\ \boldsymbol{x}^{(t)}\big] + \boldsymbol{b}_c\big) \tag{16}$$

The current cell state $\boldsymbol{c}_t$ is determined by the previous cell state $\boldsymbol{c}^{(t-1)}$ and activation of the forget gate, and the current candidate cell state $\tilde{\boldsymbol{c}}_t$ and activation of the input gate. The candidate memory content is thus added to the existing memory content, i.e.

$$\boldsymbol{c}^{(t)} = \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)} + \boldsymbol{i}^{(t)} \odot \tilde{\boldsymbol{c}}^{(t)} \tag{17}$$

The output gate is designed to decide what portion of the cell state is output. By means of a sigmoid activation layer, values are obtained for the activation of the output gate. In Equation 15, $\boldsymbol{g}^{(t)} \Leftrightarrow \boldsymbol{o}^{(t)}$, $\boldsymbol{W} \Leftrightarrow \boldsymbol{W}_o$, and $\boldsymbol{b} \Leftrightarrow \boldsymbol{b}_o$. Finally, the activation of the output gate is multiplied by the hyperbolic tangent of the cell state, and the current hidden state of the LSTM unit obtains a value, i.e.

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} \odot \tanh(\boldsymbol{c}^{(t)}) \tag{18}$$

A variation and simplification of the LSTM unit is the GRU. Unlike an LSTM unit, a GRU does not have a separate cell state and, thereby, no controlled exposure of its memory content. The hidden state is exposed without any control. An LSTM—which has input, forget, *and* output gates—can control the amount of content being added to the cell state independent of the forget gate, while a GRU—which has reset and update gates only—can control the portion of the previous hidden state that is added to the current candidate hidden state, but has no way to control the portion of the current candidate hidden state that is added to the current hidden state.

The reset gate is designed to decide what portion of the previous hidden state is added to the current candidate hidden state. In Equation 15, $\boldsymbol{g}^{(t)} \Leftrightarrow \boldsymbol{r}^{(t)}$, $\boldsymbol{W} \Leftrightarrow \boldsymbol{W}_r$, and $\boldsymbol{b} \Leftrightarrow \boldsymbol{b}_r$. The current candidate hidden state is obtained by

$$\tilde{\boldsymbol{h}}^{(t)} = \tanh\big(\boldsymbol{W}_h \cdot \big[\boldsymbol{r}^{(t)} \odot \boldsymbol{h}^{(t-1)},\ \boldsymbol{x}^{(t)}\big] + \boldsymbol{b}_h\big) \tag{19}$$

The update gate is designed to decide what portion of the previous hidden state and current candidate hidden state are added to the current hidden state. In Equation 15, $\boldsymbol{g}^{(t)} \Leftrightarrow \boldsymbol{z}^{(t)}$, $\boldsymbol{W} \Leftrightarrow \boldsymbol{W}_z$, and $\boldsymbol{b} \Leftrightarrow \boldsymbol{b}_z$. The current hidden state of the GRU is obtained by a linear interpolation between the previous hidden state and the current candidate hidden state, i.e.

$$\boldsymbol{h}^{(t)} = (1 - \boldsymbol{z}^{(t)}) \odot \boldsymbol{h}^{(t-1)} + \boldsymbol{z}^{(t)} \odot \tilde{\boldsymbol{h}}^{(t)} \tag{20}$$

The superiority of GRU over LSTM networks or vice-versa is dependent on the problem at hand. What is often stated about GRU networks and the 'compactness' of their two-gated design is that their computational cost is lower, but they can have slightly less representational power.

ANN.GRU The model that is based on a GRU network is referred to as ANN.GRU. Similarly to ANN.MLP, the hyperparameters considered for ANN.GRU are the number of hidden layers, the number of neurons in each layer, the activation function, and the dropout rate.

Complexity Complexity and computational cost of the considered ANN models is high. There is training involved that requires a sophisticated approach and there are many hyperparameters. Those factors result in the ANN models being sensitive to the forecasting procedure.

Besides that, a recurrent neural network architecture such as ANN.GRU introduces certain complexities. Firstly, ANN.GRU requires that the input is structured as three-dimensional tensors that incorporate a time-step dimension. This research utilizes a conversion from a two-dimensional structure that is required by the regression models and by ANN.MLP, to a three-dimensional structure. That conversion sorts the two-dimensional structure by time-step, and because dimensions must agree throughout the three-dimensional structure, adds zero-columns for time-steps of features that are not considered. In the following example, the 24-hour lag of the second base feature $f = 1$ is not available, such that $x_{t,1}^{-24} = 0$ for all observations $t \in T$. For features $f \in F$ and observations $t \in T$, the two-dimensional conversion is

$$\begin{bmatrix} x_{0,1}^{-24} & \cdots & x_{0,1}^{-4} & \cdots & x_{0,F}^{-24} & \cdots & x_{0,F}^{-4} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{T,1}^{-24} & \cdots & x_{T,1}^{-4} & \cdots & x_{T,F}^{-24} & \cdots & x_{T,F}^{-4} \end{bmatrix} \to \begin{array}{c} \begin{array}{ccccccccc} & \text{N/A} & & & & & & & \end{array} \\ \begin{bmatrix} x_{0,1}^{-24} & 0 & \cdots & x_{0,F}^{-24} & \cdots & x_{0,1}^{-4} & \cdots & x_{0,F}^{-4} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{T,1}^{-24} & 0 & \cdots & x_{T,F}^{-24} & \cdots & x_{T,1}^{-4} & \cdots & x_{T,F}^{-4} \end{bmatrix} \end{array}$$

That two-dimensional structure is then converted to the three-dimensional structure shown in Figure 11 where the z-dimension represents the time-step (lag/lead). In consideration of that required structure, it becomes clear that difficulties arise when sequential time features and regular features are utilized. In that case, a hybrid model can divide features between different layers [62], although that does not lie in the scope of this research.

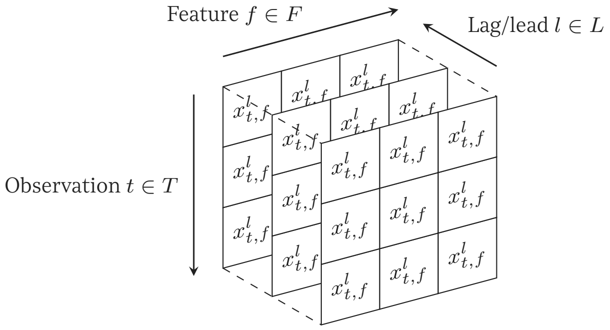


Figure 11: Three-dimensional input structure

5 Interval forecasting

This chapter formalizes interval forecasts as the region between two quantile forecasts or between two expectile forecasts, that are obtained when a bias embodied by the quantile loss function or expectile loss function is purposefully introduced to the models. It is concluded that it is generally desirable that an interval forecast has high reliability, then has high sharpness, and then has high resolution. Scoring functions help to evaluate whether an interval forecast has reliability, sharpness, and resolution. The upper and lower quantile forecasts of an interval are evaluated separately based on the PL, and all intervals are jointly evaluated based on an estimation of the CRPS for a finite grid of equidistant quantile forecasts.

On many levels of decision-making, the incorporation of uncertainty into forecasts is invaluable. Some applications require the full probability of the target variable, for instance, while others require estimation of the target variable in the tail of the distribution [97]. In contrast to a point forecast, an interval forecast can effectively represent the uncertainty that is inherently associated to the estimation of a value that is not yet realized.

An interval forecast consists of upper and lower bounds that together establish the range where the realized value is estimated to lie within, given a certain probability (*alias* nominal coverage rate). There are fundamentally different approaches to calculate bounds, although many are underlied by the distribution of the point forecasting error or by the distribution of the target value. There exist more general regressions than mean and median regressions, such as quantile regression, which has median regression as a special case but can also estimate in the tail of a distribution. In contrast to linear regressions, quantile regressions have no assumptions of a particular parametric distribution of the response variable, nor of a constant variance. Widely used for the bounds of intervals are therefore quantile forecasts. For an interval forecast with a nominal coverage rate of $(1-\tau)$, the bounds then correspond to the α-quantile forecasts where $\alpha \in \{\tau/2,\ 1-\tau/2\}$, i.e.

$$\hat{I}_t = \left[\hat{\zeta}_t^{(\tau/2)},\ \hat{\zeta}_t^{(1-\tau/2)}\right] \tag{21}$$

Given is a series of T realized values $\boldsymbol{y} = \{y_1,\ \ldots,\ y_T\}$. In a quantile regression

$$y_t = \zeta_t^{(\alpha)} + \epsilon_t^{(\alpha)} \tag{22}$$

it is assumed that the α-quantile of the residuals is zero, i.e. $P(\epsilon_t^{(\alpha)} \leq 0) = \alpha$, which is unlike in a linear regression, where it is assumed that the mean of residuals is zero, i.e. $E(\epsilon_t) = 0$. The estimate of the α-quantile $\hat{\zeta}^{(\alpha)}$ can be obtained by sorting the series in ascending order or by solving

$$\min_{\zeta^{(\alpha)}} \quad \sum_{t=1}^{T} \rho^{(\alpha)} \cdot |y_t - \zeta^{(\alpha)}| \tag{23}$$

where the check function $\rho^{(\alpha)}$ handles that penalization is asymmetric, i.e.

$$\rho^{(\alpha)} = \begin{cases} (1-\alpha) & \text{if} \quad y_t < \zeta^{(\alpha)} \\ \alpha & \text{if} \quad \textit{otherwise} \end{cases} \tag{24}$$

The α-quantile lies where $\alpha \cdot T$ realized values are smaller than $\zeta^{(\alpha)}$. From this it can be inferred that the 0.5-quantile represents the median of the series. As the quantile level represents the probability, i.e. the proportion of the observations, that are associated with a quantile, there is an innate restriction that the α_1-quantile must be greater than the α_2-quantile in the case that $\alpha_1 > \alpha_2$. Models that estimate quantiles separately might lead to this restriction being unsatisfied, such that quantiles might cross each other (*alias* quantile crossing) with an invalid distribution as a result. The problem of quantile crossing can be a addressed by simply reordering of quantiles [107], or by introducing constraints to the optimization problem [14]. The latter significantly increases complexity and computational cost. As this research focuses on a relatively limited set of quantiles, quantile crossings are not addressed. However, as the number of estimated quantiles increases, such that distances between quantiles decrease, the problem of quantile crossing becomes more severe [57]. Therefore, this research does evaluate interval forecasts on the number of quantile crossings.

Closely related to quantiles are expectiles, where the L^1 (absolute distance) term in Equation 23 is replaced by an L^2 (quadratic) term and the estimated α-expectile $\hat{\eta}^{(\alpha)}$ is obtained by solving

$$\min_{\eta^{(\alpha)}} \quad \sum_{t=1}^{T} \rho^{(\alpha)} \cdot (y_t - \eta^{(\alpha)})^2 \tag{25}$$

The 0.5-expectile represents the mean of the series. A quantile is not affected when values in the series change within the range that lies above the quantile, i.e. $[\zeta^{(\alpha)},\ \infty)$; transformation of the right half of the distribution does not affect the 0.5-quantile, nor any lower quantiles. That is fundamentally different in the case of an expectile, however, as transformation of any part of the distribution affects *all* expectiles [97, 112].

While numerically finding quantiles requires linear optimization [58], finding expectiles requires quadratic optimization [80], which is something that can become more demanding in terms of computational cost. Although the problem of quantile crossing happens similarly for expectiles (*alias* expectile crossing), it occurs less often and thus requires less attention [105, 57].

5.1 Evaluation

Evaluation of point forecasts is based around the deviation between forecasted values and realized values. It is not surprising therefore, that evaluation of interval forecasts is based around the deviation between forecasted intervals and realized intervals. This is not always self-evident however, as there is an intricate interplay between the coverage of an interval and its width.

5.1.1 Reliability and sharpness

The two main concepts that formalize the desires for interval forecasts are reliability (*alias* calibration) and sharpness (*alias* resolution). Reliability is concerned with the desire that, in essence, realized values should be indistinguishable from random draws of the estimated probability distribution, or else a systematic bias is introduced. Given a series of T forecasted quantile values $\hat{\boldsymbol{y}}^{(\alpha)} = \{\hat{y}_1^{(\alpha)},\ \dots,\ \hat{y}_T^{(\alpha)}\}$ and a series of realized values $\boldsymbol{y} = \{y_1,\ \dots,\ y_T\}$. It is calculated whether the realized value was smaller than the forecasted quantile value, i.e. whether it was “covered”

$$\xi_t^{(\alpha)} = \begin{cases} 1 & \text{if} \quad y_t < \hat{\zeta}_t^{(\alpha)} \\ 0 & \text{if} \quad \textit{otherwise} \end{cases} \tag{26}$$

for $t \in \{1, \dots, T\}$. The expected value of the binary sequence that arises $a_k^{(\alpha)} = E(\boldsymbol{\xi}^{(\alpha)})$ and the deviation from perfect reliability (*alias* probabilistic bias) $b_k^{(\alpha)} = \alpha - a_k^{(\alpha)}$, are an indication of the reliability of a series of forecasted quantile values and are easily visualized in a reliability diagram. Most diagrams in literature show the empirical coverage rate against the nominal coverage rate or the deviation from perfect reliability against the nominal coverage rate. In these diagrams, perfect reliability is given by the diagonal line $y = x$ and horizontal line $y = 0$, respectively. In any case, a reliability diagram is a summary of the calibration assessment of several quantiles or intervals so that with a single glance, one can see whether a method tends to systematically under- or overestimate the uncertainty. Judgment whether the quantiles or intervals attain satisfactory coverage is not an absolute truth but for practitioners to assess per case [85]. Sharpness is concerned with the desire that the more concentrated an interval is, at constant reliability, the better. Intuitively, sharpness is thus concerned with the width of distribution. The width of the interval between the two quantile values with nominal coverage rate $(1 - \alpha)$ can be defined as

$$\delta_t^{(\alpha)} = \hat{\zeta}_t^{(1-\alpha/2)} - \hat{\zeta}_t^{(\alpha/2)} \tag{27}$$

A measure of overall sharpness is the expected value $E(\boldsymbol{\delta}^{(\alpha)})$. It is noted that in real world applications, conditional heteroscedasticity may lead to substantial variability in the width of intervals; simply employing the average width might then be an inadequate characterization of sharpness [40]. Although sharpness and resolution are often mentioned in a single breath, they may be distinguished; sharpness is concerned with the average width of intervals, and resolution is concerned with their width variability. Interval forecasts that have low variability in width, i.e. low resolution, typically originate from simpler models, while more advanced models are able to incorporate varying confidence.

In general, reliability is regarded as the most important desire; an interval forecast with high reliability and low sharpness may have wide intervals, it is still more dependable than the unjustly concentrated interval forecast with low reliability and high sharpness. Of two forecasts with similar reliability and similar sharpness, the one with higher resolution is favored.

5.1.2 Scoring metrics

Just like the scores for point forecasts, there exist scores for quantile forecasts. The widely used quantile scoring function (*alias* pinball loss function)

$$\mathrm{QS}(\hat{\zeta}_t^{(\alpha)}, y_t, \alpha) = \begin{cases} (\hat{\zeta}_t^{(\alpha)} - y_t) \cdot (1-\alpha) & \text{if} \quad y_t \leq \hat{\zeta}_t^{(\alpha)} \\ (y_t - \hat{\zeta}_t^{(\alpha)}) \cdot \alpha & \text{if} \quad \textit{otherwise} \end{cases} \tag{28}$$

is asymmetric in the sense that it penalizes differently when the realized value is smaller or greater than the forecasted quantile value. For quantiles below the middle-quantiles, a higher penalization is for realized values that are smaller than the forecasted quantile, and vice versa for quantiles above the middle-quantile.

Unlike the quantile score, that considers a specific point in the distribution, the continuous ranked probability score (CRPS) considers the distribution as a whole, and requires no predefined classes such as quantiles. The CRPS is a quadratic metric for the difference between a theoretical cumulative distribution and an empirical cumulative distribution, and is thus formally defined for a distribution function $F(x)$

$$\mathrm{CRPS}(F, y_t) = \int_{\Re} (F(x) - \mathbf{1}\{x > y_t\})^2 \, dx \tag{29}$$

which can be equivalently denoted as a scaled integral of the quantile score over all quantiles

$$\mathrm{CRPS}(\hat{\zeta}_t^{(\alpha)}, y_t) = 2 \cdot \int_0^1 \mathrm{QS}(\hat{\zeta}_t^{(\alpha)}, y_t, \alpha) \, d\alpha \tag{30}$$

An approximation of the score can be given if one has quantile forecasts for a finite grid of quantiles $\{\alpha_i, \ \ldots, \alpha_I\}$ that are equidistant [11]. Given forecasted quantile values for a total of H equidistant quantiles $\hat{\zeta}_t^{(\boldsymbol{\alpha})} = \{\hat{\zeta}_t^{(\alpha_1)}, \ \ldots, \ \hat{\zeta}_t^{(\alpha_H)}\}$

$$\mathrm{CRPS}(\hat{\zeta}_t^{(\boldsymbol{\alpha})}, y_t) \approx \frac{2}{I} \cdot \sum_{i=1}^{I} \mathrm{QS}(\hat{\zeta}_t^{(\alpha_i)}, y_t, \alpha_i) \tag{31}$$

The CRPS is thus intrinsically an aggregate metric for all target quantiles, while the pinball loss is calculated for each target quantile individually before non-compulsory averaging [49].

5.2 Models

This section presents the considered naive, regression, and ANN interval forecasting models, and shortly describes how they arrive at a forecast. What values are considered for the hyperparameters is included in Appendix H.

5.2.1 Naive models

NVE.Q.DA In the extension of NVE.DA is a naive model for quantile forecasts, referred to as NVE.Q.DA. The residual errors $\boldsymbol{\epsilon}$ are calculated between the forecasted series $\hat{\boldsymbol{p}}$ and realized series $\boldsymbol{p}$ of the in-sample period, i.e. $\boldsymbol{\epsilon} = |\hat{\boldsymbol{p}} - \boldsymbol{p}|$. From this series of T residual errors, the quantile value is calculated, i.e. $\mathcal{Q}^{(\alpha)}(\boldsymbol{\epsilon})$, that is by definition larger than $\alpha \cdot T$ residual error values. Because it is the absolute error, the interval forecast of nominal coverage τ is then obtained by adding (and subtracting) the quantile value of the residual errors to (and from) the out-of-sample forecasted values. The forecast is formally defined as $\hat{q}_t^{(\alpha)} = \hat{p}_t + \text{sign}(\tau - 0.5)\mathcal{Q}^{(\tau)}(\boldsymbol{\epsilon})$.

NVE.Q.PNT A second naive model, referred to as NVE.Q.PNT employs the exact same approach as NVE.Q.DA, but bases itself on the forecast of the point forecasting model that attains the highest accuracy. This model can thus not be regarded as truly naive when a sophisticated point forecast must be provided. Under the assumption that such point forecasts are available, NVE.Q.PNT can be regarded as a naive interval forecasting model.

5.2.2 Regression models

Quantile regression appeals because it can describe the relationship between the price and independent variables not only on the mean, but also on the tails of the conditional price. Because a set of coefficients for each quantile is obtained, asymmetric effects of the independent variables can be investigated, which might bring insight on whether features affect the price differently at different price levels. In quantile regression, that is reflected in the cost function, which is minimized as

$$\hat{\boldsymbol{\beta}}^{\tau}(\lambda) = \underset{\boldsymbol{\beta}^{\tau} \in \mathbb{R}^F}{\arg\min} \left\{ \sum_{t=1}^{T} \Big(\tau - [p_t \le \boldsymbol{\beta}^{\tau}\mathbf{x}_t]\Big)\Big(p_t - \boldsymbol{\beta}^{\tau}\boldsymbol{x}_t\Big) \right\} \tag{32}$$

Contained in the Iverson bracket $[p_t \le \boldsymbol{\beta}^{\tau}\mathbf{x}_t]$ is the statement whether the true price value is lower than or equal to the estimated price value, i.e. whether there is a positive or negative error. Changing τ changes the ratio of the penalty terms τ and $\tau - 1$, and thus changes how severe over-predictions are penalized compared to under-predictions. It then easily becomes clear that for $\tau = 0.5$, the quantile regression is equivalent to the linear regression, for $|\tau|/|\tau - 1| = 1$.

A regression model for interval forecasting utilizes quantile regression averaging, as introduced in [81]. Unlike in regular quantile regression, not the features $\boldsymbol{x}_t = [x_t^1, \ldots, x_t^F]$ but forecasts $\boldsymbol{x}_t = [\hat{p}_t^1, \ldots, \hat{p}_t^M]$ are utilized as explanatory variables to arrive at a forecast. Selecting the more valuable point forecasts by eliminating those that are redundant by L^1-norm regularization can improve accuracy [101].

REG.Q.QRA However, this research utilizes a small number of point forecasts such that regularized quantile regression averaging is not required. Therefore, this research considers only a model based on quantile regression averaging model without regularization, referred to as REG.Q.QRA.

5.2.3 Artificial neural network models

ANN.Q.MLP The considered ANN model for point forecasting, ANN.MLP, is employed similarly for interval forecasting. To that end, the same architecture is utilized as shown in Figure I.1, but now with an output layer that contains as many nodes as quantiles, i.e. $\boldsymbol{h}^{(L)} = \hat{\boldsymbol{y}} = \{\hat{q}^{(\boldsymbol{\alpha})}\}$. That network is then trained with the quantile loss function of Equation 28, where each output node receives a unique α_l. The model is referred to as ANN.Q.MLP.

6 Forecasting procedure

This chapter establishes a forecasting procedure that contains sub-procedures of initialization, calibration, and exploitation. In initialization, the size of training and testing windows of the rolling window estimation are investigated. Also, the candidate feature set is reduced by means of an RF regression within a framework of recursive feature elimination. Although optimality is found at approximately 15 features, a reduced candidate feature set of 100 features is employed as input to more refined feature selection. In calibration, model hyperparameters are optimized by training models within a framework of TPE. It is investigated how performance holds up when forecasting multiple steps ahead.

This research establishes a forecasting procedure to systematically execute an out-of-sample test with an 'initialization' procedure, a 'calibration' procedure, and an 'exploitation' procedure. For reliable evaluation of accuracies it is vital that the procedure does not interact with observations used in exploitation before exploitation. Hence, this research divides the four-year period from 2016 through 2019 into two two-year periods, so that the period from 2016 through 2017 can be used unhesitatingly for initialization, and the period from 2018 through 2019 can be used for exploitation. A schematic walkthrough of the forecasting procedure is shown in Figure 12.

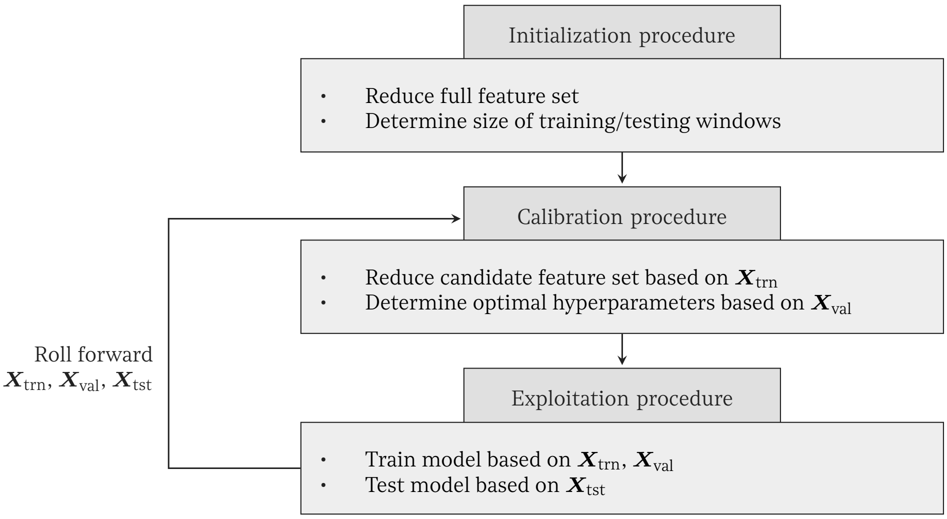

Initialization procedure
• Reduce full feature set
• Determine size of training/testing windows
Calibration procedure
• Reduce candidate feature set based on $\boldsymbol{X}_{\text{trn}}$
• Determine optimal hyperparameters based on $\boldsymbol{X}_{\text{val}}$
Roll forward
$\boldsymbol{X}_{\text{trn}}$, $\boldsymbol{X}_{\text{val}}$, $\boldsymbol{X}_{\text{tst}}$
Exploitation procedure
• Train model based on $\boldsymbol{X}_{\text{trn}}$, $\boldsymbol{X}_{\text{val}}$
• Test model based on $\boldsymbol{X}_{\text{tst}}$

Figure 12: Schematic walkthrough of the forecasting procedure.

In machine learning procedures, it is good practice to split observations in a training set, validating set, and testing set. The training set is used to fit models, and the validating set is used as guidance. The testing set does not partake in that process, so that it can be used to estimate real-world accuracy. In contrast to fixed split approaches, cross-validation approaches use multiple splits such that observations can be used more than once. Cross-validation splits are particularly useful when data is scarce, and they can give a better idea of the variability and robustness of a model fit than fixed splits. Rolling cross-validation approaches are suitable when observations are dependent. They prevent, for instance, that training sets lie after testing sets, to prevent that models can peek into the future.

In rolling cross-validation, a model is fit using all observations that lie within a training window W_{trn}, before it is used to forecast the observations that lie within a testing window W_{tst}. The windows are then 'rolled' forward one or multiple steps. In essence, a rolling cross-validation is similar to what forecasting procedures in practice are like; observations in the testing set are for delivery times that lie in the future and that are truly unknown, and windows are rolled forward as a new realized observation becomes available. In a rolling *window* cross-validation, both the head and tail of the training window advance so that the training set does not grow, while in a rolling *origin* cross-validation, only the head of the training window advances so that the training set grows. This research employs the former, because utilizing more observations for training does not necessarily result in more accurate models [33]. That is in compliance with recent research in the field of EPF [67, 3, 78].

6.1 Initialization procedure

The initialization procedure comprises stages that investigate the reduction of the full feature set and that investigate the sensitivity of accuracies to sizes of training and testing windows.

6.1.1 Stage A: Candidate feature set

The initialization procedure commences with a stage that reduces the full feature set. A RF regression is fit repeatedly to the initialization period, and after every iteration, the least important features are eliminated. It eliminates 50 features per iteration to come to 500 features, then eliminates 10 features per iteration to come to 100 features, and then eliminates 1 feature per iteration to come to 50 features. The stage of recursive feature elimination is executed in a rolling cross-validation. The model is fitted on $N_{\text{trn}} = 12 \cdot (30 \cdot 24)$ (approx. 12 months) observations to forecast $N_{\text{tst}} = 12 \cdot (30 \cdot 24)$ (approx. 1 month) observations. Scores are averaged over these folds before elimination.

6.1.2 Stage B: Training and testing windows

Although rolling cross-validation is considered as one of the standard approaches in out-of-sample tests for time-series forecasting and EPF, the convention of arbitrarily choosing the size of the rolling window, i.e. the number of observations in the training set N_{trn}, to some value between 6 and 12 months is criticized [50, 33, 70]. Findings that forecast accuracy varies considerably for varying sizes of the training window support the statement that different sizes should be considered. To the best knowledge of the author, there is no attention for this topic within the field of intraday EPF. For that reason, this research pays attention to the size of the training window, and incorporates a stage of training window sensitivity in the initialization procedure.

Research on EPF generally employs cross-validations where models are refit every hour, i.e. $N_{\text{tst}} = 1$. It is indeed plausible that the highest accuracies are attained when models are fit with observations that are as close as possible to the target observation. However, it is unclear how sensitive accuracies are to widening this gap. For that reason, this research pays attention to the size of the testing window, and incorporates a stage of testing window sensitivity in the initialization procedure.

This research evaluates all models with initial hyperparameters repeatedly on the initialization period by means of rolling window cross-validation, where every iteration employs a different size for the training window, i.e. $\boldsymbol{N}_{\text{trn}} = \{2 \cdot (24 \cdot 30),\ 6 \cdot (24 \cdot 30),\ 12 \cdot (24 \cdot 30)\}$ (approx. 2 months, 6 months, and 12 months respectively), as well as different sizes of the testing window, i.e. $\boldsymbol{N}_{\text{tst}} = \{24,\ 24 \cdot 7,\ 1 \cdot (24 \cdot 30)\}$ (approx. 1 day, 1 week, and 1 month).

6.2 Calibration procedure

Because the formation of energy price is subject to many different factors that might evolve over time, elements in the forecasting procedure that are based on the assumption that parameters are constant, may lead to accuracies turning sub-optimal over time. For that reason, this research establishes the calibration procedure that is executed before the exploitation procedure. The calibration procedure comprises stages that reduce the candidate feature set and that find optimal hyperparameters based on the observations in W_{trn} and W_{val}.

Research on EPF generally identifies a single set of features and a single set of hyperparameters for every model that are deemed as optimal, and that are utilized throughout the out-of-sample test. Instead of employing a single calibration, this research employs repeated calibration, which means that each exploitation is preceded by a calibration. Neither a single nor repeated calibration are expected to attain accuracies that are truly optimal, because only observations can be used that are available at the time of forecasting. It is expected, however, that repeated calibration attains higher accuracies than a single calibration.

6.2.1 Stage A: Input feature set

The calibration procedure commences with a stage that reduces the candidate feature set $\mathcal{F}_C$, which is very similar to the stage of reducing the full feature set $\mathcal{F}_F$ in the initialization procedure. The candidate feature set is reduced from 50 features to 15 features by eliminating 1 feature per iteration, to obtain the input feature set $\mathcal{F}_I \subseteq \mathcal{F}_C$.

To prevent that the forecasting procedure over-fits the selection of features to the RF regression model that is used in recursive feature elimination, the features of the input feature set are joined by the $[t_d - 24, t_f]$ lags of the intraday ID3 index and the $[t_d - 24, t_d + 9]$ lags and leads of the day-ahead market clearing price. Those features represent all intraday and day-ahead prices of the 24 hours before and after the time of delivery, that are available for forecasting. Together with the optimized input feature set, they represent a reliable feature set.

6.2.2 Stage B: Hyperparameters

The calibration procedure continues with a stage that finds optimal hyperparameters. A model is fit to the calibration set, and every iteration employs a different set of hyperparameters. This stage is performed with $N_{\text{trn}} = 12 \cdot (30 \cdot 24)$ (approx. 12 months) and $N_{\text{tst}} = 1 \cdot (30 \cdot 24)$ (approx. 1 month).

It is the objective to find the set of hyperparameters $\boldsymbol{h}^\star = \arg\min_{\boldsymbol{h} \in \mathcal{H}} S(\boldsymbol{h})$, that is optimal given an objective scoring function $S(\cdot)$, and given all possible hyperparameters $\mathcal{H}$. It can be a challenge, because the score as a function of a hyperparameter set, i.e. $S(\boldsymbol{h})$, is a black box function. Therefore, the search for $\boldsymbol{h}^\star$ is restricted to repeatedly evaluating the score for hyperparameter sets $\boldsymbol{h}_i \in \mathcal{H}$. Grid-search is among the algorithms that search exhaustively and in an unguided way, and is effectively applied in literature [17], although the computational cost can be substantial for a fine-grain search space. Enter informed search algorithms, that transcend the naivety of grid searches or random searches by passing on past results through the optimization procedure. In Bayesian search algorithms, a surrogate function is responsible for updating the prior probability $P\big(S(\boldsymbol{h})\big)$ with a sample $\boldsymbol{h}$ and its score $y = S(\boldsymbol{h})$ to get a better posterior probability $P\big(y|\boldsymbol{h}\big)$. An acquisition function is responsible for guiding the sampling process to where likeliness of finding the optimal solution is highest.

The surrogate function in a tree-structured parzen estimator TPE does not model $P\big(y|\boldsymbol{h}\big)$ directly, but instead uses the probability of $\boldsymbol{h}$ given y, the probability of y, and Bayes' rule to calculate

$$P\big(y|\boldsymbol{h}\big) = \frac{P\big(\boldsymbol{h}|y\big) \cdot P\big(y\big)}{P\big(\boldsymbol{h}\big)} \tag{33}$$

At the core are two kernel distribution functions $l(\boldsymbol{h})$ and $g(\boldsymbol{h})$, that are often chosen as Gaussian [10], and that are based on the samples that yield a score above or below a pre-defined threshold value y^*, respectively

$$P\big(\boldsymbol{h}|y\big) = \begin{cases} l(\boldsymbol{h}) & \text{if} \quad y < y^* \\ g(\boldsymbol{h}) & \text{if} \quad y \geq y^* \end{cases} \tag{34}$$

The acquisition function in a TPE is based on expected improvement

$$EI_{y^*}(\boldsymbol{h}) = \int_{\infty}^{y^*} (y^* - y) P\big(y|\boldsymbol{h}\big) dy \tag{35}$$

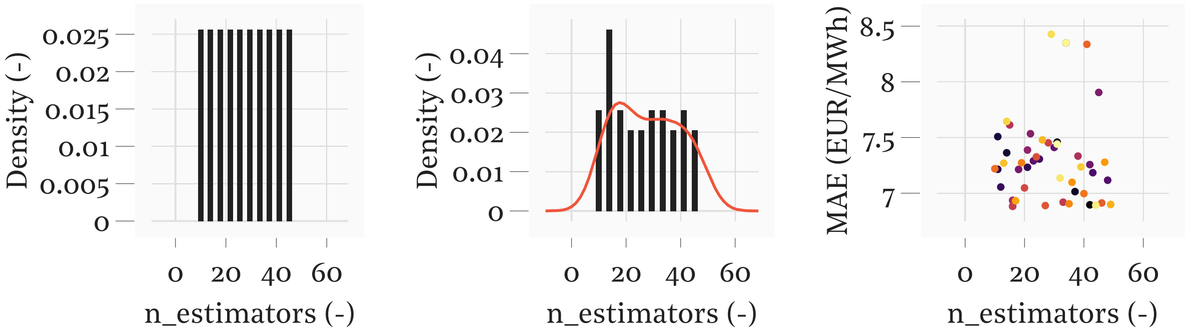
and following the derivations that are explicitly stated in Bergstra et al. [10], proportionality is found as

$$EI_{y^*}(\boldsymbol{h}) \propto \left(\gamma + \frac{g(\boldsymbol{h})}{\ell(\boldsymbol{h})} (1 - \gamma) \right)^{-1} \tag{36}$$

where $\gamma = P\big(y < y^*\big)$. Thus, for the expected improvement to grow, the term $g(\boldsymbol{h})/l(\boldsymbol{x})$ must shrink, such that a hyperparameter set $\boldsymbol{h}$ is chosen that has a high probability under $l(\boldsymbol{h})$ and low probability under $g(\boldsymbol{h})$. Because $l(\boldsymbol{h})$ is a distribution, hyperparameters that are drawn from it are likely in the neighborhood of the maximum improvement, but not exactly equal to it. In addition, selected hyperparameters might not actually improve performance as the surrogate function is merely an estimate of the objective scoring function. The expected improvement thus guides the algorithm in a way that balances exploration and exploitation.

TPE performs the optimization that is described above for all hyperparameters separately, and thus has no ability to account for interdependencies that might exist between hyperparameters. Potentially, this could lead to a hyperparameter set that is deemed optimal, although it might be slightly different than the truly optimal set identified by exhaustive search algorithms. This research still employs a Bayesian search with TPE, as many hyperparameters needs to be reconsidered many times, such that an exhaustive search would result in an excessive computational cost.

As an example, the results of this stage for a RF regression are shown in Figure 13. For one hyperparameter and 50 iterations of the TPE, Figures 13a and 13b show the distribution of the input, and the distribution of the search, respectively. Figure 13c shows the score. Color indicates the iteration; the darker, the later the iteration. The distributions shown in Figures 13a and 13b show that the TPE search does not perform a uniform search, as it evaluates some values more often than others. There is a bias towards the region that is deemed as optimal with more certainty after each iteration.



(a) Input distribution

(b) Search distribution

(c) Search scores

Figure 13: Hyperparameter optimization with TPE for RF regression

6.3 Exploitation procedure

The exploitation procedure comprises a stage that obtains forecasts for the observations in W_{tst}.

6.3.1 Stage A: Out-of-sample forecast

After the initialization procedure and the first calibration procedure, the first exploitation procedure commences. The exploitation procedure is where a model is exploited to obtain a forecast. It employs the input feature set $\mathcal{F}_I$ and the hyperparameter set decided upon in the most recent calibration procedure.

Exploitation complexity is one of the factors that separates the considered models. Exploitation complexity of the naive models is low, as there is no training or testing involved. Exploitation complexity of the regression models is moderate, as they require training and testing, although training is very straightforward. Exploitation complexity of the ANN models is high, as they require training and testing, and training is complicated. Attained accuracies are therefore not only sensitive to the utilized features and hyperparameters, but also to the approach of training.

This research utilizes the validating set, that is also utilized in the stage of hyperparameter optimization, to improve reliability of training ANN models. After every epoch, i.e. every pass through the training set, the ANN model is exploited on the validating set, to check whether gains in accuracy on the training set are due to overfitting. This concept is referred to as early stopping [111]. On the basis of initialization set it is found that and an early stopping criterion that demands an improvement of the MAE that is at least 0.1 €/MWh over every 25 epochs leads to reliable training. With those parameters, training prevents accuracy gains for the training set without accuracy gains for the validating set; training is patient enough that it progresses through local minima; and training is not unnecessarily long.

6.4 Walkthrough

A detailed walkthrough of the forecasting procedure is shown in Table G.1. The pseudocode walkthrough distinguishes the start of the out-of-sample set *start*, the end of the out-of-sample set *end*, the forecasting model *model*($\boldsymbol{h}$) with hyperparameter set $\boldsymbol{h}$, the candidate feature set $\boldsymbol{X}$, the number of observations N, the window W, the observation t, the number of observations before re-calibration T, the iteration i, and the number of iterations I.

6.5 Intricacies/Experiments

The level of intricacy that a forecasting procedure has is not a judger of quality per se. Especially not upon elucidation that and why intricacies are awarely ignored [52]. The forecasting procedure, as presented up to this point, provide enough of a framework to evaluate the overall accuracies attained by the different models. Nevertheless, there remain certain intricacies that might benefit the practical use, accuracy, and/or understanding of the forecasting procedure. The following addresses three such intricacies and indicates why they are within the scope of this research.

6.5.1 Multi-step-ahead horizon

This intricacy lies within the scope of this research because earlier intraday research does not investigate how accuracy holds up for the more practically applicable multi-step-ahead horizons. Besides that, multi-step-ahead forecasts are required for the operational context of the simulation discussed in Chapter 8.

The benefits of extending the forecasting procedure to include multi-step-ahead forecasting become evident in light of the practical use of forecasts and their role in risk estimation. It is a challenge to maintain accuracy, while there is a growing gap in time between the forecasted observations and the realized observations, i.e. 'true' information that is available to the forecaster. Most simple frameworks for multi-step-ahead forecasting rely either on recursive (*alias* iterative) or direct (*alias* independent) schemes, although others exist that combine principles of the two [9, 83].

Recursive schemes utilize one model $\hat{f}$ for all lead times, that is trained to provide a single-step-ahead (4-hour ahead) forecast based on a total of $d + 1$ previous observations. It is thus an estimation of the true model f, i.e.

$$y_t = f(y_{t-4},\, \dots,\, y_{t-4-d}) + w \tag{37}$$

where w is an error term. To obtain a forecast with a look-ahead-time of H hours into the future from a certain time N, i.e. $\hat{y}_{N+H}$, a recursive scheme utilizes model $\hat{f}$ to obtain a forecast $\hat{y}_{N+1}$, then adds this forecast as input to $\hat{f}$ to obtain a forecast $\hat{y}_{N+2}$, and continues until it provides a forecast $\hat{y}_{N+H}$, i.e.

$$\hat{y}_{N+h} = \begin{cases} \hat{f}(y_{N-4},\, \dots,\, y_{N-4-d}) & \text{if} \quad h = 0 \\ \hat{f}(\hat{y}_{N+h-1},\, \dots,\, \hat{y}_{N-1},\, y_{N-4},\, \dots,\, y_{N-4-d+h}) & \text{if} \quad h \in \{1,\, \dots,\, d-1\} \\ \hat{f}(\hat{y}_{N+h-1},\, \dots,\, \hat{y}_{N+h-d}) & \text{if} \quad h \in \{d,\, \dots,\, H\} \end{cases} \tag{38}$$

Dependent on the lead time, only realized observations $\boldsymbol{y}$, a mix of realized observations $\boldsymbol{y}$ and forecasts $\hat{\boldsymbol{y}}$, or only forecasts $\hat{\boldsymbol{y}}$ are used as input. The main drawback of recursive schemes is its sensitivity to the accumulation of errors. As future forecasts are based (partially) on earlier forecasts that are repeatedly passed along the forecasting procedure, errors in earlier forecasts propagate, what can become especially problematic for lead times that lie further away than the furthest lag, such that there are only forecasts, and no more observations, in the system.

In contrast to the single-model setup of recursive schemes, direct schemes utilize one trained model $\hat{f}_h$ per each lead time, that are separately trained to provide an h-step ahead forecast, i.e.

$$y_t = f_h(y_{t-4},\ \dots,\ y_{t-4-d}) + w \tag{39}$$

$h \in \{1,\ \dots,\ H\}$. To obtain a forecast $\hat{y}_{N+H}$, that is for a look-ahead-time of H hours, a direct scheme utilises the model $\hat{f}_H$

$$\hat{y}_{N+H} = \hat{f}_H(y_{N-4},\ \dots, y_{N-4-d}) \tag{40}$$

Although a direct scheme does not suffer from the propagation of errors, it cannot account for potential interdependencies that may exist between lead times. Besides that, the computational cost that is associated with training numerous separate models greatly exceeds that of recursive schemes. Especially for models that are computational costly, such as the considered ANN models, that might be problematic.

The naive model NVE.DA can be used directly for multi-step-ahead forecasting, albeit for a finite number of steps restricted by the number of published market clearing prices. The series could be extended by forecasts of the market clearing price, but that lies out of the scope of this research. REG.LASSO and ANN.MLP *can* be utilized for look-ahead-times further than 9 hours, however, and this research exploits them in a direct scheme for multi-step-ahead forecasting.

6.5.2 Separate models per delivery hour

This intricacy lies within the scope of this research because earlier intraday research does not investigate whether training models separately on individual delivery hours improves accuracy. Intraday price forecasting might benefit especially, as certain information that is published/gathered during the day can only be utilized for certain delivery hours. Something that can be incorporated easily when training separate models.

As shown in Figure B.8, one of the major effects of seasonality is observed in the variation of ID3 price for the hours of a day. The majority of literature address these effects to some extent by introducing calendar features. As this might not be exhaustive, Marcjasz, Lago, and Weron [69] propose to address each hour of the day separately. To that end, it proposes a single-output procedure, where separate models are trained/tested on a subset of the observations that includes only one of the 24 hours of a day, and a procedure that takes advantage of a multi-output structure of ANNs. An apparent downside to a procedure that employs separate models per hour is that it is computational costly [69]. Besides that, there is the implication that the training and testing sets—that might be reduced by a factor of 24—might not be broad enough to generalize [52]. To the best knowledge of the author, there is no attention for this topic within the field of intraday price forecasting. For this reason, this research pays attention to the intricacy of separate models per hour.

Findings of Marcjasz, Lago, and Weron [69] are that a multi-output procedure outperforms a single-output procedure, both in terms of computational cost and of accuracy. For day-ahead forecasting, where all 24 market clearing prices of the next day are forecasted at once, that research demonstrates that a multi-output procedure is beneficial. As intraday forecasting happens continuously throughout the day, however, training a multi-output ANN is not possible because the requirement of multiple target values at each point in time is not satisfied. Therefore, this research addresses the intricacy of separate models per hour with a single-output procedure.

An especially interesting benefit of utilizing a single-output procedure, is that it can underlie analysis of whether feature usefulness varies for the different hours of the day. This might be able to support, for instance, findings from Kremer, Kiesel, and Paraschiv [60] that in comparison to the prices of morning, afternoon, and evening products, the prices of night products are more driven by pure price information and less by fundamentals such as the slope of the merit order curve and expected conventional capacities.

6.5.3 Classification of extreme prices

This intricacy lies within the scope of this research because occurrences of extreme price are particularly present in aggregated Dutch intraday price. Especially in this market, participants might benefit from knowledge about these occurrences—to take advantage of extreme prices or otherwise to evade them.

This research investigates whether posing the problem as a classification problem benefits the capability of ANN.MLP to estimate occurrence of extreme prices. To that end, the target variable is modified to be a binary value that labels the price as being extreme or not. As the target variable is not continuous anymore, the architecture and training of ANN.MLP requires slight modification; activation of the output layer is on the basis of a sigmoid function so that the output of the network represents the probability of the target variable being an extreme price, and the network is trained on the basis of a binary cross-entropy loss function. That model is referred to as ANN.C.MLP.

What prices are labeled as extreme and what is an adequate confidence is highly dependent on the practical application. This research assumes that one wants to know with a confidence of 75% whether an extreme price occurs that is considered as 75 €/MWh or higher. The forecasted observations y_t from the ANN.C.MLP are binarized and get 1 if $\hat{y}_t \geq 0.75$, and 0 otherwise, and can then be evaluated against the binarized realized observations, i.e. $\hat{y}_t * y_t$ for $t \in \{1, \ldots, T\}$.

To assess whether posing the problem as a classification leads to a more accurate classification, accuracy of the forecast from the ANN.C.MLP is evaluated against that of the 0.25-quantile forecast from ANN.Q.MLP. That quantile forecast $\zeta_t^{(\alpha)}$ represents a forecast with 25% confidence that the target observation is smaller, i.e. 75% confidence that the target observation is higher. The forecasted observations are binarized and get 1 if $\zeta_t^{(\alpha)} \geq 75$ and 0 otherwise, and can then be evaluated against the binarized realized observations.

7 Results I

Many small results of the forecasting procedure constitute to the bigger result that a more detailed procedure that avoids overfitting leads to slightly higher accuracies. Besides that, it provides insight into the selection of features and hyperparameters, that vary considerably as time progresses. It is demonstrated that the year of 2018, with high volatility and extreme prices, is more challenging than the year of 2019. Overall, the multilayer perceptron ANN for point forecasting obtains the most accurate forecast with rMAEs of 0.81 and 0.77, while the multilayer perceptron ANN for quantile forecasting is outperformed by the model based on quantile regression averaging that attains CRPSs of 3.53 and 2.24. In a multi-step-ahead horizon, results demonstrate that accuracy remains steady until the point that no more market clearing prices are available, after which it rapidly deteriorates.

The point and interval forecasting models discussed in Chapters 4 and 5 are employed in the forecasting procedure described in Chapter 6, that is an execution of initialization, as well as repeated execution of calibration and exploitation. The forecasting procedure is programmed in PYTHON. The SCIKIT-LEARN library for machine learning [89] is the programming basis for the considered regression models, and the KERAS library for ANNs [56] is the programming basis for the considered ANN models.

7.1 Point forecasting

In the results of point forecasting, the initialization, calibration, and exploitation procedures are elaborately addressed. When discussing accuracies, it is explicitly noted that it concerns the accuracy attained *by the forecast* of a particular model. Referring to the accuracy attained *by the model* itself would disregard that it is very sensitive to many choices, and would demand an even more elaborate analysis.

7.1.1 Initialization procedure

This section presents the results of the initialization procedure, that is an out-of-sample forecast evaluation of 2017. Based on this procedure, it is investigated how the feature set is reduced and what the effect is of widening or narrowing the training and testing windows of the rolling cross-validation.

Candidate feature set For the candidate feature set, there is a size where near-optimal performance meets high reduction. It is shown in related research that that point generally lies at sizes smaller than 50 features [103, 59, 2]. Accuracies of the RF regression when reducing beyond 100 features are shown in Figure 14.

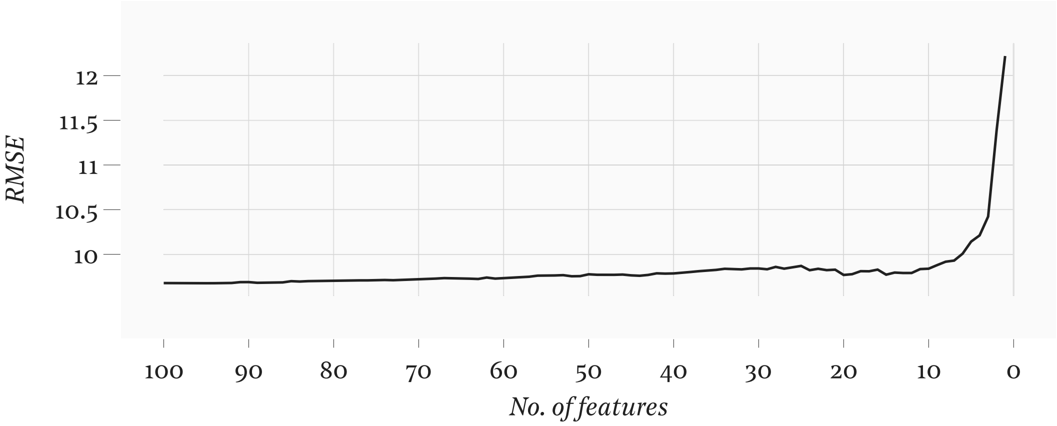


Figure 14: Accuracy in terms of RMSE as function of the number of features. *RF regression. Initialization procedure.*

Remark 1 Accuracies decrease slightly but steadily from sizes of 100 to 30 features. Only when reducing beyond a size of 10 features does accuracy decrease rapidly. At a size of approximately 15 features, near-optimal performance and high reduction meet. This research does not continue with such a limited feature set however; in a setting of recursive feature elimination, feature importance might vary considerably for different models [103] and as time progress. Also, the results shown in

Figure 14 are highly aggregated. Although more features might not benefit the overall accuracy much, they might benefit, for instance, accuracies for certain hours of the day or for certain price regimes.

This research employs a candidate feature set at a size of 50 features, that is further reduced to a size of 15 features in the calibration procedure. A candidate feature set that is broad but not so broad that computational cost is excessively high, combined with a second stage of feature reduction, is a fine-grain approach to feature selection. It can show whether the candidate feature set is utilized differently over time, for instance. Table 3 shows the 50 features included in the candidate feature set. Features are sorted by name because feature importance might not be representative of actual usefulness due to the broadness of the set, and might thus be misleading information.

(NL, ID3, -20)	(NL, LOAD_A, -4)	(NL, GEN_E1_S, -23)	(DE, LOAD_A, -5)	(DE, GEN_F1_WON, 12)
(NL, ID3, -7)	(NL, LOAD_F, -24)	(NL, GEN_E1_S, -18)	(DE, LOAD_F, -23)	(DE, GEN_F1_WOFF, -3)
(NL, ID3, -4)	(NL, LOAD_F, -21)	(DE, ID3, -168)	(DE, LOAD_F, -17)	(DE, GEN_F2_S, -9)
(NL, MCP, -1)	(NL, LOAD_F, 0)	(DE, ID3, -10)	(DE, LOAD_F, -12)	(DE, GEN_F2_WON, -20)
(NL, MCP, 0)	(NL, GEN_A_WON, -13)	(DE, ID3, -4)	(DE, LOAD_F, 8)	(DE, GEN_F2_WOFF, -21)
(NL, MCP, 1)	(NL, GEN_A_WON, -11)	(DE, MCP, -168)	(DE, LOAD_E, -4)	(DE, GEN_F2_WOFF, -6)
(NL, MCP, 2)	(NL, GEN_F1_WOFF, -21)	(DE, MCP, -18)	(DE, GEN_A_WOFF, -11)	(DE, GEN_F2_WOFF, 3)
(NL, LOAD_A, -15)	(NL, GEN_F1_WOFF, -7)	(DE, MCP, -10)	(DE, GEN_F1_WON, -17)	(DE, GEN_F2_WOFF, 6)
(NL, LOAD_A, -12)	(NL, GEN_F1_WOFF, 4)	(DE, MCP, -4)	(DE, GEN_F1_WON, -3)	(DE, GEN_F2_WOFF, 11)
(NL, LOAD_A, -7)	(NL, GEN_F1_WOFF, 5)	(DE, MCP, 0)	(DE, GEN_F1_WON, 4)	(DE, GEN_E2_S, -8)

Table 3: Candidate feature set $\mathcal{F}_C$. *Initialization procedure.*

Remark 2 Many types of features are represented, and slightly more than half of the features are German features. The Dutch and German features of the most recent ID3 are included, i.e. (NL, ID3, -4) and (DE, ID3, -4), as well as the non-lagged MCP, i.e. (NL, MCP, 0) and (DE, MCP, 0). What strikes is that the Dutch features of slightly-lagged and slightly-leading MCP are included, i.e. (NL, MCP, -1), (NL, MCP, 1), and (NL, MCP, 2), but not their German counterparts. What also strikes is that the German features of ID3 and MCP of one week ago are included, i.e. (DE, ID3, -168) and (DE, MCP, -168), but not their Dutch counterparts. Wind generation and load are represented by many types of features, while solar generation is represented mostly by features of solar generation forecast errors, i.e. (NL, GEN_E1_S, -23), (NL, GEN_E1_S, -18), and (DE, GEN_E2_S, -8).

Size of the training and testing windows Table 4 shows the accuracies that the forecasts from REG.LASSO and ANN.MLP attain in single-step-ahead forecasting as function of the sizes of training and testing windows. Sizes range from relatively narrow to relatively broad, i.e. from 1 month to 1 year for training, and from 1 week to 6 months for testing.

Remark 1 As naive models obtain forecasts without training and testing, their accuracy is unaffected by the sizes of training and testing windows.

Remark 2 For the regression and ANN models, the size of the testing window becomes insignificant as the training window widens. For a narrow training window it is beneficial to employ a narrow testing window, i.e. refit models more often. A wide training window thus benefits accuracy significantly.

Period	*2017*								
Scoring metric	**MAE (€/MWh)**								
Training window	*1m*			*6m*			*1y*		
Testing window	*1w*	*1m*	*6m*	*1w*	*1m*	*6m*	*1w*	*1m*	*6m*
NVE.DA	6.82								
NVE.ID	8.95								
REG.LASSO	5.92	5.95	6.31	5.81	5.84	5.92	5.86	5.85	5.89
ANN.MLP	5.99	6.02	6.19	5.68	5.69	5.81	5.64	5.64	5.65

Table 4: Accuracy in terms of MAE as function of sizes of training and testing windows. *Initialization procedure.*

Transformation Two conclusions regarding feature transformations are that normalization results in slightly more accurate forecasts than standardization, and that reducing spikes does not benefit accuracy at all. Normalization is thus the only transformation that is utilized in calibration and exploitation.

7.1.2 Calibration & Exploitation procedures

This section presents the results of the calibration and exploitation procedures, that is a systematic out-of-sample test of the period from 2018 through 2019. Based on this procedure, it is investigated what accuracies are attained by the considered models. In consideration of the results shown in Table 4, this research utilizes a training window of 1 year and a testing window of 1 month for the exploitation procedure. These windows result in near-optimal accuracies at a computational cost that is acceptable considering the available resources.

After $1 \cdot (30 \cdot 24)$ observations (approx. 1 month) are forecasted, the windows are rolled forward and the observations in the training and validating windows are utilized to find out what features and what hyperparameters should be used for the model that is exploited to forecast the next $1 \cdot (30 \cdot 24)$ observations. The following presents results from the stages of calibration regarding feature selection and hyperparameter optimization.

Calibration of input feature set In calibration, an RF regression is employed in a recursive feature elimination to determine which of the features included in the candidate feature set should be included in the input feature set. Figure 15 shows which features are included after each calibration. Bars represent the relative feature importance as determined in initialization. This stage of calibration is unaffected by the model that is exploited; these results are thus universal for all regression and ANN models. Again, features are sorted by name.

Remark 1 Of the 50 features in the candidate feature set, there are a total of six dominant features, that are selected in all or almost all calibrations. Those are the Dutch features of non-lagged, slightly-lagged, and slightly-leading MCP, the German feature of non-lagged MCP, and the Dutch feature of most-recent ID3. All dominant features are thus features of price. Dutch features of price are more useful than their German counterparts, as they are selected more frequently.

Remark 2 Findings in [103] demonstrate that many German and Belgian features of renewable generation and load are among the 15 most useful features for Dutch day-ahead price forecasting, and their feature importance is of similar magnitude as that of features of price. Figure 15 shows that German features of renewable generation and load are also selected, and are even more useful than their Dutch counterparts, as they are selected more frequently. Among the Dutch and German features of generation and load, there seems to be no dominant features that are selected repeatedly, however, as many features of generation and load are selected at least in three calibrations. That result for the Dutch intraday market, combined with the results in [103] for the Dutch day-ahead market, suggests that the effects of exogenous variables on the price of energy are already largely reflected in the price of the day-ahead market, or in past prices of the intraday market. A similar conclusion to that drawn in [52], for the German intraday market.

Remark 3 Of the 50 features in the candidate feature set, there are a total of six features that are not included in any of the input feature sets. That includes three of the four features of solar generation.

A less elaborate procedure might base its choice of the input feature set on a fixed cut after the 15 highest ranked features in the candidate feature set, for instance. In consideration of feature importances, the six dominant features would be included, and then nine other features. The information contained in the many features of generation and load would then not be used optimally.

Remark 4 As time progresses, Figure 15 shows that different features of generation and load become more or less useful, and either make or do not make it into the input feature set. Despite the clear dominance of a limited number of features, the feature with the fifth-highest feature importance in the candidate set, i.e. (DE, MCP, 0), does not make the input feature set in two calibrations. Calibration thus tailors the forecasting procedure to the inconstant distribution of feature importance. This might be even more essential as the share of renewable generation grows further in the energy mix, or in markets where features are less dominant.

Remark 5 There is no clear trend in the selection of features of renewable generation. Although the

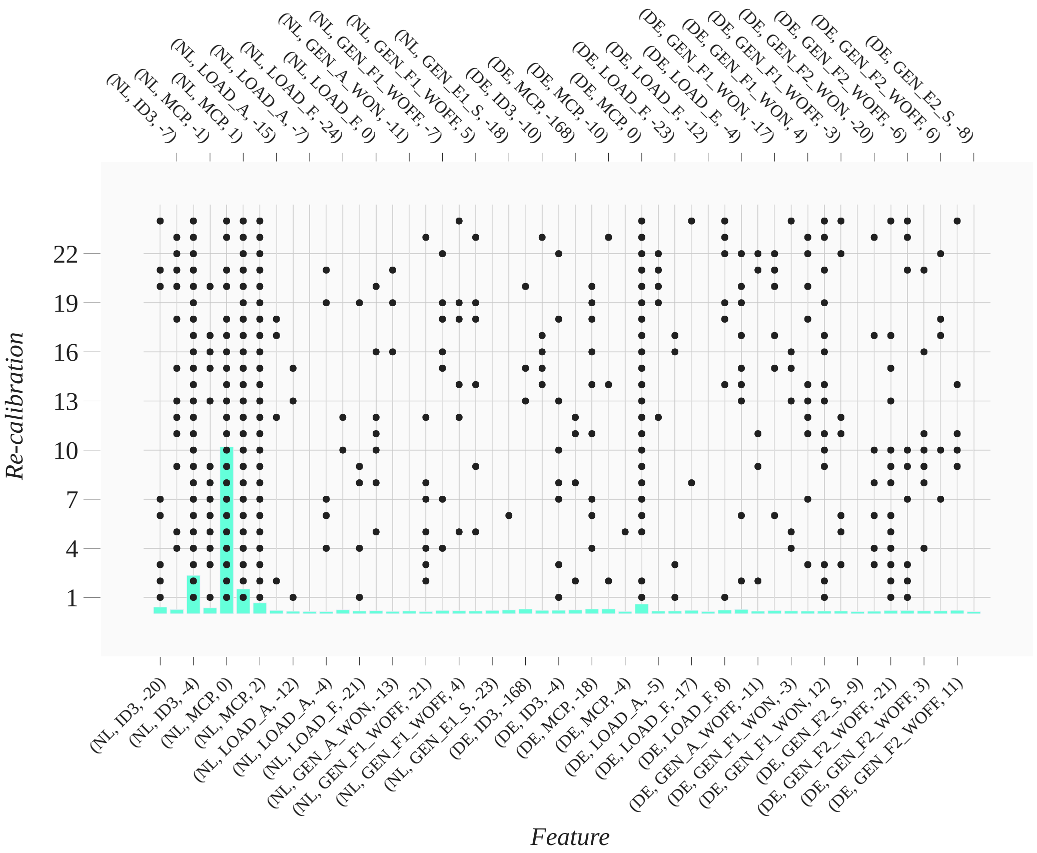


Figure 15: Input feature set $\mathcal{F}_I$ as function of calibration iteration. *Exploitation procedure.*

share of renewable generation has increased significantly over 2018 and 2019, features of renewable generation are not selected more often as time progresses.

Remark 6 Findings of [103] show that feature importances of an RF regression can be more concentrated than that of other regression models. Thus, the feature importances shown in Figure 15 do not represent an absolute truth.

Calibration of hyperparameters In calibration, it is then determined what values should be employed for the hyperparameters of the model. To that end, Bayesian hyperparameter optimization is employed with a tree-structured parzen estimator and 100 iterations. Figure 16 shows what accuracy REG.LASSO attains in terms of MAE, for every iteration of the tree-structured parzen estimator. Color indicates the calibration; the darker, the earlier the calibration. Figure 17 shows which regularization parameter λ is selected as optimal after each calibration.

Remark 1 Figure 17 shows that what is estimated to be the optimal hyperparameter varies considerably for different calibrations. For the search space of the regularization parameter λ (alpha), which is from 0 to 10, the variation of the optimal hyperparameter between consecutive calibrations exceeds 0.1 several times.

Effect of calibration As that result does not necessarily mean that more accurate forecasts are obtained, the effect of calibration on the overall accuracy that REG.LASSO attains for the period of 2017 is evaluated in different scenarios of feature and hyperparameter calibration. Similar results are found for other models. A calibration of 12 indicates that the calibration is performed after every $1 \cdot (30 \cdot 24)$ observations on the validating set, i.e. the $1 \cdot (30 \cdot 24)$ observations that precede each testing set. A calibration of 1 indicates that the calibration is performed only once, and on the $12 \cdot (30 \cdot 24)$ observations that precede the first testing set.

Remark 1 Repeated calibration leads to a slightly more accurate forecast with a MAE that is 0.14

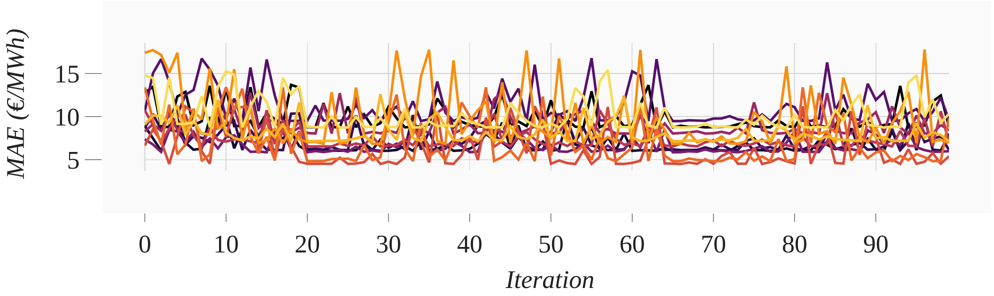


Figure 16: Accuracy in terms of MAE as function of TPE iteration. *REG.LASSO. Calibration procedure.*

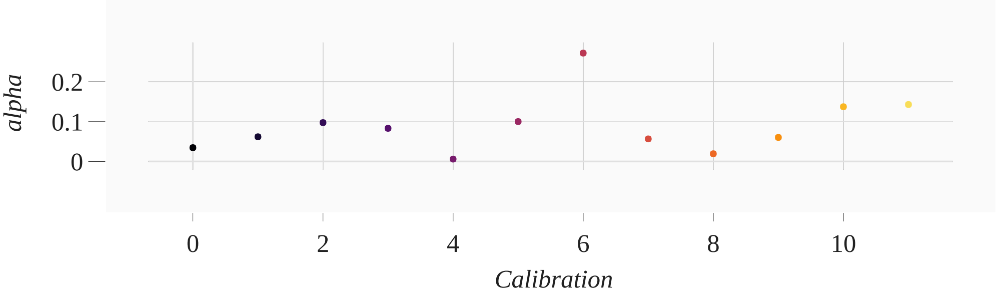


Figure 17: Value of the regularization parameter λ (alpha) as function of calibration iteration. *REG.LASSO. Calibration procedure.*

€/MWh lower than with single calibration. That result suggests that incorporating repeated calibration in the forecasting procedure is beneficial.

Period	*2017*			
Scoring metric	**MAE (€/MWh)**			
Feature calibration	*1*	*1*	*12*	*12*
Hyperparameter calibration	*1*	*12*	*1*	*12*
REG.LASSO	5.72	5.80	5.77	5.86

Table 5: Accuracy in terms of MAE as function of calibration. *Single-step-ahead point forecasting. Exploitation procedure.*

Overall accuracy Table 6 shows the accuracies that the considered point forecasting models attain for single-step-ahead (4-hour-ahead) forecasting. A visualizations of the point forecasts, for a short slice of the out-of-sample test, is shown in Figure J.1.

Remark 1 The accuracies attained by all models are higher, i.e. lower MAE values are attained, for the period of 2019 than for the period of 2018. This result demonstrates that accuracies vary considerably when viewed on a yearly scale, which is also true when viewed on monthly, daily, or hourly scales. In consideration of the significant price deviation shown in Figures 5 (yearly) and B.8 (monthly, daily, hourly), that is not surprising. A factor that might contribute to better accuracies for 2019 is that the average ID3 value over 2019 (41.85 €/MWh) is considerably lower than that over 2018 (52.93 €/MWh). Moreover, prices in 2018 are more volatile, and there are more occurrences of extreme price.

Remark 2 Of the naive models, NVE.DA attains the highest overall accuracy. As most intraday trading is incited by conditions that are unforeseen during day-ahead trading, the market clearing price largely defines the ID3 price, especially under stable conditions. NVE.DA thus represents the model to beat, and this research utilizes the forecasted series of NVE.DA to calculate the rMAE score of other models, i.e. NVE.DA attains an rMAE of 1.00 for both periods.

Remark 3 The rMAE compensates for variations of price from year-to-year, because the reference model that is NVE.DA also attains a better accuracy for 2019. Still, the rMAEs attained by all regression and ANN models are considerably lower for 2019 than for 2018. This result demonstrates that the gains from using more sophisticated models than NVE.DA are higher for 2019 than for 2018.

Remark 4 Of the considered models NVE.ID clearly attains the worst overall accuracies. That result reflects that the ID3 price generally deviates so much within a few hours, that the price of four hours ago does not hold up as an accurate estimate. The 4-hour gap between the time of forecasting and time of delivery is thus a major limiting factor for the accuracy of NVE.ID. As NVE.ID offers better accuracies than all other naive approaches based on intraday price, it is concluded that naive approaches based on intraday price do not provide reliable forecasts.

Remark 5 The forecasts from all regression and ANN models attain better accuracies than NVE.DA and thus are capable to infer to a varying degree what drives intraday prices to deviate from day-ahead prices.

Remark 6 ANN.MLP achieves the highest overall accuracy on both periods, and attains a MAE of 6.55 €/MWh for 2018 and 4.33 €/MWh for 2019.

Period	*2018*				*2019*			
Scoring metric	**rMAE (-)**	**sMAPE (%)**	**MAE (€/MWh)**	**RMSE (-)**	**rMAE (-)**	**sMAPE (%)**	**MAE (€/MWh)**	**RMSE (-)**
NVE.DA	1.00	15.19	8.06	12.83	1.00	14.33	5.63	9.30
NVE.ID	1.36	20.76	10.99	16.44	1.33	18.43	7.50	12.26
REG.LASSO	0.84	12.96	6.80	11.11	0.79	11.38	4.47	8.17
REG.SVR	0.89	13.43	7.17	12.93	0.81	11.52	4.57	8.49
ANN.MLP	0.81	12.47	6.55	10.82	0.77	11.04	4.33	8.02
ANN.GRU	0.90	13.64	7.21	11.89	0.84	12.27	4.72	8.59

Table 6: Accuracy in terms of rMAE, sMAPE, MAE, and RMSE. *Single-step-ahead point forecasting. Exploitation procedure.*

Similarly to the year of 2018, the year of 2020 is also characterised by many extreme prices. Figure 9 shows that there are more in 2018 but that they are even more severe in 2020, however. Besides that, Figure 10 shows that there are only two occurrences of (consecutive) negative prices in 2018, while there are many more in 2020. The year of 2020 is considered as a period that is extremely far from normal; it is left out of consideration in the results that follow as it would influence the results to a great extent. But a separate out-of-sample test of 2020 demonstrates exactly how challenging that year is; forecasts from NVE.DA, REG.LASSO, and ANN.MLP attain sMAPEs of 24.55, 19.70, and 19.85 %.

In-depth accuracy Table 6 is an aggregated representation of accuracy. More insightful are Figures 18 and 19, that show the accuracy that the considered models attain for the period of 2019 as function of delivery hour and of price range, respectively. Figures K.1 and K.2 show similar results, for the period of 2018. To improve readability, many figures in this chapter connect markers by lines. It should be noted that lines do not suggest interpolation, as most figures have discrete x-axes such as delivery hours or look-ahead-times.

Remark 1 Figure 18 shows that accuracy varies considerably for different delivery hours. Besides absolute accuracies, relative accuracies also vary considerably; the forecast from NVE.ID attains the best accuracy of all forecasts at 04:00, for instance, while it generally attains the worst accuracies. Already with the limited number of forecasts that this research considers, that result suggests that a model based on composite forecasting might attain a better overall accuracy than the overall accuracies attained by its individual forecasts, such as demonstrated in [77].

Remark 2 Figure 18 demonstrates the benefits of sophisticated models, as the naive model NVE.ID based on intraday price does not provide a reliable forecast. For certain delivery hours, e.g. at 04:00 and at 15:00, the forecast is rather accurate. For other hours, the forecast is extremely inaccurate, e.g. at 09:00 and during night hours. The naive model NVE.DA based on day-ahead price generally provides a more accurate forecast, although it is still unreliable for certain delivery hours. That variability is less severe for the regression and ANN models. Although their accuracy is still worse during day hours than during night hours, they provide forecasts that are more accurate and that remain steadier across delivery hours. Those are thus regarded as more reliable point forecasts.

Remark 3 Figure 18 demonstrates that the general shape of the forecasts from all models is similar, and that accuracy converges during night hours. It is especially during the morning and evening, that the simpler REG.LASSO and ANN.MLP models attain higher accuracies than the more complex REG.SVR and ANN.GRU models. That result suggests that in the forecasting procedure, the more complex models do not reach their full potential. They might be limited by the amount of training data or lose generalizability despite the efforts to fight overfitting.

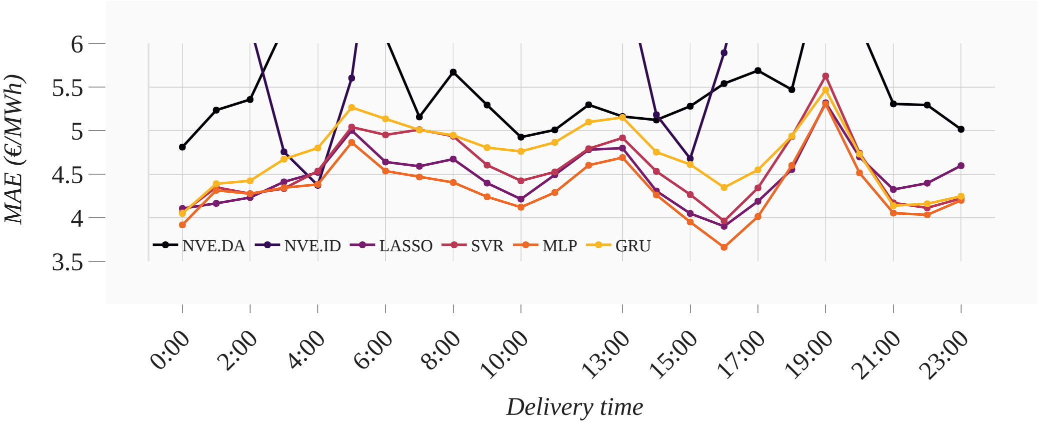


Figure 18: Accuracy in terms of MAE as function of deliver hour. *Single-step-ahead point forecasting. Exploitation procedure. 2019.*

Remark 4 As there are an equal number of observations for every delivery hour in training, validating, and testing sets, the learning process of models is not biased towards improving accuracy for certain delivery hours. There *are* many more observations for price ranges close to the average price, however. The regression and ANN models are therefore biased towards learning for those "common" price ranges. Furthermore, dynamics in extreme price regions might be much less explainable than in the common price ranges. As a consequence, all accuracy profiles shown in Figure 19 are U-shaped, from which it becomes apparent that accuracy deteriorates for occurrences of extreme price. Still, accuracies are in the lower part of the U-shape for the large majority of observations, as more than 90% of observations lie in the range $[20,\ 60)$, and more than 95% of observations lie in the range $[10,\ 70)$.

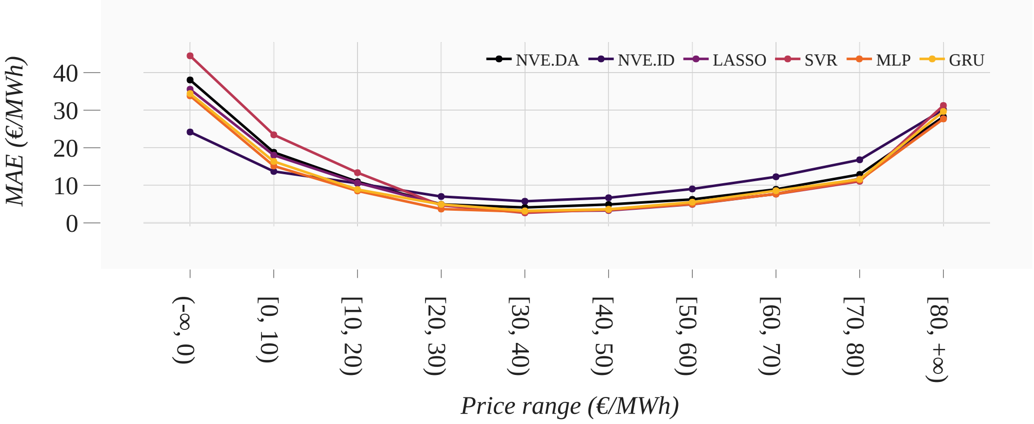


Figure 19: Accuracy in terms of MAE as function of price range. *Single-step-ahead point forecasting. Exploitation procedure. 2019.*

Superiority Figure 20 shows the results of the GW-test for the forecasts from the considered models. It supports what is reported in terms of accuracy. The forecast from ANN.MLP is superior to all other forecasts. Furthermore, the forecast from all the regression and ANN models are superior to those from the naive models. For the regression models, the null hypothesis that the competing forecasts of REG.LASSO and REG.SVR have equal conditional predictive ability cannot be rejected.

Residuals Figure 21 shows the residual errors $\boldsymbol{\epsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ of the forecast from REG.LASSO for the period of 2019, and Figure 22 shows how they are distributed.

Remark 1 The mean of that forecast lies very close to zero, which demonstrates that there is almost no bias toward over- or underestimation. Figure 22 shows that the distribution has positive kurtosis, i.e. the tails of the distribution are heavier than if the residuals were normally distributed. The distribution also has positive (right) skewness, i.e. the positive residuals are more extreme than the negative residuals, where positive residuals represent underestimation. Figure 8 shows that there are many

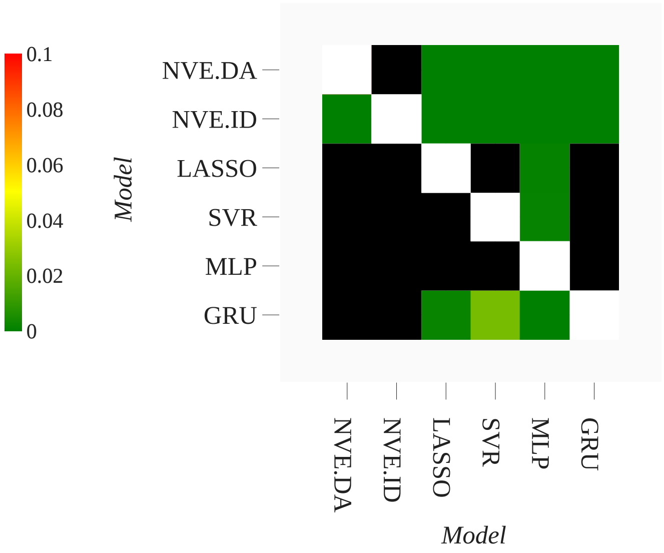


Figure 20: GW-test. *Single-step-ahead point forecasting. Exploitation procedure. 2018 & 2019.*

more extremely *high* prices that lead to relatively large under-estimations than extremely *low* prices that lead to relatively large over-estimations.

Remark 2 Although the mean is close to zero, Figure 21 shows that there is significant variation in residuals around the mean. Moreover, Figure 23 shows that there is still serial correlation in the residual time-series. For a white noise series with no seasonality, at least 95% of the coefficients should lie within the confidence interval $\pm 2/\sqrt{T}$. Figure 23 shows that some coefficients lie considerably outside that range.

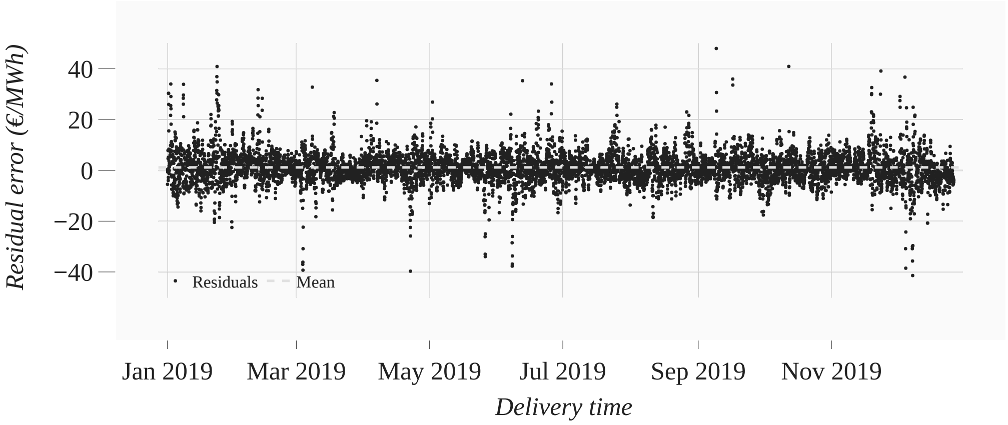


Figure 21: Residuals. *ANN.MLP. Single-step-ahead point forecasting. Exploitation procedure. 2019.*

Figure 24 shows the quantile values of the residual errors, aggregated by 7*24 observations (approx. 1 week), as an indication of variance.

Remark 3 The variance is certainly not constant. Especially at the edges of the distribution, i.e. for quantiles that lie outside the interquartile range, the average residual error varies considerably, and even the center quantile varies up to 5 €/MWh from week to week.

The mean of the residuals lies close to zero, but residuals are distributed with fat tails and still show some degree of serial correlation. Furthermore, there is some degree of heteroscedastic. The error term thus does not resemble white noise, which suggests that a more sophisticated approach might be able to capture the intraday price dynamics to a larger extent than the models considered in this research.

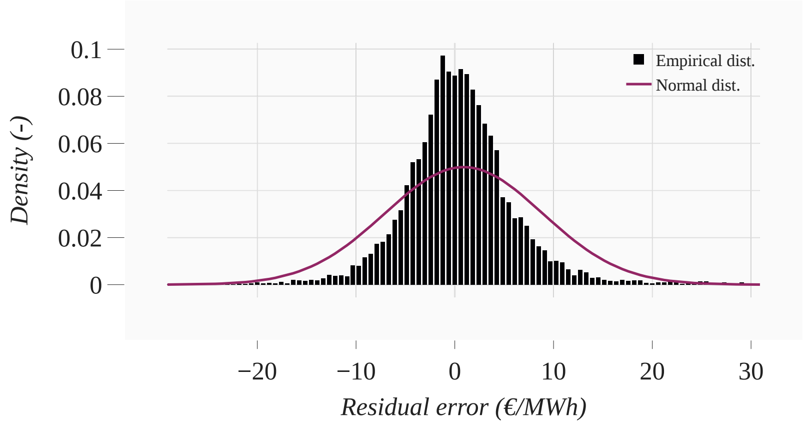
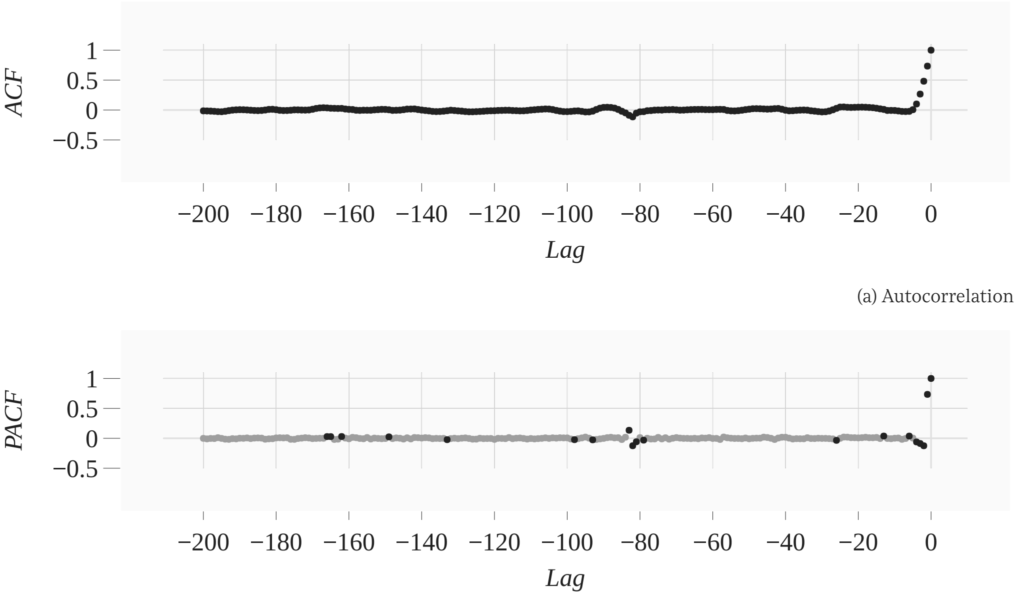


Figure 22: Distribution of residuals. *ANN.MLP. Single-step-ahead point forecasting. Exploitation procedure. 2019.*



(a) Autocorrelation

(b) Partial autocorrelation

Figure 23: Autocorrelation and partial autocorrelation of residuals. *Coefficients that are within the 95% confidence interval are a lighter shade. ANN.MLP. Single-step-ahead point forecasting. Exploitation procedure. 2019.*

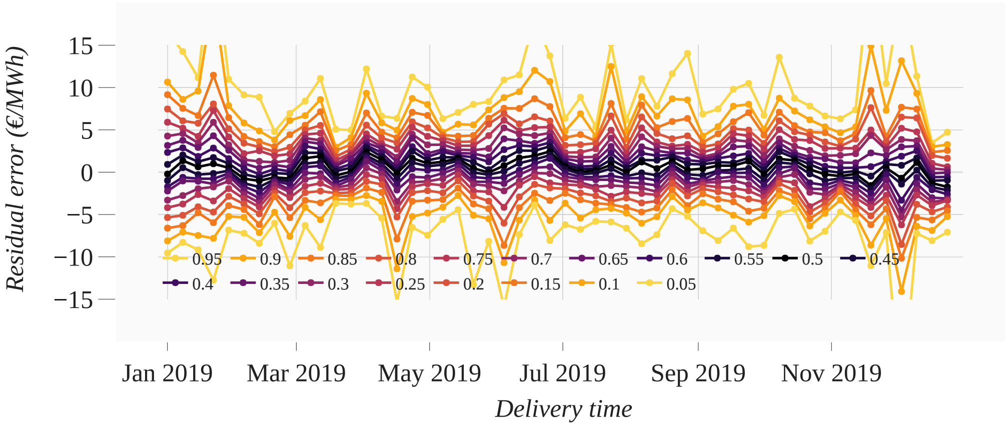


Figure 24: Quantile values of residuals. *ANN.MLP. Single-step-ahead point forecasting. Exploitation procedure. 2019.*

7.2 Interval forecasting

The previous section addresses the initialization and calibration procedures elaborately. As these procedures are very similar for interval forecasting, this section presents results of the exploitation procedure only.

7.2.1 Exploitation procedure

Overall accuracy Table 7 shows the scores that the interval forecasts of the considered interval forecasting models attain for a single-step-ahead horizon. A visualization of the interval forecast from ANN.Q.MLP is shown J.2. The CRPS is an aggregated score for all intervals at once, while the two pinball losses are calculated for each interval separately. This research utilizes an equidistant quantile grid $\alpha = \{0.05,\ 0.1,\ \dots, 0.95\}$ that gives rise to an interval grid with a step size of 10%, i.e. $\{0.1,\ 0.2,\ \dots, 0.9\}$. Results concentrate on the 0.1, 0.5, and 0.9 intervals, that represent narrow, medium, and wide intervals, respectively.

Remark 1 From the CRPSs and PLs it becomes clear that just like for point forecasting, 2018 is a relatively challenging year for interval forecasting. The results in Table 7 show that the attained CRPSs and PLs are worse for 2018 than for 2019. In addition to that and something that would not be visible when considering the average PL, is that the PLs of the upper and lower quantile forecasts of an interval are also more segregated for 2018.

Remark 2 Of the naive models, the forecast from NVE.Q.DA attains the worst accuracies and a CRPS of 3.64 for the period of 2019. NVE.Q.PNT utilizes the same approach as NVE.Q.DA but on the basis on the more sophisticated forecast from REG.MLP. That forecast attains a considerably better CRPS of 2.51. That result demonstrates that the considered naive approach of interval forecasting can be successful, although it relies on the accuracy attained by the underlying point forecast.

Remark 3 Of the more sophisticated models, the forecast from REG.Q.QRA attains the best accuracies and a CRPS of 2.24 for the period of 2019.

Remark 4 For 2018, the forecast from ANN.Q.MLP is not able to outperform NVE.Q.PNT. Its ability to generalize on observations with high volatility and many extreme prices is thus limited.

Period	*2018*							*2019*						
Scoring metric	**CRPS** (-)	**PL** (-)						**CRPS** (-)	**PL** (-)					
Interval		*0.1*		*0.5*		*0.9*			*0.1*		*0.5*		*0.9*	
Quantile		*0.45*	*0.55*	*0.25*	*0.75*	*0.05*	*0.95*		*0.45*	*0.55*	*0.25*	*0.75*	*0.05*	*0.95*
NVE.Q.DA	4.45	3.97	4.05	3.22	3.60	1.21	1.80	3.64	2.78	2.83	2.30	2.48	0.97	1.10
NVE.Q.PNT	3.58	3.21	3.29	2.55	3.00	0.92	1.56	2.72	2.11	2.14	1.76	1.88	0.75	0.91
REG.Q.QRA	3.53	3.20	3.30	2.51	3.04	0.88	1.46	2.24	2.13	2.17	1.72	1.90	0.69	0.88
ANN.Q.MLP	4.11	3.97	4.05	3.22	3.60	1.21	1.80	2.53	2.18	2.20	1.79	1.90	0.72	0.86

Table 7: Accuracy in terms of CRPS and PL. *Single-step-ahead interval forecasting. Exploitation procedure.*

In depth accuracy For interval forecasts, there are more desires to bear in mind than for point forecasts; an interval forecast represents a closer approximation of reality that must not portray uncertainty incorrectly. Therefore, interval forecasts are evaluated in terms of reliability and sharpness as well. Table 8 shows different metrics that are more directly interpretable than those of Table 7. There is the average width of the interval, the percentage of realized observations that are contained in the interval (*alias* coverage), and the number of occurrences of non strictly increasing quantile values (*alias* quantile crossings).

Remark 1 For all forecasts, empirical coverages are reasonably close to nominal coverages. Only the empirical coverages of NVE.Q.DA and NVE.Q.MLP deviate more than 4%. The 0.5 interval deviates the most, as that interval is the most challenging of the three due to having a very high density of observations.

Remark 2 Accuracies and widths reported in [54], that contains a similar analysis but of the German intraday market for the period of July 2014 through September 2016 (approx. 3 years), are of a similar order of magnitude as those shown in Tables 7 and 8. Accuracies reported in [54] are slightly better, although the naive forecast of that research attains slightly better accuracies as well. In contrast to the results shown in Table 7, results of [54] show that forecasts from a model equivalent to NVE.Q.QRA are not superior to those from a model equivalent to NVE.Q.PNT.

Remark 3 The problem of quantile crossings becomes particularly apparent in view of the forecast from ANN.Q.MLP; for 2019, there are a total of 157 observations that have one or more quantile crossings. While that result is for the quantile grid of $\boldsymbol{\alpha} = \{0.05,\ 0.1,\ \ldots, 0.95\}$, more accurate estimations of the true distribution would require an even finer grid of quantiles, and many more quantile crossings. That result demonstrates that although the forecast from ANN.Q.MLP is rather certain for 2019, demonstrated by a relatively low CRPS and relatively narrow widths, the problem of quantile crossings is more severe than in its less certain and wider interval forecast for 2018.

Period	*2018*							*2019*						
Scoring metric	**Width (€/MWh)**			**Coverage (%)**			**QC (-)**	**Width (€/MWh)**			**Coverage (%)**			**QC (-)**
Interval	*0.1*	*0.5*	*0.9*	*0.1*	*0.5*	*0.9*		*0.1*	*0.5*	*0.9*	*0.1*	*0.5*	*0.9*	
NVE.Q.DA	1.56	9.38	31.53	9	45	88	0	1.70	9.60	30.87	13	58	94	0
NVE.Q.PNT	1.31	7.52	25.78	9	45	88	0	1.31	5.51	23.45	13	59	95	0
REG.Q.QRA	1.70	10.13	32.74	11	54	91	5	1.09	6.24	20.30	10	51	92	17
ANN.Q.MLP	1.80	9.98	31.81	11	53	91	23	1.20	6.62	18.55	10	50	88	157

Table 8: Accuracy in terms of width, coverage, and quantile crossings. *Single-step-ahead interval forecasting. Exploitation procedure.*

7.3 Intricacies

Multi-step-ahead Figure 25 shows the MAE scores that the considered point forecasting models attain for multi-step-ahead forecasting.

Remark 1 Figure 25 demonstrates that accuracy rapidly deteriorates when no more published market clearing prices are available, and forecasts are based mostly on features of generation and load. It is difficult to conclude based on accuracy whether look-ahead-times further than 9 hours represent usable forecasts. For that reason, Chapter 9 investigates whether forecasts of look-ahead-times up to and including 20 hours (far) lead to higher profits than forecasts of look-ahead-times up to and including 9 hours (moderate) or 5 hours (near).

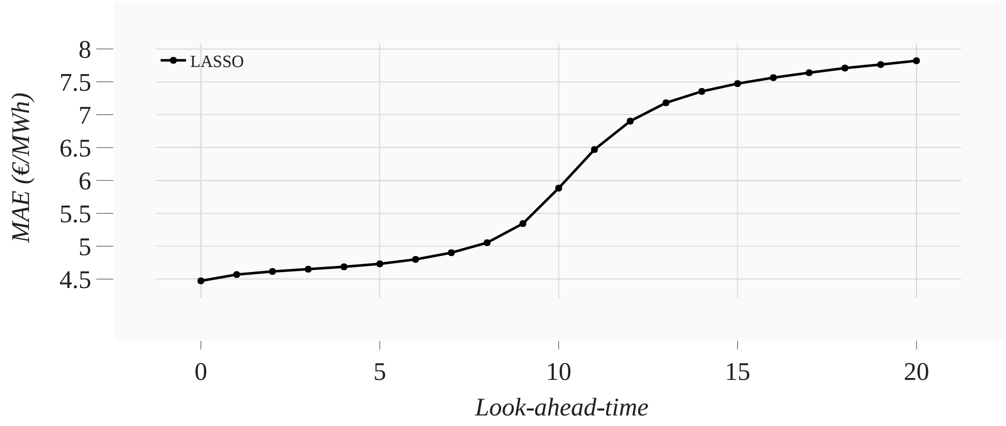


Figure 25: Accuracy in terms of MAE as function of look-ahead time. *REG.LASSO. Multi-step-ahead point forecasting. Exploitation procedure. 2019.*

Figure 26 shows the CRPS scores that the considered interval forecasting models attain for multi-step-ahead forecasting.

Remark 1 Figure 26 shows that the CRPS increases monotonically with increasing look-ahead-time. This effect follows an S-shaped curve. A relatively steep regime starts at a look-ahead-time of 9. Left of this regime, the model can base itself on realized MCP values, while right of this regime, the model cannot base itself on realized MCP values.

Remark 2 Figure 26 shows that the number of quantile crossings decreases with increasing look-ahead-time. Figure 27 shows that the intervals widen with increasing look-ahead-time. As the separation of quantiles increases, the number of quantile crossings decreases.

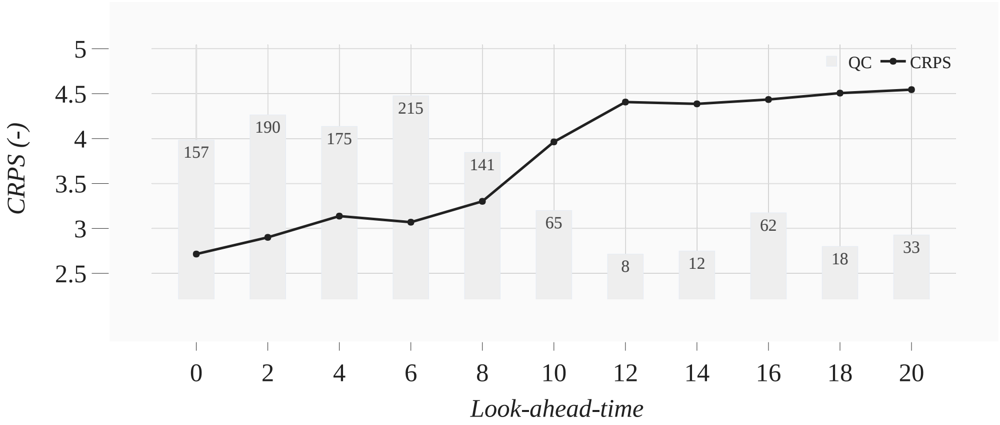


Figure 26: Accuracy in terms of CRPS as function of look-ahead time. *ANN.Q.MLP. Multi-step-ahead interval forecasting. Exploitation procedure. 2019.*

Figures 27 and 28 show the width and coverage deviation of the intervals, respectively, as function of look-ahead-time.

Remark 1 Figure 27 shows that the width increases monotonically with increasing look-ahead-time, which indicates that the model is less certain about the estimation, the further the target variable is into the future. Like the CRPS shown in Figure 26, this effect follows an S-shaped curve. Within the flatter regimes, the look-ahead-time has minor effects on the width.

Remark 2 Figure 28 shows that the coverage deviation of the wider intervals (70–90%) generally decreases as look-ahead-time increases and the average width of the intervals increase. The wider intervals at nearer look-ahead-times are generally too wide, as too many observations are covered. The opposite is true for those intervals at further look-ahead times. Empirical coverages of the wider intervals do not diverge more than 3% from nominal coverages, which suggests that reliability is good even at relatively far look-ahead-times of 20.

Remark 3 For the narrower intervals (10–30%), coverage deviation varies more considerably and is less a function of look-ahead-time. By definition, those intervals cover less observations, which is why they are more sensitive to slight misestimations. Empirical coverages of the narrower intervals often diverge more than 4% from nominal coverages, which demonstrates that the reliability of narrow intervals at further look-ahead-times is questionable.

Separate models per delivery hour The aim of training separate models per delivery hour is to investigate whether more accurate forecasts can be attained, but also to stress that separate models can incorporate different features reflecting updated information, although that does not lie within the scope of this research.

Remark 1 The accuracy that the forecasts from the 24 REG.LASSO models jointly attain for the year of 2019 is 4.45 €/MWh, and thus only slightly more accurate than the forecast from only one REG.LASSO model that attains 4.47 €/MWh. That result suggests that—if not provided with updated information—none of the 24 REG.LASSO models are more capable to learn what drives intraday prices in their corresponding delivery hour than the single REG.LASSO model. A single set of regression coefficients is not more optimal for some delivery hours than for others.

Remark 2 The 24 separate models might be limited by the number of observations in the severely reduced training sets. Further research is needed to assess whether the 24 separate models can outperform the single model when given more observations for training and/or updated information.

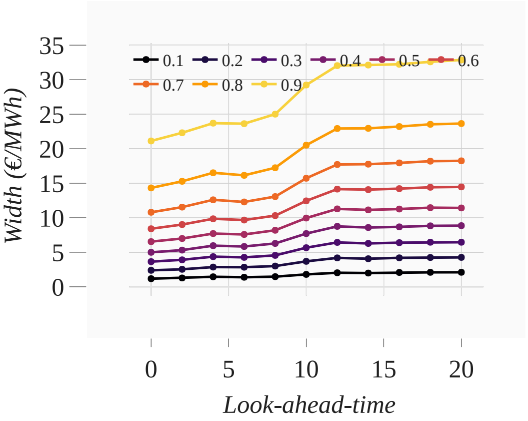


Figure 27: Width as function of look-ahead-time. *ANN.Q.MLP. Multi-step-ahead interval forecasting. Exploitation procedure. 2019.*

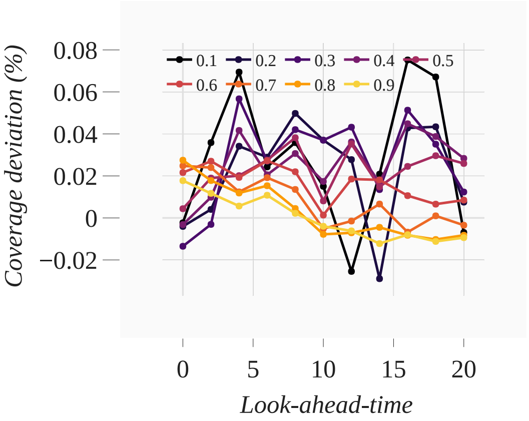


Figure 28: Coverage deviation as function of look-ahead-time. *ANN.Q.MLP. Multi-step-ahead interval forecasting. Exploitation procedure. 2019.*

Classification of extreme prices The aim of the classification problem is to estimate with 75% confidence whether prices exceed 75 €/MWh. The classification-forecast from ANN.C.MLP and the quantile-forecast from ANN.Q.MLP are evaluated for the year of 2018 as it contains many occurrences of extreme prices.

Remark 1 Both approaches to classification do not result in adequate accuracy, which is reflected mostly by the attained number of false negatives that can be detrimental to risk strategies; of the total number of observations with a price higher than 75 €/MWh, 83% and 86% were incorrectly classified by ANN.Q.MLP and ANN.C.MLP, respectively. ANN.Q.MLP performs slightly better in terms of false positives, with a total of 16 false positives against 131 from ANN.C.MLP. That result demonstrates that extreme prices in the Dutch market are very difficult to forecast with a simple approach, and that posing the problem as a classification is not an approach that outperforms a regression-based classification.

8 Point forecasting and model predictive control: Energy plant with storage capacity

This chapter introduces a simulation of an energy plant with storage capacity that smartly dispatches generated energy to the Dutch intraday market. The aim is to investigate whether superiority that is evaluated theoretically is also found when evaluated practically. Model predictive control handles optimization of the dispatch schedule by charging and discharging the battery with generated energy, and is supplied with information on future price and generation. A base case is established based on a hypothetical wind plant and battery, as well as additional cases that show whether outcomes are sensitive to system parameters.

Only in a practical scenario does the value of a point forecast become apparent; when it translates into actual profit, for instance. In a mostly flat market, sophisticated price forecasts—no matter how accurate—might not lead to more profitable schedules than naive forecasts, while in a volatile market, even mildly accurate forecasts might lead to more profitable schedules. Therefore, it is beneficial to assess not only accuracy but also the impact of a forecast on decision costs and profit [96].

This research establishes a simulation that investigates an energy plant with storage capacity. The system has access to price forecasts of the intraday market and uses model predictive control to settle on a dispatch schedule. The objective of the simulation is not to offer an as-realistic-as-possible outlook on real-world performance. Rather, it is used to see whether and how a system that is a simulation of reality reacts to different price forecasts. The system is therefore a simplified representation of reality while being sophisticated enough to offer insights. While [101] has an aim similar to that of this research, with a trading strategy that buys and sells on the Polish day-ahead and balancing markets, results are purely profit oriented, and it does not state elaborate details about the system. [82] investigates a system similar to that of this research, but does not aim to deduce the impact of price forecasts.

8.1 System description

Because the system is reduced to only the core components, case parameters are mostly universal, and conversion to different energy plants with different types of storage capacities is straightforward. What follows is a description of the system, where attention is given to several assumptions that might affect results when altered. As it is the aim to infer the effect of price forecasts, full knowledge about future generation, for instance, eliminates the effects that uncertainty would have on the results. The downside of that assumption is that absolute results of the simulation might not fully represent real-world results, however.

Figure 29 shows a schematic overview of the system, with the various subsystems that interact. There are the system parameters, that define the specifications of the battery. There is the forecasting procedure that provides a multi-step-ahead point forecast of the ID3 index to the model predictive control, and the energy plant that provides the volume of generated energy to the model predictive control. The model predictive control minimizes an objective function, based on system constraints, and controls how the battery should discharge. The battery provides the model predictive control with an update of its level of charge. Finally, there is the intraday market that provides the true price of the ID3 index.

8.1.1 Base case

The component of the energy plant in the base case is a hypothetical wind plant, because wind energy is the largest variable renewable energy source in terms of volume in the Dutch energy mix, as shown in Figure 1. Because high-resolution data of historic generation from actual wind plants in the Netherlands are not publicly available, this research employs the aggregated generation of all wind farms installed on land [30], i.e. all onshore wind plants. Two transformations are performed on that series. Firstly, it is averaged for each hour of 2018 and 2019, so that it becomes a one-year series. Secondly, it is normalized so that for every month there is a total generation of 300 MWh. Because an averaged and normalized one-year generation profile is utilized, the effects of varying generation volumes from

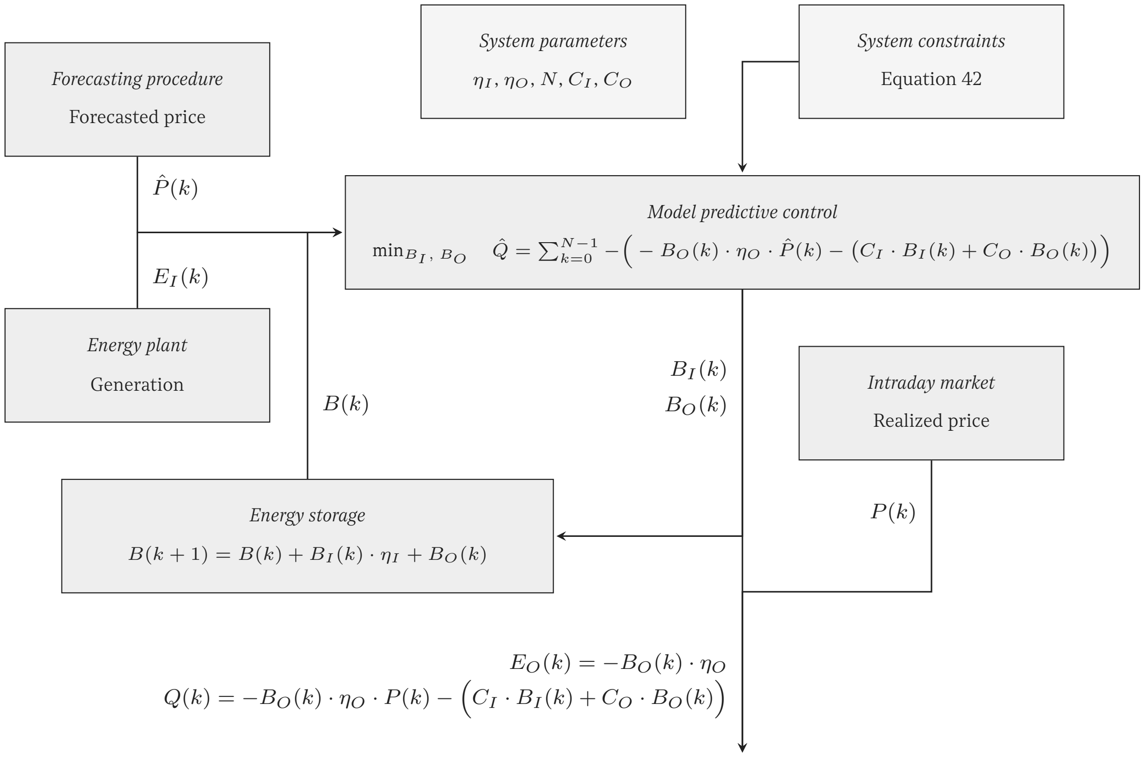

System parameters
$\eta_I, \eta_O, N, C_I, C_O$
System constraints
Equation 42
Forecasting procedure
Forecasted price
$\hat{P}(k)$
Model predictive control
$\min_{B_I,\, B_O} \quad \hat{Q} = \sum_{k=0}^{N-1} -\Big(- B_O(k) \cdot \eta_O \cdot \hat{P}(k) - \big(C_I \cdot B_I(k) + C_O \cdot B_O(k)\big)\Big)$
$E_I(k)$
Energy plant
Generation
$B(k)$
$B_I(k)$
$B_O(k)$
Intraday market
Realized price
$P(k)$
Energy storage
$B(k+1) = B(k) + B_I(k) \cdot \eta_I + B_O(k)$
$E_O(k) = -B_O(k) \cdot \eta_O$
$Q(k) = -B_O(k) \cdot \eta_O \cdot P(k) - \Big(C_I \cdot B_I(k) + C_O \cdot B_O(k)\Big)$

Figure 29: Schematic overview of the system.

month to month and from year to year are eliminated, and the effect of different price forecasts, and that of varying price characteristics of 2018 and 2019, can be assessed. The transformed generation profile has a nominal power of approx. 1.3 MWh, which coincides with the specifications of the relatively small onshore wind turbine SIEMENS SWT-1.3-62, six of which are operational for the wind farm Beabuorren in Friesland, The Netherlands [86]. The subsystem of the energy plant is thus reduced to a generation profile, of which the system has full knowledge.

The component of energy storage in the base case is a large-scale stationary energy storage product in the form of a rechargeable lithium-ion battery. The base case follows the specifications of the TESLA MEGAPACK [98]. The battery is assumed to have a fixed maximum capacity of 10 MWh, and a fixed minimum capacity of 0 MWh, fixed maximum charge and discharge rates of 5 MWh per hour, and fixed charging and discharging efficiencies of 95%, i.e. a fixed round-trip efficiency of approx. 90%. There is also a cost of usage, that incorporates that the lifetime of the battery is affected when repeatedly charging and discharging the battery, that is assumed to have a fixed value of 3 €/MWh. The subsystem of energy storage is thus reduced to a capacity range, charging and discharging rates and efficiencies, and a cost of usage [82].

Another important consideration is the look-ahead-time. Lower look-ahead-times mean that the provided forecasts are relatively accurate, but that the system might be limited to identify low and high prices within that time. Higher look-ahead-times mean that the provided forecasts are relatively inaccurate, and the system might base choices on incorrect information. The base case utilizes multi-step-ahead forecasts with look-ahead-times up to and including 9 hours.

8.1.2 Additional cases

In the Netherlands, the photovoltaic power potential is relatively low. Besides that, wind and solar plants have inherently different generation profiles, as many evening/night/morning hours can provide wind energy but no solar energy. As prices are generally higher during the day, which is also when there is solar generation, the system might work well on the basis of solar generation as well. To investigate the effect of the energy plant, this research investigates an additional case based on a solar generation profile. The optimal schedule might be sensitive to the specifications of the battery as well. First and foremost, the capacity of the battery might limit the amount of time that the system can wait before it offers energy to the market. Furthermore, high charging and discharging constraints might limit the system to employ an aggressive schedule, so that it is more reliant on longer periods of high prices. Profits are also sensitive to the specifications of the battery, as cost of usage and charging and discharging inefficiencies are a significant cost to the system. Finally, the optimal schedule might be sensitive to the look-ahead-time. It is investigated whether the moderate look-ahead-time of 9 hours that is utilized in the base case is indeed an optimal balance between a lack of information and misinformation, and it is investigated what the effect is of increasing and decreasing the look-ahead-time. This research investigates additional cases with look-ahead-times up to and including 20 hours (long) and 5 hours (short).

8.2 Optimization problem

Output variables of the system constitute the dispatch schedule, and are the amount of energy supplied to the battery $B_I(k)$ and the amount of energy that is drawn from the battery $B_O(k)$. The volume that is dispatched is $-B_O(k)\cdot\eta_O$. Input variables of the system are what the system bases its optimal dispatch schedule on, and are the amount of energy generated by the energy plant $E_I(k)$, the amount of energy stored in the battery $B(k)$, and the forecast of the Dutch ID3 price index $\hat{P}(k)$.

The objective of the system is to maximize an objective variable, the estimated profit $\hat{Q}$, by controlling the amount of energy supplied to the battery $B_I(k)$ and drawn from the battery $B_O(k)$. The energy drawn from the battery, minus what is lost due to discharge efficiency, is dispatched according to a hypothetical transaction on the intraday market. At a sampling instant k, $\hat{Q}(k)$ is defined as the amount of energy dispatched, i.e. $-B_O(k)\cdot\eta_O(k)$, times the estimated ID3 price $\hat{P}(k)$, minus the costs of the battery, i.e. $C_I \cdot B_I(k) + C_O \cdot B_O(k)$. This research refers to $Q(k)$ and $\hat{Q}(k)$ as profit and estimated profit, although a more delicate approach that takes into account trading activity would be necessary for them to be more representative of real world profits.

In state space description, the system is formalized as an explicit discrete time-invariant system, where the state vector $\boldsymbol{x}(\cdot)$, input vector $\boldsymbol{u}(\cdot)$, state matrix $\boldsymbol{A}$, and input matrix $\boldsymbol{B}$ are defined as

$$\underbrace{\begin{bmatrix}B(k+1)\end{bmatrix}}_{\boldsymbol{x}(k+1)} = \underbrace{\begin{bmatrix}1\end{bmatrix}}_{\boldsymbol{A}}\underbrace{\begin{bmatrix}B(k)\end{bmatrix}}_{\boldsymbol{x}(k)} + \underbrace{\begin{bmatrix}\eta_I & 1\end{bmatrix}}_{\boldsymbol{B}}\underbrace{\begin{bmatrix}B_I(k) & B_O(k)\end{bmatrix}^\top}_{\boldsymbol{u}(k)} \tag{41}$$

Model predictive control (*alias* receding horizon control) is a control scheme that optimizes for a finite number of future states in the prediction horizon, then executes the first step of the found solution, and then re-optimizes on a horizon that is shifted one step into the future. That makes it particularly suitable for the problem at hand, as updated forecasts and system states can be provided continuously. Model predictive control utilizes an objective function and constraints that together formalize the problem for the prediction horizon K. The optimization must satisfy all equality and inequality constraints. The energy stored in the battery at the next sampling instant $B(k+1)$ is equal to the energy stored in the battery at the current sampling instant $B(k)$ plus the energy supplied to the battery $B_I(k)$ times the charge efficiency η_I, plus the energy drawn from the battery $B_O(k)$. The amount of energy that exits the system $E_O(k)$ is equal to the energy generated $E_I(k)$, minus the energy supplied to the battery $B_I(k)$, minus the energy drawn from the battery $B_O(k)$ times the discharging efficiency η_O. The energy stored in the battery $B(k)$ cannot lie outside capacity constraints. The energy drawn from the battery $B_O(k)$ cannot be absolutely larger than the energy stored in the battery $B(k)$. Finally, the amount of energy that is supplied to the battery $B_I(k)$ and drawn from the battery $B_O(k)$ cannot be more than the maximum charge/discharge rate.

$$\begin{aligned}
\min_{B_I,\, B_O} \quad & \hat{Q} = \sum_{k=1}^{K} -\Big(- B_O(k) \cdot \eta_O \cdot \hat{P}(k) - \big(C_I \cdot B_I(k) + C_O \cdot B_O(k)\big)\Big) \\
\text{s.t.} \quad & B(k+1) = B(k) + B_I(k) \cdot \eta_I + B_O(k) \\
& E_O(k) = E_I(k) - B_I(k) - B_O(k) \cdot \eta_O \\
& 0 \leq B(k) \leq 10 \\
& 0 \leq B_I(k) \leq 5 \\
& -5 \leq B_O(k) \leq -B(k)
\end{aligned} \tag{42}$$

8.2.1 Assumptions

In the base case, a constraint is added to this problem that all energy that is generated enters the battery before it is dispatched, i.e. there is an equality constraint $B_I(k) = 100\% \cdot E_I(k)$. All generated energy is thus dispatched, and the system has no destination for energy other than dispatch after transacting on the intraday market. Without that constraint, the system would charge the battery only when deeming that it increases estimated profit. Constraining $B_I(k) = 100\% \cdot E_I(k)$ eliminates the effects of the cost of usage and of inefficiencies of the battery on choices that the system makes for charging, such that the battery specifications affect the absolute profit only. That constraint is thus very important in forcing all schedules to start with the same initial situation, so that the effect of changing the price forecast is separated from other effects.

It is assumed that all offers that are hypothetically placed for dispatch are accepted for the realized ID3 price of the product in question, although in reality that would require matching counteroffers. The 4-hour gap between forecasting and delivery provides a broad window wherein offers of the ID3 price can be be accepted. Also, the system never dispatches more than $95\% \cdot 2.5$ MWh. Offers of such volumes are very common on the Dutch intraday market, demonstrated by the distribution of offer volumes in 2018 and 2019, shown in Figure A.2. Therefore, it is assumed that the gross income from dispatch is simply the amount of energy dispatched times the realized ID3 price, i.e. $\sum_{t=1}^{T} E_O(t) \cdot P(t)$.

8.2.2 Evaluation

When all realized system outputs, i.e. $B_I(t)$ and $B_O(t)$ for $t = \{1, \ldots, T\}$, are determined from the optimization problem of Equation 42, the profit Q can be calculated, i.e.

$$Q = \sum_{t=1}^{T} -B_O(t) \cdot \eta_O \cdot P(t) - \Big(C_I \cdot B_I(t) + C_O \cdot B_O(t)\Big) \tag{43}$$

The schedules are evaluated on the profit Q and on rel. profit, i.e. the profit normalized by that attained by a reference schedule. Furthermore, they are evaluated on the percentage of hours that the system dispatches (*alias* dispatch frequency), and the average volume of dispatch (*alias* dispatch volume). The dispatch frequency and volume are linearly related because all generated volume is dispatches, and thus shine a different light on the same result.

Two reference schedules are added to the evaluation. The first reference schedule, referred to as *Direct feed*, involves no optimization or forecasting. That schedule simply dispatches all energy generated during an hour for the ID3 price of that hour, such that $E_O(t)$ simply equals $E_I(t)$ for all t and the profit is $\sum_{t=1}^{T} E_I(t) \cdot P(t)$. The second reference schedule, referred to as *True price*, is optimized based on the true intraday price. The system thus has full knowledge of price and generation and thus provides an upper limit for schedules that are optimized based on price forecasts. Direct feed is utilized to normalize the profits of other schedules, i.e. the rel. profit of *Direct feed* is 100%.

9 Results II

This chapter discusses results from the case study. It is concluded that the system attains higher profits when exploiting a schedule that is based on a more accurate forecast. With a schedule based on the most accurate point forecast, that from ANN.MLP, the system attains a profit that is 94.5% of the reference profit that is based on a practically infeasible direct feed strategy. Even based on the least accurate forecast, that from NVE.DA, the system is able to attain 93.1% of the reference profit. It is concluded that it is not so much the accuracy of the forecast than its shape, that is important to attain profit. In terms of dispatch frequency and volume, the schedules are more divergent, however, and especially for a smaller storage capacity, the schedules based on forecasts from REG.LASSO and ANN.MLP become relatively high frequency and low volume schedules, with more than 8% higher dispatch frequencies than the schedule based on forecasts from NVE.DA. It is concluded that the considerable reduction of point forecast accuracies for further look-ahead-times does not mislead the system and deteriorate profits, as it amasses approx. 3% more profit with schedules based on forecasts for further look-ahead-times.

The multi-step-ahead point forecasts discussed in Chapter 7 are employed in the case study described in Chapter 8. The case study is programmed in MATLAB. It is formulated as a linear programming that is solved with the YALMIP toolbox for modeling and optimization [110].

Results are discussed as much as possible in relative terms, as absolute results are sensitive to system parameters. Of the in Chapter 7 considered forecasts, only the forecasts from NVE.DA, REG.LASSO, and ANN.MLP are considered in this chapter, as they have relatively low computational costs and very similar conclusions would be drawn based on the forecasts from other models. The schedules are evaluated for the full period of 2018 through 2019, as well as per year separately. The period of 2018 is identified earlier to be a more challenging year with higher volatility and more extreme prices, such that forecasts attain worse sMAPE scores for 2018 than for 2019.

9.1 Base case

Table 9 shows how the schedules based on the different price forecasts perform.

Remark 1 Table 9 shows that the schedule based on the true price attains the highest rel. profit of 104.5% over the full period. As this schedule can take advantage of extreme prices of which it has perfect knowledge, this schedule performs unrealistically well and provides an upper limit for the other schedules. That schedule is the only to attain a rel. profit higher than 100% given the parameters of the base case.

Remark 2 Intraday prices are higher for the period of 2018, which is reflected by the fact that all schedules attain profits that are more than 30% higher for 2018 than for 2019.

Remark 3 Intraday prices are also more volatile for the period of 2018, which is reflected by the fact that the schedule based on the true price attains a rel. profit that is approx. 7% higher for 2018 than for 2019. The schedules based on NVE.DA and REG.LASSO are only able to convert the higher volatility into a 5% higher rel. profit for 2018 than for 2019, which might reflect the fact that forecast accuracies are relatively slightly better for 2019 (sMAPE of 14.33%, 11.38%) than for 2018 (sMAPE of 15.19%, 12.96%). All schedules are thus able to take advantage of the stronger presence of minimum-maximum pairs in more volatile market conditions and therefore come closer to the generated value.

Remark 4 Absolute and relative profits are only marginally sensitive to the three different forecasts, and deviate only about 1%. Based on the forecast from NVE.LASSO, the system is able to obtain a profit in 2018 that is 0.7% higher than based on the forecast from NVE.DA, although the forecast from NVE.LASSO attains an accuracy that is 2% lower in terms of sMAPE. A factor that contributes to the insensitivity is that the system bases its decisions largely on the estimation where low and high occurrences of price are during the day, which follows a seasonal profile that is captured quite well by all forecasts. Next to accuracy, it is thus important that the profile of the forecasted and realized prices are of a similar shape. Something that is also concluded in [82]. Another factor that contributes to the insensitivity is that discharging of the battery, i.e. dispatching of energy, happens at a limited discharge rate. The system is thus forced to 'spread out' dispatch over a region of high prices instead of concentrating all dispatch on a single occurrence of high price which would introduce sensitivity caused by slight misestimations of local maxima.

Remark 5 In terms of dispatch frequency and offered average volume, the schedules based on the different forecasts are more dissimilar however. Based on REG.LASSO, the system dispatches energy in almost 8% more instances than that based on NVE.DA. With that higher offer frequency comes a reduced average volume. The schedule based on REG.LASSO is thus characterized as a relatively high frequency and low volume schedule.

Remark 6 With the parameters of the base case, none of the profits attained by schedules based on forecasts are higher than that of direct feed. However, that schedule represents consistently offering all generated energy on the intraday market, which is assumed to be infeasible in reality because it may violate market trading constraints and may be subject to a high degree of unaccepted offers.

	Profit (k€)			**Rel. profit (%)**			**Dispatch frequency (%)**			**Dispatch volume (MWh)**		
Period	*Full*	*2018*	*2019*	*Full*	*2018*	*2019*	*Full*	*2018*	*2019*	*Full*	*2018*	*2019*
Direct feed	333.1	186.2	147.0	100	100	100	100	100	100	0.42	0.42	0.42
True price	348.2	200.2	148.0	104.5	107.6	100.7	23.9	24.0	23.9	1.57	1.57	1.57
NVE.DA	310.1	177.5	132.6	93.1	95.4	90.2	21.2	20.5	21.9	1.77	1.83	1.72
REG.LASSO	311.9	178.8	133.1	93.6	96.1	90.6	28.8	29.1	28.5	1.30	1.29	1.32
ANN.MLP	314.8	180.0	134.8	94.5	96.7	91.7	27.5	29.0	25.9	1.37	1.30	1.45

Table 9: Performance of schedule in terms of profit, rel. profit, dispatch frequency, and dispatch volume. *Wind plant. Storage capacity of 10 MWh.*

For a one-week period in 2018, for the system based on a wind plant, and for the schedule based on REG.LASSO, the system variables as function of time are shown in Figure 30. What generally holds true for this system is that a large portion of energy is generated during night hours, when the ID3 price index is usually lower than during day hours.

Remark 7 What the schedule thus successfully does is store the energy generated during night hours, to increase the volume that it offers on the intraday market at estimated local maxima during day hours. This results in repeated cycles of slow charge and fast discharge.

Remark 8 Because generation is rather continuous throughout the day, the system can discharge and then charge the battery in a rapidly repeating cycle. In the one-week period shown in Figure 30, there are no occurrences where the battery is exploited near its maximum capacity of 10 MWh.

9.2 Additional cases

Impact of storage capacity The storage capacity can greatly impact the schedules. With a smaller battery, the schedule must dispatch energy more frequently and might need to settle with smaller price differences. With a larger battery, the system is more capable to hold on to a lot of energy that it can potentially hold onto until it estimates very high prices. It is still constrained by look-ahead-time and by maximum charge/discharge rates, however. Table 10 shows how the schedules based on the different price forecasts perform, when the system has a smaller battery with a maximum capacity of 5 MWh and maximum charge/discharge rates of 2.5 MWh/hour or a larger battery with a maximum capacity of 20 MWh and maximum charge/discharge rates of 10 MWh/hour.

Remark 1 Table 10 shows that with a smaller battery, the system is indeed forced to exploit schedules that dispatch more often. The dispatch frequencies of all schedules increase more than 9%. Besides that, the rel. profit of the schedules based on NVE.DA and REG.LASSO decrease more than 2%. The rel. profit of the schedule based on the true price, which is now less capable to amplify volume during high prices, is hit even harder by the smaller battery, however, and decreases more than 6%.

Remark 2 With a larger battery, the schedules are able to attain higher profits, at lower dispatch frequencies. The system is less limited by the storage capacity, and can keep charging the battery more

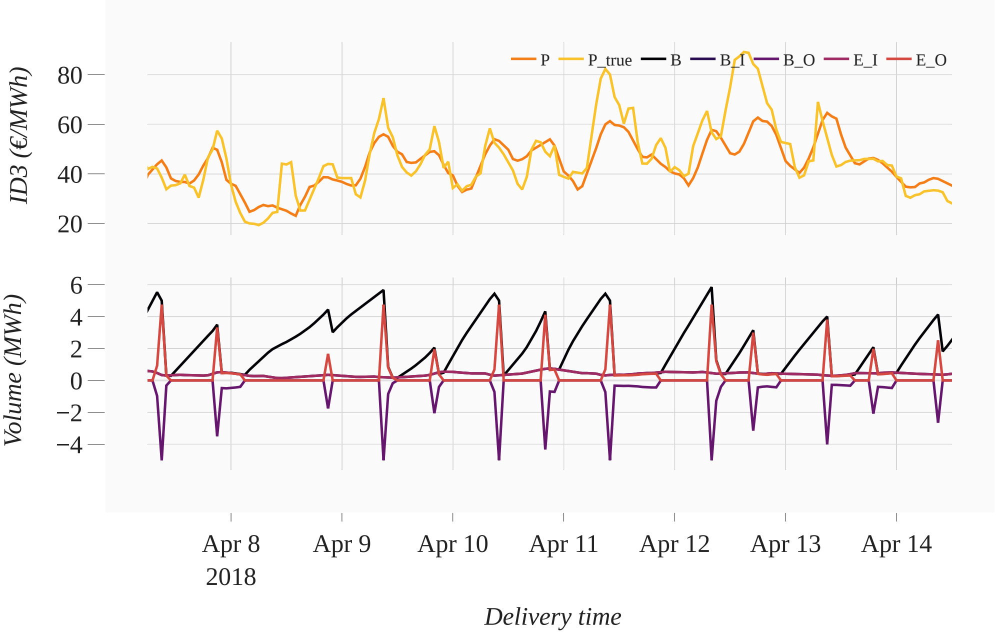


Figure 30: System variables as function of time. *REG.LASSO. Wind plant.*

often, so that more energy can be dispatched at high prices. The gains are clearly less when utilizing a 20 MWh instead of a 10 MWh battery, than when utilizing a 10 MWh instead of a 5 MWh battery.

Remark 3 With a smaller battery, the schedules diverge even more in terms of dispatch frequency. With a storage capacity of 5 MWh, the schedule based on REG.LASSO dispatches energy in approx. 11% more instances than that based on NVE.DA, while that is approx. 7% with storage capacities of 10 and 20 MWh.

Figure 31 shows the system variables for the schedules based on NVE.DA and REG.LASSO. In that slice of the full period, three gray areas highlight two occurrences where the schedules diverge and an occurrence where the schedules are exactly the same.

Remark 1 The first gray area (2019.08.09 at 18:00) highlights an occurrence where the REG.LASSO forecast (orange, dotted) provides an accurate estimation of the local maximum, and thus the system chooses to amplify the volume dispatched at that time (red, dotted). The NVE.DA forecast (orange, line) does estimate a local maximum, but does not provide an accurate estimation and thus the system chooses to dispatch energy continuously (red, line). That occurrence demonstrates that the schedule is affected by how well the vertical location, i.e. price, of true local maxima/minima are estimated.

Remark 2 The second gray area (2019.08.11 at 20:00) highlights another occurrence where the REG.LASSO forecast provides a more accurate estimation of the local maximum. Although both schedules dispatch the same amount of energy, the dispatch of the schedule based on REG.LASSO coincides with the true local maximum much better. That occurrence demonstrates that the schedule is affected by how well the horizontal location, i.e. time, of true local maxima are estimated.

Remark 3 The third gray area (2019.08.12 at 07:00) highlights an occurrence where the REG.LASSO and NVE.DA forecasts are different but of a similar shape, and lead to the exact same choice of charging and dispatch in both schedules. That occurrence demonstrates that the shape of the forecast can be enough for the system to base its choices on.

Impact of generation profile The generation profile of a solar plant is very different to that of a wind plant. Both have dynamics that are unexpected, but the solar generation profile brings spikes in generation and many hours that no generation occurs, while a wind generation profile is generally more smooth and there are far less occurrences of no generation. The employed generation profiles of solar and wind for in summer and winter are shown in Figures L.1 and L.2. The volumes of these profiles are averaged over 2018 and 2019, and normalized to a monthly generation of 300 MWh. Therefore, the

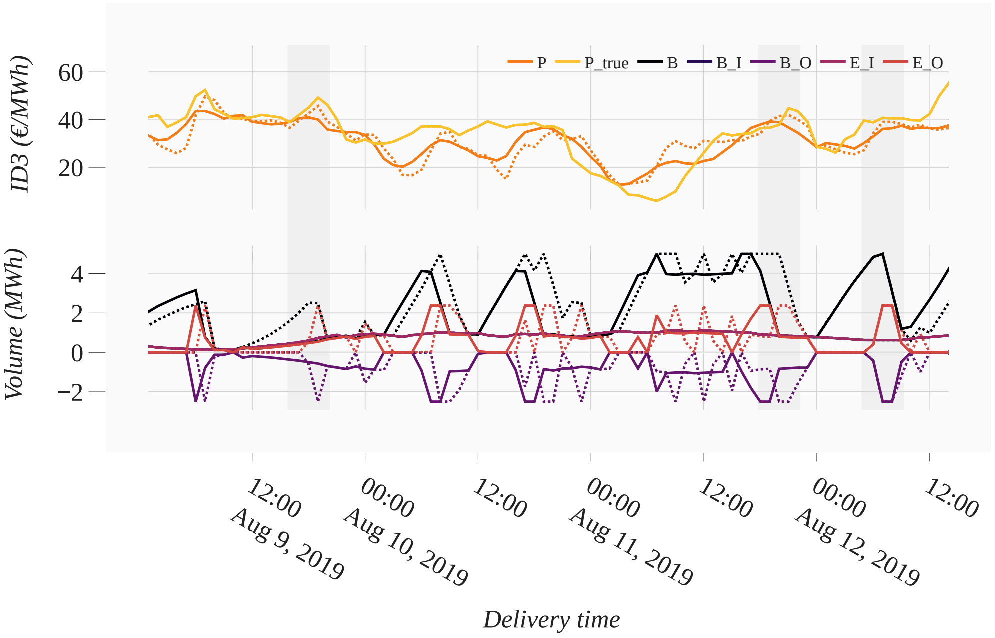


Figure 31: System variables as function of time. *NVE.DA (line) & REG.LASSO (dotted). Wind plant. Storage capacity of 5 MWh.*

	Profit (k€)			**Rel. profit (%)**			**Dispatch frequency (%)**			**Dispatch volume (MWh)**		
Period	*Full*											
Storage capacity	*5*	*10*	*20*	*5*	*10*	*20*	*5*	*10*	*20*	*5*	*10*	*20*
Direct feed	333.1			100			100			0.42		
True price	328.9	348.2	356.9	98.0	104.5	106.9	36.0	23.9	20.2	1.04	1.57	1.86
NVE.DA	300.1	310.1	313.8	90.1	93.1	94.2	31.5	21.2	18.0	1.19	1.77	2.10
REG.LASSO	302.8	311.9	315.3	90.9	93.6	94.7	41.9	28.8	24.5	0.90	1.30	1.54
ANN.MLP	304.3	314.8	318.3	91.4	94.5	95.6	39.5	27.5	23.6	0.95	1.37	1.59

Table 10: Performance of schedule as function of storage capacity. *Wind plant.*

generated volume per spike is very high for the solar plant. Table 11 shows how the schedules based on the different price forecasts perform, when the system is based on generation from a solar plant.

Remark 1 Because of the many hours where there is no solar generation at all, the system dispatches less often, which is reflected in lower dispatch frequencies for all schedules. The schedules based on forecasts from REG.LASSO and ANN.MLP are affected most, as they dispatch more than 5% less often than in the case of a wind plant.

Remark 2 Both profits and dispatch frequencies for the different schedules are much more even in this case, which demonstrates that the solar generation profile reduces the gains from employing sophisticated forecasts.

	Profit (k€)	Rel. profit (%)	Dispatch frequency (%)	Dispatch volume (MWh)
Period	*Full*			
Direct feed	353.7	100	100	0.42
True price	345.8	97.8	21.0	1.79
NVE.DA	316.4	89.5	19.0	1.97
REG.LASSO	318.6	90.1	22.9	1.64
ANN.MLP	318.9	90.2	22.4	1.67

Table 11: Performance of schedule. *Solar plant. Storage capacity of 10 MWh.*

For a one-week period in 2018, for the system based on a solar plant, and for the schedule based on REG.LASSO, the system variables as function of time are shown in Figure M.2. What generally holds true for this system is that no energy is generated during night hours, when the ID3 price index is usually lower than during day hours.

Remark 3 Therefore, hours where solar generation is high generally coincide with high prices and the system is not often forced to store energy for very long. Figure M.2 shows it does occur, however, that not all of the stored energy can be dispatched during a local maximum, such that remaining energy is stored until the next occurrence of high price.

Remark 4 Because generation is rather sporadic, the schedules are very dependent on the battery capacity. In the one-week period shown in Figure M.2, there are already two occurrences where the battery is exploited near its maximum capacity of 10 MWh.

Impact of look-ahead-time With a higher storage capacity and greater look-ahead-time, the schedule becomes less of a repeated cycle that does not bump into constraints of battery storage as often, and that can spot high prices further away in time. Table 12 shows how the schedules based on the different look-ahead-times perform. The results are discussed only for the forecast from REG.LASSO, as very similar conclusions would be drawn for the forecasts from other models.

Remark 1 With further look-ahead-times and based on true price, the system amasses approx. 5% more profit as it becomes more capable to spot local minimum-maximum pairs that are more separated. The opposite is true for nearer look-ahead-times. Results regarding sensitivity to the storage capacity, shown in Table 10, suggest that that effect would be even stronger in the case of more storage capacity, as 10 MWh might constrain the ability to store energy over the full horizon.

Remark 2 Results of the forecasting procedure show that accuracies at look-ahead-times beyond 9 hours deteriorate rapidly, demonstrated by the S-shaped curve shown in Figure 25. While the forecast attains a MAE of 4.47 €/MWh for a look-ahead-time of 0 hours, the forecast attains MAEs of 5.50 €/MWh or higher at look-ahead-times greater than 9 hours, and MAEs of 7.00 €/MWh at look-ahead-times further than 12 hours. Despite that considerable reduction of accuracy, the system is able to amass approx. 3% more profit based on a forecast with a look-ahead-time of 20.

Remark 3 An important contribution to that effect is the way that the model predictive control optimization works. Even though the inputs that are deemed optimal for the steps with look-ahead-times further than 12 hours, i.e. $\{\boldsymbol{u}(13), \ldots, \boldsymbol{u}(20)\}$ might be rather inaccurate, the system does not execute those inputs right away, and executes only the first input $\boldsymbol{u}(0)$ before re-optimizing on a horizon that is shifted one step into the future. The step with a look-ahead-time of 13 hours now coincides with a look-ahead time of 12 hours and the employed price forecast is thus already more accurate. The system is thus slightly less sensitive to inaccuracies further along the prediction horizon, as long as they offer a reasonable indication of local maxima. Upon inspection of the forecasted series for near, moderate, and far look-ahead-times, shown in Figures J.3 through J.6, it is concluded that although the forecast becomes less detailed for further look-ahead-times, the estimation of the daily price cycle remains reasonably accurate.

	Profit (k€)			**Rel. profit (%)**			**Dispatch frequency (%)**			**Dispatch volume (MWh)**		
Period	*Full*											
Look-ahead-time	*5*	*9*	*20*	*5*	*9*	*20*	*5*	*9*	*20*	*5*	*9*	*20*
Direct feed	333.1			100			100			0.42		
True price	327.5	348.2	364.5	98.3	104.5	109.4	32.5	23.9	19.3	1.16	1.57	1.95
REG.LASSO	303.7	311.9	321.2	91.2	93.6	96.4	37.2	28.8	20.8	1.01	1.30	1.80

Table 12: Performance of schedule as function of look-ahead-time. *Wind plant. Storage capacity of 10 MWh.*

10 Conclusion

This research investigates the accuracies that naive, regression, and ANN models attain when forecasting an aggregated price of the Dutch intraday market. Six models for point forecasting and four models for interval forecasting are considered in single- and multi-step-ahead horizons, that employ Dutch and German features of price, calendar, generation, and load. A systematic forecasting procedure is established that limits the number of arbitrary choices and ensures that the out-of-sample test is similar to how forecasting is carried out in practice. To investigate whether theoretical and practical superiority coincide, point forecasts are deployed in an operational context where a battery dispatches generated energy according to a schedule optimized with model predictive control.

Dutch intraday price Dutch aggregated intraday price has statistical properties that are challenging for accurate forecasting. That includes high volatility, negative prices, and spikes, some of which are slightly more pronounced than for the often-studied German intraday market. The out-of-sample test is on the years of 2018 and 2019; the former represents a year with many high prices and high volatility, while the latter represents a year where prices exhibit less extreme dynamics.

Initialization The forecasting procedure starts with execution of an initialization procedure on the period from 2016 through 2017, that reveals how the full feature set can be reduced in size to improve generalizability and decrease computational cost. Results of correlation and reduction demonstrate that although features of generation and load slightly benefit the accuracy of intraday price forecasts, several features of price are clearly dominant in terms of usefulness. Besides that, results show that certain German features are useful—or even more useful than their Dutch counterparts. Results demonstrate that the size of the testing window hardly affects the accuracy of the regression and ANN models, given a large enough training window. It is thus concluded that monthly retraining suffices, and that evolution of the Dutch intraday market is not so rapid that regression and ANN models cannot reach (close to) their full potential if not retrained hourly, daily, or weekly.

Calibration & Exploitation The forecasting procedure continues with repeated execution of calibration and exploitation procedures on the period from 2018 through 2019, which is an out-of-sample test that systematically trains models and obtains forecasts in exploitation, and that repeatedly reconsiders the candidate feature set and hyperparameters in calibration. What is considered as the optimal model is thus not deemed to be a fixed entity. Results demonstrate that several dominant features of price are identified as most useful in almost every calibration, while the usefulness of other features varies considerably in each calibration. Results also show that the optimal values of hyperparameters vary considerably in each calibration.

Point forecasting Results for single-step-ahead (4-hour ahead) point forecasting demonstrate that day-ahead price dynamics offer a reasonable indication of intraday price dynamics. Still, the forecasts from the considered regression and ANN models attain better accuracies than the naive models, and are especially more dependable throughout day-hours that generally coincide with high prices. The regression and ANN models are thus capable to infer—to a greater or lesser extent—what drives intraday (ID3) prices to deviate from day-ahead (MCP) prices. In the out-of-sample test and the volatile year of 2018, the forecasts from the multilayer perceptron ANN attains the highest accuracies with an rMAE of 0.81 and a MAE of 6.55 €/MWh. Accuracies are much better for the less extreme year of 2019, where the same model attains an rMAE of 0.77 and a MAE of 4.33 €/MWh. Although at least 90% of the observations in the out-of-sample test lie in a price region for which accuracies do not deteriorate, results demonstrate that accuracies of all forecasts deteriorate for occurrences of extreme price. Although the mean of the residuals lies close to zero, residuals are distributed with fat tails, they still show some degree of serial correlation, and are heteroscedastic. It is thus concluded that the error is not random, which suggests that a more sophisticated approach might be able to capture the intraday price dynamics to a larger extent. For multi-step-ahead point forecasting, the main limiting factor becomes the fact that market clearing prices of the forecasted delivery hour are not available for look-ahead-times greater than 9 hours and results demonstrate that accuracies deteriorate to an increasing degree after that point, while remaining rather steady up to that point.

Interval forecasting Single-step-ahead interval forecasting requires a slightly more detailed evaluation. Results show that all the regression and ANN models obtain interval forecasts that attain an empirical coverage rate that deviate at most 4% (2018) and 2% (2019) from the nominal coverage rate. Results also show that the regression and ANN models are capable of achieving interval forecasts where the width varies considerably, and thus are capable to reveal more or less certainty about future prices. For the volatile year of 2018, the interval forecast from the naive model based on the most accurate point forecast provides a better interval forecast than both the regression and ANN models. That result suggests that in a challenging period with many high prices, a more specifically trained point forecasting model might represent a more dependable foundation for interval forecasting. For the more temperate year of 2019, results demonstrate that it becomes beneficial to utilize a more sophisticated interval forecasting model as both the regression and ANN models outperform the naive models. That comes at the cost of more quantile crossings, however, and the employed grid of 18 quantiles already leads to 157 observations with one or more quantile crossings in the forecast from ANN.Q.MLP. A model based on quantile regression averaging achieves the highest accuracy in the out-of-sample test, equal to a CRPS of 3.83 for 2018 and 2.24 for 2019. It is concluded that the regression and ANN models provide forecasts with adequate reliability and sharpness.

Simulation This research establishes a simulation to investigate the practical applicability of intraday price forecasts and to assess whether theoretically superior point forecasts result in more profitable dispatch schedules. To that end, point forecasts are deployed in the operational context of a generic system comprising an energy plant with storage capacity that dispatches generated energy after transacting on the Dutch intraday market. The dispatch strategy is optimized by means of model predictive control with full knowledge of generation and forecasted price in a receding horizon. The setup is deliberately simple, but does take efficiencies, cost of usage, and charge/discharge limitations into account. Results show that the dispatch strategy exploits especially estimation of local minimum-maximum pairs, and charges the battery when prices are relatively low to amplify the volume that is dispatched when prices are relatively high. Results demonstrate that the forecasts from all considered models are generally able to detect where local minima and maxima lie. As the strategy is largely dependent on the estimation of the timing of such pairs, and less so on the estimation of their absolute height, the dispatch strategies that arise from the different point forecasts result in very similar profits, such that the system attains a profit in 2018 with a schedule based on a more accurate forecast that is only 0.7% higher than that based on the least accurate forecast, despite the fact that the more accurate forecast attains a more than 2% higher accuracy in terms of sMAPE. Trading behaviour varies considerably when employing a different forecast, however, as the dispatch frequency with those forecasts varies more than 8% or more than 10% with a smaller battery. This research concludes that for the considered system, gains in accuracy are not reflected linearly in profit, but differences between forecasts do provoke the system to alter dispatch behaviour. Results demonstrate that the offered value is very sensitive to costs of using the battery; improvements of battery efficiency can enhance profits, and particularly higher storage capacities and charge/discharge speeds can improve the gains from employing a forecast-driven strategy. Furthermore, it is concluded that the considerable reduction of point forecast accuracies for longer look-ahead-times does not mislead the system and deteriorate profits, as the system amasses approx. 3% more profit with schedules based on forecasts for longer look-ahead-times.

11 Discussion

This research demonstrates that an aggregated price for the Dutch intraday market can be forecasted with more or less accuracy, by the considered naive, regression, and ANN models. Performance is evaluated in the established forecasting procedure on the out-of-sample period from 2018 through 2019, and can serve as benchmarks for future research. In the field of electricity price forecasting, many intricacies can easily blur what is deemed optimal in a real-world scenario. This research attempts be elaborate by considering both point and interval forecasting, in single- and multi-step-ahead horizons. Also, it considers various models that are more or less complex in terms of interpretability or in terms of training, and evaluation is taken well beyond the metrics of accuracy. As forecasting is often the means to a goal, intentions can lie far apart. There might still be discrepancies with the interest of practitioners for whom accuracy might be a less important consideration than model interpretability, or for whom it may be critical to recognize extreme prices, for instance. Nevertheless, approaches and results presented in this research can offer a foundation for those and for other specific real-world interests.

11.1 Avenues for future research

The following addresses several avenues for future research.

Individual trades To establish benchmark performance for the Dutch intraday market, this research employs a price index that is an aggregated price which is frequently employed for intraday price forecasting. Performance can thus be easily compared, also with results from (past or future) studies where availability of data on individual trades is limited. However, this research demonstrates that prices of individual trades can vary to a great extent within such an aggregated price, however. Because practical implementation of intraday price forecasts might depend on individual trades, future research might consider forecasting the distribution of individual trades, such that it can be estimated, for instance, when prices of individual trades are more or less scattered around the aggregated price.

Other and/or more advanced models and approaches This research offers a foundation for an intraday market that is not earlier studied. For that reason, this research does not employ the most advanced models and approaches in all stages of the forecasting procedure. Therefore, it is encouraged that future research considers other and/or more advanced models and approaches that might outperform the benchmark accuracies put forward in this research. Explicit examples of this are approaches that combine forecasts from multiple models [20], that ultimately utilize those forecasts that are assumed to be most accurate at the time of forecasting, such that they often outperform individual models. Another example is the recently proposed LASSO regularized quantile regression averaging for interval forecasting [101] that smartly “selects” what point forecasting models to include in the quantile regression.

Feature transformations Future research might investigate to a greater extent the possible transformations of features. Although many transformations might benefit performance of models similarly and therefore might not change the perspectives on superiority, they might represent straightforward and undemanding means to more accurate forecasts. An explicit example are the many different variance stabilizing transformations investigated in [102].

Residual errors Results of residual error analysis demonstrate that the point forecast that attains the highest overall accuracy is not able to capture all dynamics of the target variable. It is encouraged that future research utilizes a more iterative procedure, where model adequacy is fed back and model building is repeated until it provides satisfactory forecasts [90]. It is also encouraged that future research investigates whether results from a similar analysis can be used to improve accuracy, or whether different models lead to different residual error characteristics.

Stochastic model predictive control For the case study, this research employs deterministic point forecasts as input to an optimization in a setting of model predictive control. Future research might consider optimization that incorporates uncertainty. An explicit example is using the bounds of interval forecasts as an indication of certainty, as in [101]. More advanced is deriving scenario forecasts from interval forecasts [76] and utilizing two-stage or multi-stage stochastic model predictive control [44, 42]. A stochastic approach complicates the optimization, but results in a case study that is more representative of reality, wherein more sophisticated strategies can be exploited and evaluated.

More realistic trading In the case study, this research does not take into account actual trading on the intraday market. Future research might incorporate a more realistic simulation of trading. An explicit example is the incorporation that the strategy obtains profit only when asks/bids are matched by actual orders—both in price and in volume—and that otherwise it might be necessary to enter the balancing market. The day-ahead market could be integrated into the system as well, so that the strategy enters the intraday market only when there are opportunities on top of accepted day-ahead offers, for instance. When more realistic trading is incorporated in the case study, future research might also consider more sophisticated metrics of profit and risk. A relevant metric might then be the Sharpe ratio, which summarizes the first two moments of the return distribution as the ratio of return over volatility.

Unconstrained charging When the constraint on charging $B_I(k) = 100\% \cdot E_I(k)$ is removed from the case study, it can be investigated whether the dispatch schedules diverge if the system is free to choose whenever to charge and discharge the battery. In that case, energy that the system chooses not to supply to the battery, i.e. $E_I(t) - B_I(t)$, can be considered as energy that may be dispatched to a different destination.

Close the loop This research stresses that differences in accuracy, and superiority of forecasts that are confirmed not to be by chance, do not necessarily translate to more or less profitable trading or dispatch strategies. Rather, strategies can be highly dependent on certain time ranges, price ranges, or other detailed accuracies that undermine the superiority that is evaluated on the basis of overall accuracy. Future research might therefore consider catering the forecasting procedure towards obtaining forecasts that benefit a simulated strategy. For instance, by utilizing a metric of profit as loss function that would 'close the loop' between the forecast and the strategy, which might lead to novel insights on superiority.

11.2 Limitations

The following addresses several limitations of this research.

Computational power This research pays attention to the repeated optimization of hyperparameters, as well as to proper training of neural networks. Due to limited availability of computational power it is not able to be exhaustive in every regard, however. In hyperparameter optimization, for instance, more computational power could underlie a finer-grain search space, as well as more iterations for the search. It is expected that that might slightly improve accuracy.

Separate models per delivery hour This research does provide extensive analysis of separate models per delivery hour, although it might improve overall accuracy, as well as improve accuracy for particularly the time ranges where the occurrence of extreme prices are relatively high. Moreover, separate models for every hour of the day could utilize a separate input feature set, which could particularly improve the accuracies of multi-step-ahead forecasts. The availability of market clearing prices could then be variable, instead of being restricted at the minimum of 9 hours for all delivery hours.

Coarse grid of quantiles This research does not evaluate a fine grid of quantile forecasts, but instead focuses evaluation on only one narrow, one medium, and one wide interval. They are assumed to be representative of the performance across the whole distribution, although a more accurate estimation of the true distribution requires more quantiles.

Quantile crossing The crossing of quantiles is often addressed by re-sorting quantiles, although more sophisticated approaches exist. It is assumed that resolving the quantile crossings has only a minor effect on the evaluation of accuracy contained in this research, thus the number of quantiles crossings is mentioned but they are not resolved. In the case that a finer grid of quantiles is utilized, the effect becomes more severe, however. What is more, quantiles that are exploited in a stochastic optimization, for instance, should never cross.

Raw data This research employs a lot of raw data on Dutch and German generation and load that originates from one source [30]. Besides examining whether there is strong correlation among the features of these connected markets, which might suggest that data is reliable, there are no explicit checks of trustworthiness. As this research does not harness any second-opinion on reliability, it can not be guaranteed and might be limited.

Intentionally blank

Bibliography (114)

[1] Nima Amjady and Farshid Keynia. “A new prediction strategy for price spike forecasting of day-ahead electricity markets”. In: *Applied Soft Computing* 11.6 (Sept. 2011), pp. 4246–4256. DOI: 10.1016/j.asoc.2011.03.024 (Ref. on page XV).

[2] Nima Amjady and Farshid Keynia. “Day-ahead price forecasting of electricity markets by a new feature selection algorithm and cascaded neural network technique”. In: *Energy Conversion and Management* 50.12 (Dec. 2009), pp. 2976–2982. DOI: 10.1016/j.enconman.2009.07.016 (Ref. on pages 13, 33).

[3] José Andrade et al. “Probabilistic price forecasting for day-ahead and intraday markets: Beyond the statistical model”. In: *Sustainability* 9.11 (Oct. 2017), p. 1990. DOI: 10.3390/su9111990 (Ref. on pages 26, XVII).

[4] Riccardo R. Appino et al. “Energy-based stochastic MPC for integrated electricity-hydrogen VPP in real-time markets”. In: *Electric Power Systems Research* 195 (June 2021), p. 106738. DOI: 10.1016/j.epsr.2020.106738 (Ref. on page XV).

[5] Muhammad Ardalani-Farsa and Saeed Zolfaghari. “Residual analysis and combination of embedding theorem and artificial intelligence in chaotic time series forecasting”. In: *Applied Artificial Intelligence* 25.1 (Jan. 2011), pp. 45–73. DOI: 10.1080/08839514.2011.529263 (Ref. on page 16).

[6] J. Scott Armstrong. “Evaluating forecasting methods”. In: *International Series in Operations Research & Management Science*. Springer US, 2001, pp. 443–472. URL: http://dx.doi.org/10.1007/978-0-306-47630-3_20 (Ref. on page 2).

[7] J. Scott Armstrong and Kesten C. Green. “Forecasting methods and principles: Evidence-based checklists”. In: *Journal of Global Scholars of Marketing Science* 28.2 (Mar. 2018), pp. 103–159. DOI: 10.1080/21639159.2018.1441735 (Ref. on page 2).

[8] Mariette Awad and Rahul Khanna. “Support Vector Regression”. In: *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress, Apr. 2015, pp. 67–80. URL: https://doi.org/10.1007/978-1-4302-5990-9_4 (Ref. on page 18).

[9] Souhaib Ben Taieb et al. “A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition”. In: *Expert Systems with Applications* 39.8 (June 2012), pp. 7067–7083. DOI: 10.1016/j.eswa.2012.01.039 (Ref. on page 30).

[10] James Bergstra et al. “Algorithms for hyper-parameter optimization”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor J. Shawe-Taylor. Vol. 24. Curran Associates, Inc., 2011, pp. 2546–2554 (Ref. on page 28).

[11] Jonathan Berrisch and Florian Ziel. “CRPS Learning”. In: 2021. URL: https://arxiv.org/abs/2102.00968 (Ref. on page 24).

[12] Svetlana Borovkova and Maren Diane Schmeck. “Electricity price modeling with stochastic time change”. In: *Energy Economics* 63 (Mar. 2017), pp. 51–65. DOI: 10.1016/j.eneco.2017.01.002 (Ref. on page 9).

[13] BP. *Statistical Review of World Energy*. 2021 (Ref. on page 4).

[14] Alex J. Cannon. “Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes”. In: *Stochastic Environmental Research and Risk Assessment* 32.11 (June 2018), pp. 3207–3225. DOI: 10.1007/s00477-018-1573-6 (Ref. on page 22).

[15] Álvaro Cartea and Marcelo G. Figueroa. “Pricing in electricity markets: A mean reverting jump diffusion model with seasonality”. In: *Applied Mathematical Finance* 12.4 (Dec. 2005), pp. 313–335. DOI: 10.1080/13504860500117503 (Ref. on pages 1, 8).

[16] Mauro Costantini and Robert M. Kunst. “On using predictive-ability tests in the selection of time-series prediction models: A Monte Carlo evaluation”. In: *International Journal of Forecasting* 37.2 (Apr. 2021), pp. 445–460. DOI: 10.1016/j.ijforecast.2020.06.010 (Ref. on page 15).

[17] Sumeyra Demir et al. “Introducing technical indicators to electricity price forecasting: A feature engineering study for linear, ensemble, and deep machine learning models”. In: *Applied Sciences* 10.1 (Dec. 2019), p. 255. DOI: 10.3390/app10010255 (Ref. on page 28).

[18] Francis X. Diebold. “Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests”. In: *Journal of Business & Economic Statistics* 33.1 (Jan. 2015), pp. 1–1. DOI: 10.1080/07350015.2014.983236 (Ref. on page 15).

[19] Francis X. Diebold and Roberto S. Mariano. “Comparing predictive accuracy”. In: *Journal of Business & Economic Statistics* 13.3 (July 1995), p. 253. DOI: 10.2307/1392185 (Ref. on page 15).

[20] Francis X. Diebold and Minchul Shin. “Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives”. In: *International Journal of Forecasting* 35.4 (Oct. 2019), pp. 1679–1691. DOI: 10.1016/j.ijforecast.2018.09.006 (Ref. on page 58).

[21] Zhao Y. Dong, Tapan K. Saha, and Kit P. Wong. “Artificial intelligence in electricity market operations and management”. In: *Business Applications and Computational Intelligence*. Ed. by Kevin E. Voges and Nigel K. Ll. Pope. IGI Global, 2005, pp. 131–154. URL: http://dx.doi.org/10.4018/978-1-59140-702-7.ch008 (Ref. on page XV).

[22] EEX. *Annual Report 2015*. 2016 (Ref. on page 5).

[23] EEX. *Annual Report 2016*. 2017 (Ref. on pages 1, 5).

[24] EEX. *Annual Report 2017*. 2018 (Ref. on pages 1, 5).

[25] EEX. *Annual Report 2018*. 2019 (Ref. on pages 1, 5, 7).

[26] EEX. *Annual Report 2019*. 2020 (Ref. on pages 1, 5).

[27] EEX. *Annual Report 2020*. 2021 (Ref. on pages 1, 7, 10).

[28] EEX. *Data of EPEX SPOT markets*. Webshop EEX Group. URL: https://webshop.eex-group.com/epex-spot-public-market-data (Ref. on pages 6, 7, 8, 9, 10, 11, IX, X, XI, XII, XIII, XIV).

[29] Energy Charts. *Data of German energy price*. Energy Charts Dashboard. URL: https://energy-charts.info/index.html?l=en&c=DE (Ref. on pages 11, XI, XII).

[30] ENTSO-E. *Data of European energy price, generation, and load*. ENTSO-E Transparency Platform. URL: https://transparency.entsoe.eu/ (Ref. on pages 11, 46, 60).

[31] EPEX SPOT. *Description of EPEX SPOT indices*. 2020 (Ref. on page 12).

[32] Raphael Espinoza, Fabio Fornari, and Marco J. Lombardi. “The Role of Financial Variables in predicting economic activity”. In: *Journal of Forecasting* 31.1 (Feb. 2011), pp. 15–46. DOI: 10.1002/for.1212 (Ref. on page 15).

[33] Carlo Fezzi and Luca Mosetti. “Size matters: Estimation sample length and electricity price forecasting accuracy”. In: *The Energy Journal* 41.4 (Oct. 2020). DOI: 10.5547/01956574.41.4.cfez (Ref. on pages 26, 27).

[34] Hélyette Geman and Andrea Roncoroni. “Understanding the fine structure of electricity prices*”. In: *The Journal of Business* 79.3 (May 2006), pp. 1225–1261. DOI: 10.1086/500675 (Ref. on page 13).

[35] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. “Learning to Forget: Continual Prediction with LSTM”. In: *Neural Computation* 12.10 (Oct. 2000), pp. 2451–2471. DOI: 10.1162/089976600300015015 (Ref. on page 20).

[36] Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. “Learning Precise Timing with LSTM Recurrent Networks”. In: *Journal of Machine Learning Research* 3 (Aug. 2002), pp. 115–143. DOI: 10.1162/153244303768966139 (Ref. on page 20).

[37] Raffaella Giacomini and Halbert White. “Tests of conditional predictive ability”. In: *Econometrica* 74.6 (Nov. 2006), pp. 1545–1578. DOI: 10.1111/j.1468-0262.2006.00718.x (Ref. on page 15).

[38] Angelica Gianfreda, Francesco Ravazzolo, and Luca Rossini. “Comparing the forecasting performances of linear models for electricity prices with high RES penetration”. In: *International Journal of Forecasting* 36.3 (July 2020), pp. 974–986. DOI: 10.1016/j.ijforecast.2019.11.002 (Ref. on page 1).

[39] Tilmann Gneiting. “Making and evaluating point forecasts”. In: *Journal of the American Statistical Association* 106.494 (June 2011), pp. 746–762. DOI: 10.1198/jasa.2011.r10138 (Ref. on page 14).

[40] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. “Probabilistic forecasts, calibration and sharpness”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (Apr. 2007), pp. 243–268. DOI: 10.1111/j.1467-9868.2007.00587.x (Ref. on page 24).

[41] Fazil Gökgöz and Fahrettin Filiz. “Electricity price forecasting in Turkey with artificial neural network models”. In: *Investment Management and Financial Innovations* 13.3 (Sept. 2016), pp. 150–158. DOI: 10.21511/imfi.13(3-1).2016.01 (Ref. on page 19).

[42] Edwin González et al. “A comparative study of stochastic model predictive controllers”. In: *Electronics* 9.12 (Dec. 2020), p. 2078. DOI: 10.3390/electronics9122078 (Ref. on page 59).

[43] Shadi Goodarzi, H. Niles Perera, and Derek Bunn. “The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices”. In: *Energy Policy* 134 (Nov. 2019), p. 110827. DOI: 10.1016/j.enpol.2019.06.035 (Ref. on page 1).

[44] Arne Groß et al. “Using probabilistic forecasts in stochastic optimization”. In: *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, Aug. 2020. URL: http://dx.doi.org/10.1109/pmaps47429.2020.9183575 (Ref. on page 59).

[45] José Guajardo, Richard Weber, and Jaime Miranda. “A forecasting methodology using support vector regression and dynamic feature selection”. In: *Journal of Information & Knowledge Management* 05.04 (Dec. 2006), pp. 329–335. DOI: 10.1142/s021964920600158x (Ref. on page 18).

[46] Lars Ivar Hagfors et al. “Prediction of extreme price occurrences in the German day-ahead electricity market”. In: *Quantitative Finance* 16.12 (Sept. 2016), pp. 1929–1948. DOI: 10.1080/14697688.2016.1211794 (Ref. on page XV).

[47] Rodrigo Herrera and Nicolás González. “The modeling and forecasting of extreme events in electricity spot markets”. In: *International Journal of Forecasting* 30.3 (July 2014), pp. 477–490. DOI: 10.1016/j.ijforecast.2013.12.011 (Ref. on page 15).

[48] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (Ref. on page 20).

[49] Tao Hong and Shu Fan. “Probabilistic electric load forecasting: A tutorial review”. In: *International Journal of Forecasting* 32.3 (July 2016), pp. 914–938. DOI: 10.1016/j.ijforecast.2015.11.011 (Ref. on page 24).

[50] Katarzyna Hubicka, Grzegorz Marcjasz, and Rafal Weron. “A note on averaging day-ahead electricity price forecasts across calibration windows”. In: *IEEE Transactions on Sustainable Energy* 10.1 (Jan. 2019), pp. 321–323. DOI: 10.1109/tste.2018.2869557 (Ref. on page 27).

[51] Joanna Janczura et al. “Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling”. In: *SSRN Electronic Journal* (2012). DOI: 10.2139/ssrn.2081738 (Ref. on pages 9, 10).

[52] Tim Janke and Florian Steinke. “Forecasting the price distribution of continuous intraday electricity trading”. In: *Energies* 12.22 (Nov. 2019), p. 4262. DOI: 10.3390/en12224262 (Ref. on pages 6, 15, 30, 31, 35).

[53] Orhan Karabiber and George Xydis. “Electricity price forecasting in the danish day-ahead market using the TBATS, ANN and ARIMA methods”. In: *Energies* 12.5 (Mar. 2019), p. 928. DOI: 10.3390/en12050928 (Ref. on page 16).

[54] Christopher Kath and Florian Ziel. “Conformal prediction interval estimation and applications to day-ahead and intraday power markets”. In: *International Journal of Forecasting* 37.2 (Apr. 2021), pp. 777–799. DOI: 10.1016/j.ijforecast.2020.09.006 (Ref. on pages i, 42, 43).

[55] Dogan Keles et al. “Comparison of extended mean-reversion and time series models for electricity spot price simulation considering negative prices”. In: *Energy Economics* 34.4 (July 2012), pp. 1012–1032. DOI: 10.1016/j.eneco.2011.08.012 (Ref. on pages 8, 9).

[56] *KERAS library for ANNs, for PYTHON*. URL: https://keras.io/ (Ref. on page 33).

[57] Thomas Kneib. “Beyond mean regression”. In: *Statistical Modelling* 13.4 (Aug. 2013), pp. 275–303. DOI: 10.1177/1471082x13494159 (Ref. on page 23).

[58] Roger Koenker. *Quantile regression*. Cambridge University Press, 2005. URL: http://dx.doi.org/10.1017/cbo9780511754098 (Ref. on page 23).

[59] Irena Koprinska, Mashud Rana, and Vassilios G. Agelidis. “Correlation and instance based feature selection for electricity load forecasting”. In: *Knowledge-Based Systems* 82 (July 2015), pp. 29–40. DOI: 10.1016/j.knosys.2015.02.017 (Ref. on pages 13, 33).

[60] Marcel Kremer, Rüdiger Kiesel, and Florentina Paraschiv. “Intraday electricity pricing of night contracts”. In: *Energies* 13.17 (Sept. 2020), p. 4501. DOI: 10.3390/en13174501 (Ref. on page 31).

[61] Dheepak Krishnamurthy et al. “Energy storage arbitrage under day-ahead and real-time price uncertainty”. In: *IEEE Transactions on Power Systems* 33.1 (Jan. 2018), pp. 84–93. DOI: 10.1109/tpwrs.2017.2685347 (Ref. on page 5).

[62] Jesus Lago, Fjo De Ridder, and Bart De Schutter. “Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms”. In: *Applied Energy* 221 (July 2018), pp. 386–405. DOI: 10.1016/j.apenergy.2018.02.069 (Ref. on pages 16, 19, 21).

[63] Jesus Lago et al. “Forecasting day-ahead electricity prices in Europe: The importance of considering market integration”. In: *Applied Energy* 211 (Feb. 2018), pp. 890–903. DOI: 10.1016/j.apenergy.2017.11.098 (Ref. on page 11).

[64] Jesus Lago et al. “Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark”. In: *Applied Energy* 293 (July 2021), p. 116983. DOI: 10.1016/j.apenergy.2021.116983 (Ref. on pages 2, 14).

[65] R. Laref et al. “On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications”. In: *Chemometrics and Intelligent Laboratory Systems* 184 (Jan. 2019), pp. 22–27. DOI: 10.1016/j.chemolab.2018.11.011 (Ref. on page 18).

[66] Kun Li et al. “The significance of calendar effects in the electricity market”. In: *Applied Energy* 235 (Feb. 2019), pp. 487–494. DOI: 10.1016/j.apenergy.2018.10.124 (Ref. on page 9).

[67] Katarzyna Maciejowska, Weronika Nitka, and Tomasz Weron. “Day-Ahead vs. intraday—forecasting the price spread to maximize economic benefits”. In: *Energies* 12.4 (Feb. 2019), p. 631. DOI: 10.3390/en12040631 (Ref. on pages 2, 26, XV).

[68] Reinhard Madlener and Oliver Ruhnau. “Variable renewables and demand flexibility: Day-ahead versus intraday valuation”. In: *Variable Generation, Flexible Demand*. Elsevier, 2021, pp. 309–327. URL: http://dx.doi.org/10.1016/b978-0-12-823810-3.00005-4 (Ref. on page 4).

[69] Grzegorz Marcjasz, Jesus Lago, and Rafał Weron. “Neural networks in day-ahead electricity price forecasting: single vs. multiple outputs”. In: 2020. URL: https://arxiv.org/abs/2008.08006 (Ref. on page 31).

[70] Grzegorz Marcjasz, Tomasz Serafin, and Rafał Weron. “Selection of calibration windows for day-ahead electricity price forecasting”. In: *Energies* 11.9 (Sept. 2018), p. 2364. DOI: 10.3390/en11092364 (Ref. on page 27).

[71] Grzegorz Marcjasz, Bartosz Uniejewski, and Rafał Weron. “Beating the naïve—combining LASSO with naïve intraday electricity price forecasts”. In: *Energies* 13.7 (Apr. 2020), p. 1667. DOI: 10.3390/en13071667 (Ref. on pages 2, 6, 15, 16).

[72] Rodrigo de Marcos, Antonio Bello, and Javier Reneses. “Short-Term electricity price forecasting with a composite fundamental-econometric hybrid methodology”. In: *Energies* 12.6 (Mar. 2019), p. 1067. DOI: 10.3390/en12061067 (Ref. on page 14).

[73] Henry Martin and Scott Otterson. “German intraday electricity market analysis and modeling based on the limit order book”. In: *2018 15th International Conference on the European Energy Market (EEM)*. IEEE, June 2018. URL: http://dx.doi.org/10.1109/eem.2018.8469829 (Ref. on page 5).

[74] Robert May, Graeme Dandy, and Holger Maier. “Review of input variable selection methods for artificial neural networks”. In: *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. InTech, Apr. 2011. URL: http://dx.doi.org/10.5772/16004 (Ref. on page XV).

[75] Klaus Mayer, Thomas Schmid, and Florian Weber. “Modeling electricity spot prices: Combining mean reversion, spikes, and stochastic volatility”. In: *The European Journal of Finance* 21.4 (Sept. 2012), pp. 292–315. DOI: 10.1080/1351847x.2012.716775 (Ref. on pages 8, 10).

[76] Dennis van der Meer, Guang Chao Wang, and Joakim Munkhammar. “An alternative optimal strategy for stochastic model predictive control of a residential battery energy management system with solar photovoltaic”. In: *Applied Energy* 283 (Feb. 2021), p. 116289. DOI: 10.1016/j.apenergy.2020.116289 (Ref. on page 59).

[77] Atom Mirakyan, Martin Meyer-Renschhausen, and Andreas Koch. “Composite forecasting approach, application for next-day electricity price forecasting”. In: *Energy Economics* 66 (Aug. 2017), pp. 228–237. DOI: 10.1016/j.eneco.2017.06.020 (Ref. on page 38).

[78] Michał Narajewski and Florian Ziel. “Econometric modelling and forecasting of intraday electricity prices”. In: *Journal of Commodity Markets* 19 (Sept. 2020), p. 100107. DOI: 10.1016/j.jcomm.2019.100107 (Ref. on pages 2, 15, 16, 26).

[79] Bijay Neupane, Wei Woon, and Zeyar Aung. “Ensemble prediction model with expert selection for electricity price forecasting”. In: *Energies* 10.1 (Jan. 2017), p. 77. DOI: 10.3390/en10010077 (Ref. on page 11).

[80] Whitney K. Newey and James L. Powell. “Asymmetric least squares estimation and testing”. In: *Econometrica* 55.4 (July 1987), p. 819. DOI: 10.2307/1911031 (Ref. on page 23).

[81] Jakub Nowotarski and Rafał Weron. “Computing electricity spot price prediction intervals using quantile regression and forecast averaging”. In: *Computational Statistics* 30.3 (Aug. 2014), pp. 791–803. DOI: 10.1007/s00180-014-0523-0 (Ref. on page 25).

[82] Amparo Núñez-Reyes et al. “Optimal scheduling of grid-connected PV plants with energy storage for integration in the electricity market”. In: *Solar Energy* 144 (Mar. 2017), pp. 502–516. DOI: 10.1016/j.solener.2016.12.034 (Ref. on pages 46, 47, 50).

[83] Yicun Ouyang and Hujun Yin. “Multi-Step time series forecasting with an ensemble of varied length mixture models”. In: *International Journal of Neural Systems* 28.04 (Mar. 2018), p. 1750053. DOI: 10.1142/s0129065717500538 (Ref. on page 30).

[84] Pablo M Pincheira. “Conditional predictive ability of exchange rates in long run regressions”. In: *Revista de análisis económico* 28.2 (Oct. 2013), pp. 3–35. DOI: 10.4067/s0718-88702013000200001 (Ref. on page 15).

[85] Pierre Pinson, Christophe Chevallier, and George N. Kariniotakis. “Trading wind generation from short-term probabilistic forecasts of wind power”. In: *IEEE Transactions on Power Systems* 22.3 (Aug. 2007), pp. 1148–1156. DOI: 10.1109/tpwrs.2007.901117 (Ref. on pages 2, 23).

[86] The Wind Power. *Data of wind farm Beabuorren in Friesland, The Netherlands*. Wind farm database. URL: https://www.thewindpower.net/windfarm_en_6264_beabuorren.php (Ref. on page 47).

[87] Alessandro Sapio. “Econometric modelling and forecasting of wholesale electricity prices”. In: *Handbook of Energy Economics and Policy: Fundamentals and Applications for Engineers and Energy Planners*. Elsevier, May 2021, pp. 595–640. URL: https://doi.org/10.1016/B978-0-12-814712-2.00015-4 (Ref. on page 8).

[88] Simon Schnürch and Andreas Wagner. “Electricity price forecasting with neural networks on EPEX order books”. In: *Applied Mathematical Finance* 27.3 (May 2020), pp. 189–206. DOI: 10.1080/1350486x.2020.1805337 (Ref. on page XVII).

[89] *SCIKIT-LEARN library for machine learning, for PYTHON*. URL: https://scikit-learn.org/stable/ (Ref. on page 33).

[90] Mohammad Shahidehpour, Hatim Yamin, and Zuyi Li. *Electricity Price Forecasting*. John Wiley & Sons, Inc., Apr. 2002. URL: http://dx.doi.org/10.1002/047122412x (Ref. on pages 13, 58).

[91] Leili Shahriyari. “Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma”. In: *Briefings in Bioinformatics* 20.3 (Nov. 2017), pp. 985–994. DOI: 10.1093/bib/bbx153 (Ref. on page 13).

[92] Deepak Singhal and K.S. Swarup. “Electricity price forecasting using artificial neural networks”. In: *International Journal of Electrical Power & Energy Systems* 33.3 (Mar. 2011), pp. 550–555. DOI: 10.1016/j.ijepes.2010.12.009 (Ref. on page 19).

[93] Fereidoon Sioshansi. *Variable generation, flexible demand*. Academic Press, Nov. 2020. ISBN: 9780128241912 (Ref. on page 5).

[94] E. Snieder, R. Shakir, and U.T. Khan. “A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models”. In: *Journal of Hydrology* 583 (Apr. 2020), p. 124299. DOI: 10.1016/j.jhydrol.2019.124299 (Ref. on pages 13, XV).

[95] Miriam Steurer, Robert J. Hill, and Norbert Pfeifer. “Metrics for evaluating the performance of machine learning based automated valuation models”. In: *Journal of Property Research* 38.2 (Apr. 2021), pp. 99–129. DOI: 10.1080/09599916.2020.1858937 (Ref. on page XX).

[96] Akylas Stratigakos, Andrea Michiorri, and Georges Kariniotakis. “A value-oriented price forecasting approach to optimize trading of renewable generation”. In: *2021 IEEE Madrid PowerTech*. IEEE, June 2021. URL: http://dx.doi.org/10.1109/powertech46648.2021.9494832 (Ref. on page 46).

[97] James W. Taylor. “Evaluating quantile-bounded and expectile-bounded interval forecasts”. In: *International Journal of Forecasting* 37.2 (Apr. 2021), pp. 800–811. DOI: 10.1016/j.ijforecast.2020.09.007 (Ref. on pages 22, 23).

[98] Tesla. *Specifications of Tesla Megapack*. Tesla Megapack. URL: https://www.tesla.com/megapack (Ref. on page 47).

[99] Bartosz Uniejewski, Grzegorz Marcjasz, and Rafał Weron. “Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO”. In: *International Journal of Forecasting* 35.4 (Oct. 2019), pp. 1533–1547. DOI: 10.1016/j.ijforecast.2019.02.001 (Ref. on pages 6, 15).

[100] Bartosz Uniejewski, Jakub Nowotarski, and Rafał Weron. “Automated variable selection and shrinkage for day-ahead electricity price forecasting”. In: *Energies* 9.8 (Aug. 2016), p. 621. DOI: 10.3390/en9080621 (Ref. on page XV).

[101] Bartosz Uniejewski and Rafał Weron. “Regularized quantile regression averaging for probabilistic electricity price forecasting”. In: *Energy Economics* 95 (Mar. 2021), p. 105121. DOI: 10.1016/j.eneco.2021.105121 (Ref. on pages i, 25, 46, 58, 59).

[102] Bartosz Uniejewski, Rafal Weron, and Florian Ziel. “Variance stabilizing transformations for electricity spot price forecasting”. In: *IEEE Transactions on Power Systems* 33.2 (Mar. 2018), pp. 2219–2229. DOI: 10.1109/tpwrs.2017.2734563 (Ref. on page 58).

[103] Lennard Visser, Tarek AlSkaif, and Wilfried van Sark. “The importance of predictor variables and feature selection in day-ahead electricity price forecasting”. In: *2020 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, Sept. 2020. URL: http://dx.doi.org/10.1109/sest48500.2020.9203273 (Ref. on pages i, 11, 13, 33, 35, 36, XV).

[104] Dao H. Vu et al. “Short-Term forecasting of electricity spot prices containing random spikes using a time-varying autoregressive model combined with kernel regression”. In: *IEEE Transactions on Industrial Informatics* 15.9 (Sept. 2019), pp. 5378–5388. DOI: 10.1109/tii.2019.2911700 (Ref. on page 15).

[105] Linda Schulze Waltrup et al. “Expectile and quantile regression—David and Goliath?” In: *Statistical Modelling* 15.5 (Dec. 2014), pp. 433–456. DOI: 10.1177/1471082x14561155 (Ref. on page 23).

[106] Jianzhou Wang, Ling Xiao, and Jun Shi. “The combination forecasting of electricity price based on price spikes processing: A case study in South Australia”. In: *Abstract and Applied Analysis* 2014 (2014), pp. 1–12. DOI: 10.1155/2014/172306 (Ref. on page 13).

[107] Yi Wang et al. “Combining probabilistic load forecasts”. In: *IEEE Transactions on Smart Grid* 10.4 (July 2019), pp. 3664–3674. DOI: 10.1109/tsg.2018.2833869 (Ref. on page 22).

[108] Rafał Weron. “Market price of risk implied by Asian-style electricity options and futures”. In: *Energy Economics* 30.3 (May 2008), pp. 1098–1115. DOI: 10.1016/j.eneco.2007.05.004 (Ref. on page 13).

[109] Lei Wu and Mohammad Shahidehpour. “A hybrid model for day-ahead price forecasting”. In: *IEEE Transactions on Power Systems* 25.3 (Aug. 2010), pp. 1519–1530. DOI: 10.1109/tpwrs.2009.2039948 (Ref. on page 16).

[110] *YALMIP toolbox for modeling and optimization, for MATLAB*. URL: https://yalmip.github.io/ (Ref. on page 50).

[111] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. “On early stopping in gradient descent learning”. In: *Constructive Approximation* 26.2 (Apr. 2007), pp. 289–315. DOI: 10.1007/s00365-006-0663-2 (Ref. on page 30).

[112] Thomas W Yee. “Quantile and Expectile Regression”. In: *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, 2015, pp. 415–445. URL: https://doi.org/10.1007/978-1-4939-2818-7_15 (Ref. on page 23).

[113] Guo-Bing Zhou et al. “Minimal gated unit for recurrent neural networks”. In: *International Journal of Automation and Computing* 13.3 (June 2016), pp. 226–234. DOI: 10.1007/s11633-016-1006-2 (Ref. on page 20).

[114] H. Zhou et al. “Study on probability distribution of prices in electricity market: A case study of zhejiang province, china”. In: *Communications in Nonlinear Science and Numerical Simulation* 14.5 (May 2009), pp. 2255–2265. DOI: 10.1016/j.cnsns.2008.04.020 (Ref. on page 7).

Intentionally blank

Appendix

A Order book

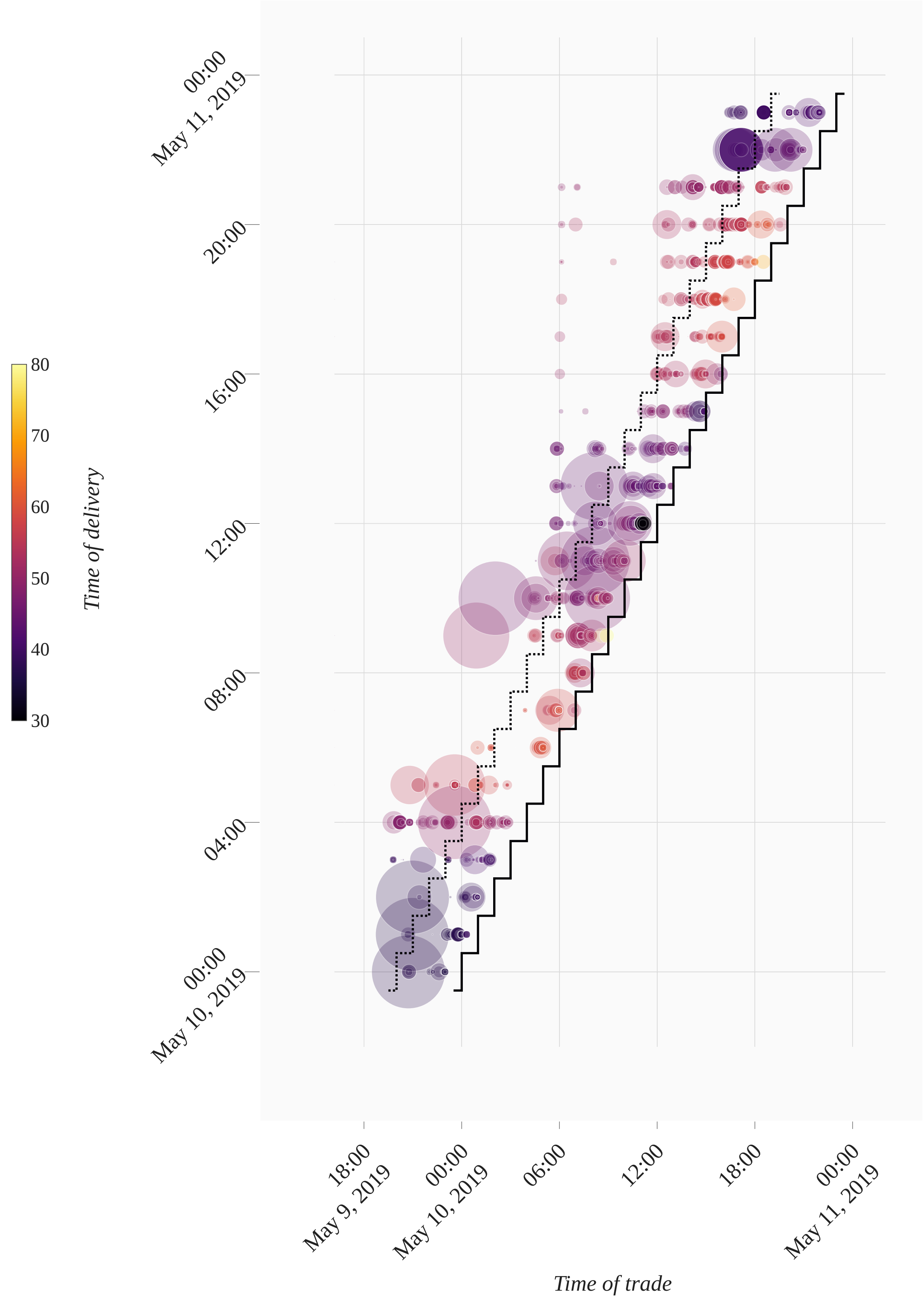


Figure A.1: Time, volume, and price of Dutch continuous hourly intraday trades. *Included are trades for all products with delivery on 2019.05.10. Size represents volume, color represents price. The line represents the time of delivery* t_d. *The dotted line represents the time of forecasting* $t_f = t_d - 4$. *Data from EEX [28].*

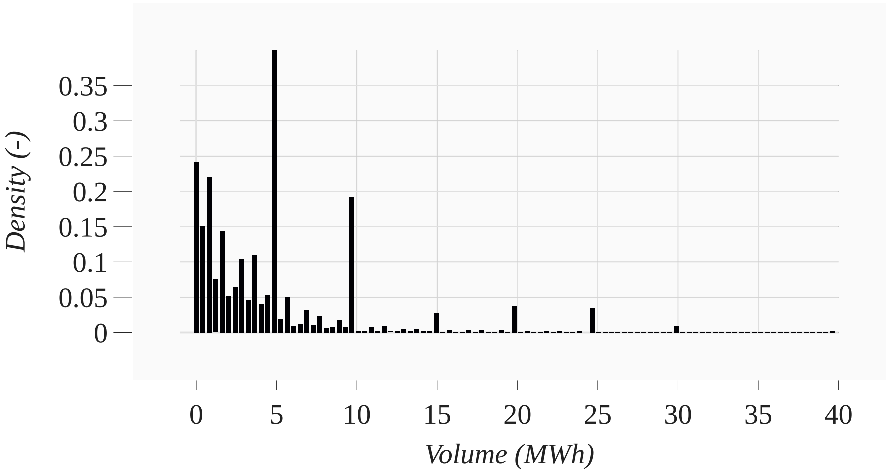


Figure A.2: Empirical distribution of the volume of offers for continuous hourly products of the Dutch intraday market. *2015–2020. Data from EEX [28].*

B Dutch and German intraday markets

Statistical property		
Region	*NL*	*DE*
Min.	-121.31	
Max.	523.68	
Mean	40.933	
Std.	18.027	
Skew.	3.1521	
Kurt.	43.348	
Obs.	52608	

Table B.1: Statistical properties of the Dutch and German ID3 indices *2015–2020. Data from EEX [28] and Energy Charts [29].*

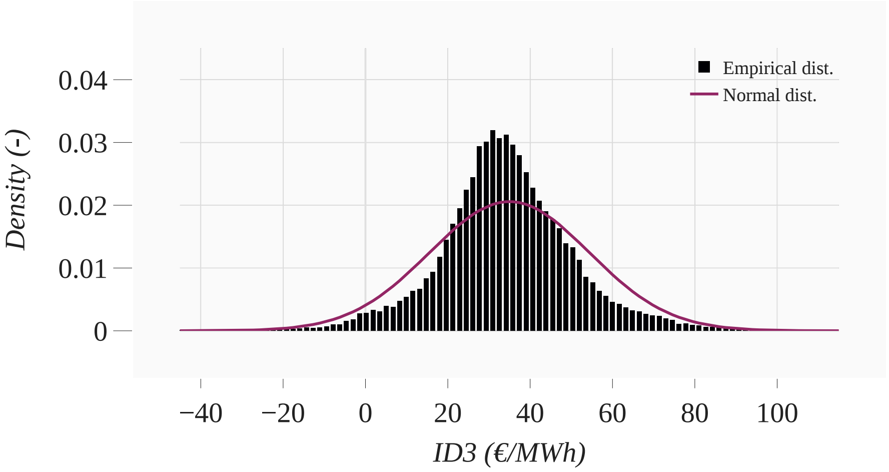


Figure B.1: Empirical distribution of the German ID3 index. *2015–2020. Data from Energy Charts [29]*

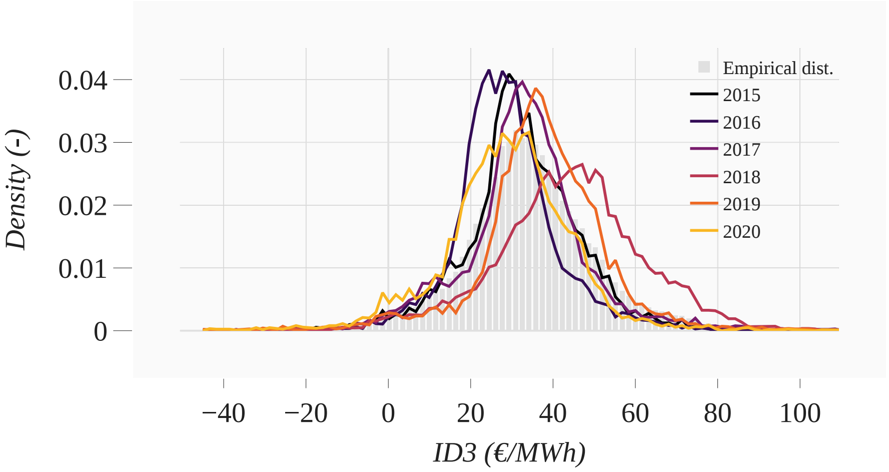


Figure B.2: Empirical distribution of the German ID3 index as function of year. *Data from Energy Charts [29]*

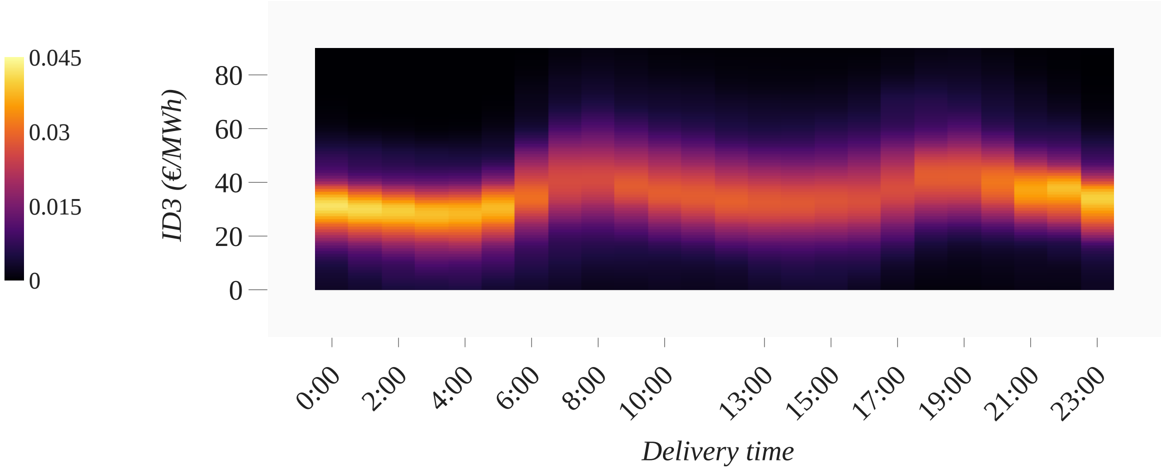
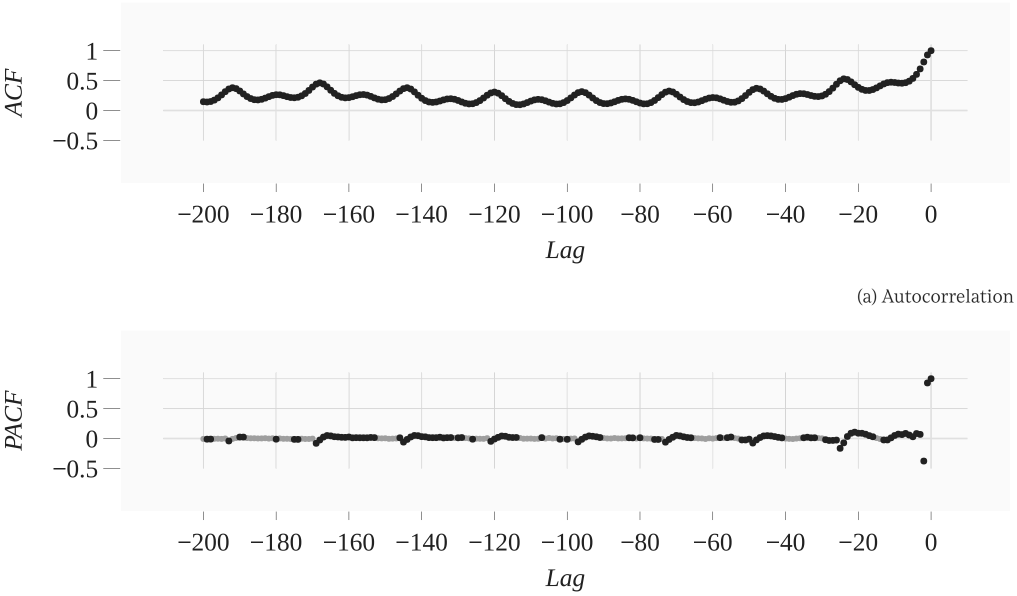


Figure B.3: Kernel density estimates of the German ID3 index as function of delivery time. *Color represents density. 2015–2020. Data from Energy Charts [29]*



(a) Autocorrelation

(b) Partial autocorrelation

Figure B.4: Autocorrelation and partial autocorrelation of the German ID3 index. *Coefficients that are within the 95% confidence interval are a lighter shade. 2015–2020. Data from Energy Charts [29]*

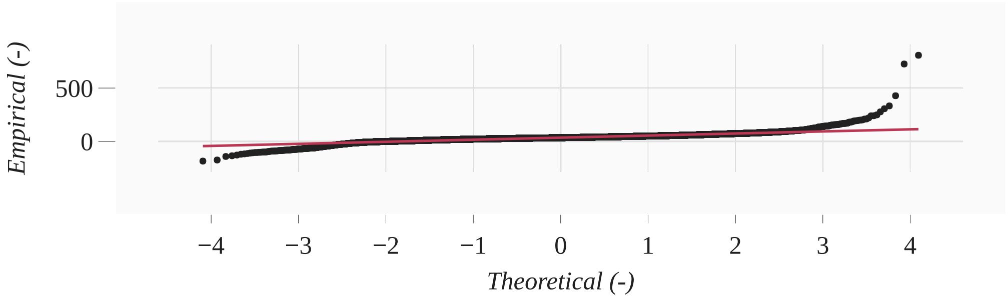


Figure B.5: Normality test for the German ID3 index. *2015–2020. Data from EEX [28].*

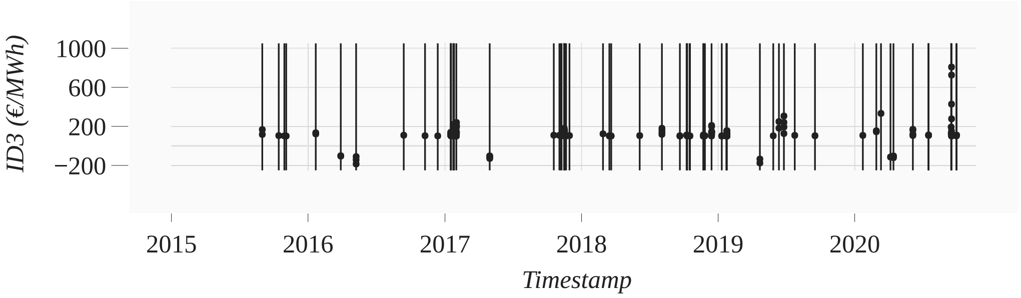


Figure B.6: Occurrence of extreme prices in the German ID3 index. *Data from EEX [28].*

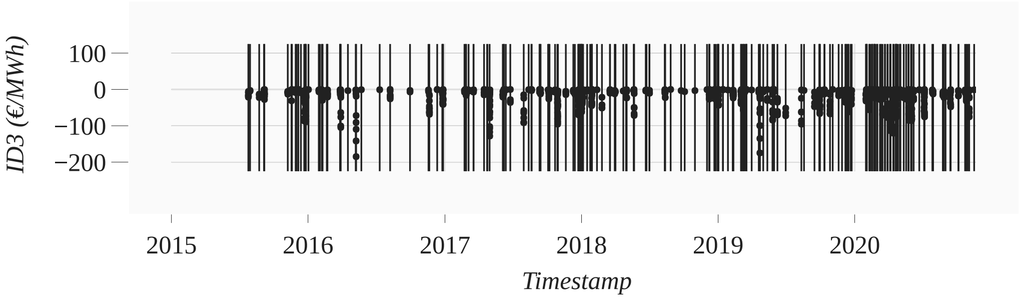
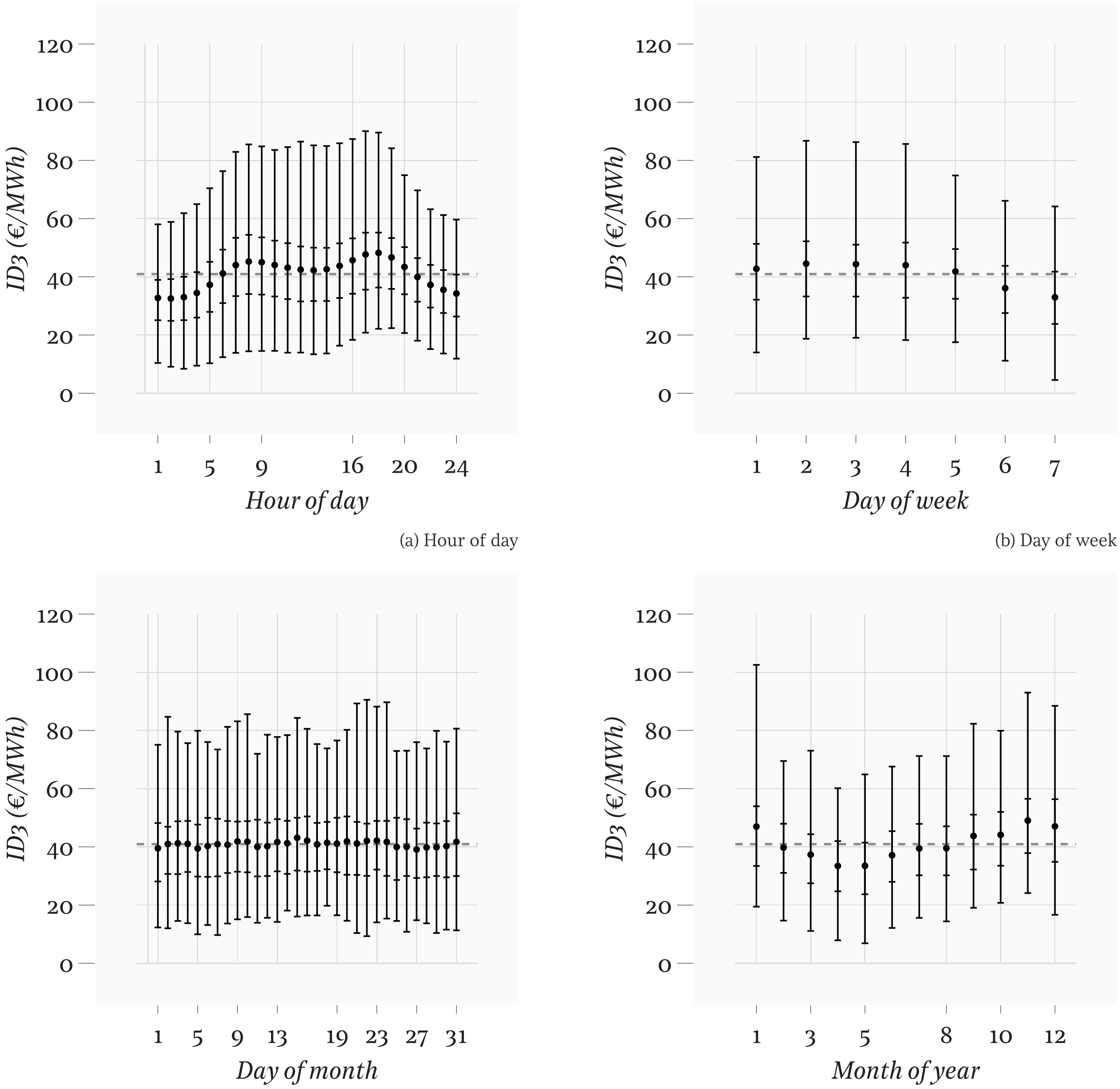


Figure B.7: Occurrence of negative prices in the German ID3 index. *Data from EEX [28].*



(a) Hour of day

(b) Day of week

(c) Day of month

(d) Month of year

Figure B.8: Seasonality of intraday price on different timescales. *Data from EEX [28].*

C Features

Scaling Two often-found scaling approaches are referred to as normalization and standardization. In scaling by means of normalization, all observations t of a feature f are scaled so that they all lie in the interval $[0, 1]$, i.e. $x'_{f,t} = (x_{f,t} - \min \boldsymbol{x}_f)/(\max \boldsymbol{x}_f - \min \boldsymbol{x}_f)$. This involves a linear scaling of the distribution, which might introduce vulnerabilities when outliers are so extreme that the majority of the values are pushed into a short interval. In scaling by means of standardization, all observations of a feature are scaled so that they have zero mean and unit variance, i.e. $x'_{f,t} = (x_{f,t} - \bar{\boldsymbol{x}}_f)/\sigma_{\boldsymbol{x}_f}$.

Scaling Two often-found scaling approaches are referred to as normalization and standardization. In scaling by means of normalization, all observations t of a feature f are scaled so that they all lie in the interval $[0, 1]$, i.e. $x'_{f,t} = (x_{f,t} - \min \boldsymbol{x}_f)/(\max \boldsymbol{x}_f - \min \boldsymbol{x}_f)$. This involves a linear scaling of the distribution, which might introduce vulnerabilities when outliers are so extreme that the majority of the values are pushed into a short interval. In scaling by means of standardization, all observations of a feature are scaled so that they have zero mean and unit variance, i.e. $x'_{f,t} = (x_{f,t} - \bar{\boldsymbol{x}}_f)/\sigma_{\boldsymbol{x}_f}$.

Spikes Energy prices are characterized by the occurrence of spikes, and Figure 9 confirms that to be true for the Dutch intraday market. Spikes can affect forecasting accuracies negatively because many forecasting models are sensitive to extreme observations [1]. The presence of spikes might therefore limit models to infer relationships properly not only in the tails, but in the centre of the distribution as well. That is not to say that spikes are always strictly unwanted in EPF; the ability to forecast spikes can be a major source of profits, and literature underlines the importance to certain market participants [67, 46, 21, 4]. Because spikes can also cause major losses in the slightest amount of time, it seems compulsory for market participants to arm themselves. There is no necessity for a one-fits-all approach; forecasting procedures that have a stage for 'normal' price forecasting and a stage for spike forecasting are not uncommon [21]. The specific design and evaluation of models around capability to forecast spikes lies outside the scope of this research, however.

Spikes Energy prices are characterized by the occurrence of spikes, and Figure 9 confirms that to be true for the Dutch intraday market. Spikes can affect forecasting accuracies negatively because many forecasting models are sensitive to extreme observations [1]. The presence of spikes might therefore limit models to infer relationships properly not only in the tails, but in the centre of the distribution as well. That is not to say that spikes are always strictly unwanted in EPF; the ability to forecast spikes can be a major source of profits, and literature underlines the importance to certain market participants [67, 46, 21, 4]. Because spikes can also cause major losses in the slightest amount of time, it seems compulsory for market participants to arm themselves. There is no necessity for a one-fits-all approach; forecasting procedures that have a stage for 'normal' price forecasting and a stage for spike forecasting are not uncommon [21]. The specific design and evaluation of models around capability to forecast spikes lies outside the scope of this research, however.

Feature selection Typically, feature selection approaches are based on identifying what the most useful input features are from a set of candidate features. 'Usefulness' referring to what features have the highest relevancy to the output, subject to a condition of minimal redundancy to other input features [94, 74]. A proper reduction of the full feature set has the potential of lowering computational cost, enhancing the interpretability of results, combating the curse of dimensionality, and limiting output variability. The large feature space of the full feature set shown in Table 2 is bound to result in problems with regard to the curse of dimensionality and computational cost. An initial stage of feature selection is thus imperative. Particularly [100] emphasizes the importance of embedded feature selection in the form of least absolute shrinkage and selection operator (LASSO) regression and elastic net regression. Approaches similar to those utilised in LASSO and elastic net regression might not be optimal for models that are of a nonlinear and non-parametric nature like ANNs, however [74]. One of the very few researches devoted to the Dutch electricity market, [103], contains an extensive analysis of feature selection. It utilises four models in a framework of recursive feature elimination (RFE). A finding of [103] is that the input feature set derived from the relative feature importance of a support vector regression (SVR) is optimal for all considered models, and that the model that attains the highest accuracy, regardless of the input feature set, is based on random forest (RF) regression. Absolutely optimal is thus the RF regression combined with the input feature set derived from the relative feature importance of an SVR.

D Raw data to full feature set

Price features F_{p2} through F_{p5} contain the average values of the ID3 index during the same day last week, during the 24 hours preceding t_f, during the 168 hours (7 days) preceding t_f, and during the 5040 hours (30 days) preceding t_f, respectively.

The MCP of the day-ahead market is published by EPEX SPOT for every hour of the day. It is therefore very straightforward to obtain F_{p6}, the MCPs of the day-ahead market, from R_{p3}. The number of MCPs that are available depends on whether the time of forecasting is before or after the time that R_{p3} is updated (12:00); the number of future values that are available is minimal (equal to 12) when the time of forecasting t_f is 11:00 and maximal (equal to 36) when t_f is 12:00. The features contained in F_{p6} are the lags of one week ago and of two days ago $\{t_d - 168,\ t_d - 48\}$ and all lags/leads from one day ago through 12 hours after the time of forecasting $[t_d - 24,\ t_f + 12]$.

F_{p7} through F_{p10} contain the average values of the MCP, and are analogous to those of the ID3 index.

Calendar features F_{c1} through F_{c4}, the hour of day, day of week, month of year, and day of year, should not be represented by their numerical values (e.g. 0–23 for hour of day) because that leads to a discontinuity between the end and start of successive cycles. One option is to encode these features by a one-hot-encoding, where every value except one becomes a separate binary column. It is important to leave one value out to prevent that the correlation matrix of features is singular. In such a categorical encoding all information of successive hours and days is lost. Although that is not necessarily problematic for low cardinality and a large enough training set, it leads to many additional features, 29 features to represent only the hour of the day and the day of the week. To limit the amount of features at this stage, one can also use a numerical encoding with built-in cyclicity, using a two dimensional circle-projection by means of sine and cosine transformations, for instance.

For features related to time, one-hot-encoding is more often found in literature, although examples of cyclical transformations can be found as well [3, 88]. To the best knowledge of the author, EPF literature lacks a comparison of these two encoding approaches. The research employs cyclical transformations for F_{c1} through F_{c4}.

Generation & Load features F_{g1} and F_{l1}, the actual generation and load, follow without calculation from R_{g1} and R_{l1}. The features contained in F_{g1} and F_{l1} are the lags from one day ago to one hour before the time of forecasting $[t_d - 24,\ t_f - 1]$.

F_{g2} and F_{l2}, the forecasted generation and load, follow without calculation from R_{g2} and R_{l2}. The number of lags/lead that are available depends on whether the time of forecasting is before or after the time that R_{g2} and R_{l2} are updated (18:00); the number of future values that are available is minimal (equal to 5) when the time of forecasting t_f is 17:00 and maximal (equal to 29) when t_f is 18:00.

The features contained in F_{g2} and F_{l2} are all lags/leads from one hour through 5 hours after the time of forecasting $[t_f + 1, t_d + 5]$. F_{g4} and F_{l3}, the forecast error of generation and load, are the difference between actual and forecasted generation and load. To take into account whether there was over- and underestimation, they are calculated by subtraction, *not* by absolute difference. The features contained in F_{g4} and F_{l3} are all lags from one day before delivery through one hour before the time of forecasting $[t_d - 24,\ t_f - 1]$.

There is also a second type of raw data on generation R_{g3}, which has only a German variant. This is an updated forecast of generation that is published during the day of delivery. F_{g3} and F_{g5}, the intraday variants of generation forecast and error, follow from R_{g1} and R_{g3}. The number of lags/lead that are available depends on whether the time of forecasting is before or after the time that R_{g3} is updated (08:00); the number of future values that are available is minimal (equal to 0) when the time of forecasting t_f is anywhere from 23:00 through 07:00 and maximal (equal to 15) when t_f is 08:00. The features contained in F_{g3} are all lags from one day before delivery through the time of forecasting $[t_d - 24,\ t_f]$. The features contained in F_{g5} are all lags from one day before delivery through one hour before the time of forecasting $[t_d - 24,\ t_f - 1]$.

In the full feature set, all features on generation (actual, forecast, and error) are separate for solar, onshore wind, and offshore wind plants.

E Candidate feature set

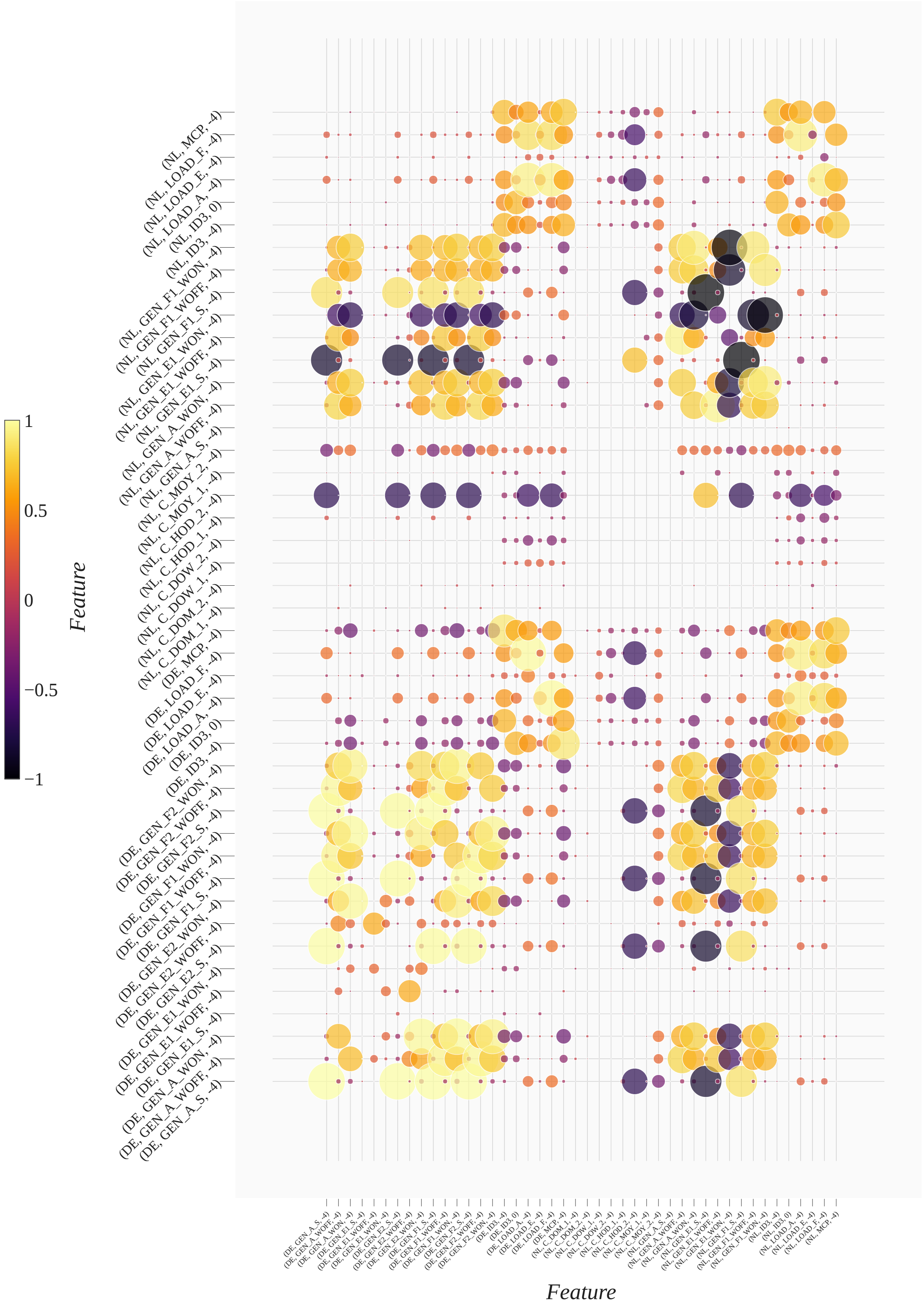


Figure E.1: 2016 & 2017. Correlation of features in the candidate set with lag of -4. *Size represents absolute correlation coefficient, color represents correlation coefficient. For size reference, the auto-correlation coefficient (equal to 1) of (DE, GEN_F1_S, 0) is included. (NL, ID3, 0) and (DE, ID3, 0) are not in the candidate feature set, but is included to investigate correlation.*

F Properties of scoring functions

One of the properties of scoring functions is whether errors are scored proportionally. The squared error, for instance, has a slope that grows ever-steeper as forecasted and realized values lie further apart, and it is therefore disproportional in penalization. Strategies based on minimization of the squared error are thus biased towards reducing errors in the tails as opposed to reducing errors in the centre of the distribution. A second property is whether the score has the same dimension as the input. Of the scores discussed earlier, only the absolute error has that property. In the field of EPF, it can thus be interpreted directly as a €/MWh score. Furthermore, there is the property whether two problems that are different absolutely but equal relatively are scored equally, i.e. whether $S(a,\ b) = S(c,\ d)$ given $a/b = c/d$. That property might be important to consider when, for instance, the score should be independent of the currency of the problem. Finally, then there is the property whether over or under predictions that are equal in absolute terms are scored by equal magnitude, i.e. whether $S(a,\ b) = |S(b,\ a)|$ [95].

G Forecasting procedure

```
Require: model($\boldsymbol{h}$) and $\boldsymbol{X}$
Require: $T$, $I$
Require: $N_{\text{tst}} > 0$, $N_{\text{val}} > 0$, $N_{\text{trn}} > 0$, start, and end

$W_{\text{tst}} \leftarrow [\ start,\ start + N_{\text{tst}}\ )$                                   ▷ Set testing window
$W_{\text{val}} \leftarrow [\ W_{\text{tst}}[0] - N_{\text{val}},\ W_{\text{tst}}[0]\ )$            ▷ Set validating window
$W_{\text{trn}} \leftarrow [\ W_{\text{val}}[0] - N_{\text{trn}},\ W_{\text{val}}[0]\ )$            ▷ Set training window

$t \leftarrow 0$                                                                                    ▷ Set observation

while $W_{\text{tst}}[0] \leq end$ do

    $\boldsymbol{X}_{\text{tst}} \leftarrow \boldsymbol{X}(W_{\text{tst}})$                        ▷ Set testing set
    $\boldsymbol{X}_{\text{val}} \leftarrow \boldsymbol{X}(W_{\text{val}})$                        ▷ Set validating set
    $\boldsymbol{X}_{\text{trn}} \leftarrow \boldsymbol{X}(W_{\text{trn}})$                        ▷ Set training set

    if $t \pmod T = 0$ then

        while $\boldsymbol{X}_{\text{tst}}$ has more than 15 features do

            fit model($\boldsymbol{h}$) on $\boldsymbol{X}_{\text{trn}}$
            reduce $\boldsymbol{X}_{\text{tst}}$, $\boldsymbol{X}_{\text{val}}$, $\boldsymbol{X}_{\text{trn}}$ by 1 feature    ▷ RFE

        end while

        $i \leftarrow 0$                                                                            ▷ Set iteration

        while $i < I$ do

            select $\boldsymbol{h}_i$                                                               ▷ TPE
            fit model($\boldsymbol{h}_i$) on $\boldsymbol{X}_{\text{trn}}$
            employ model($\boldsymbol{h}_i$) on $\boldsymbol{X}_{\text{val}}$

            $i \leftarrow i + 1$                                                                    ▷ Update iteration

        end while

        $\boldsymbol{h} \leftarrow$ optimal $\boldsymbol{h}_i$

    end if

    fit model($\boldsymbol{h}$) on $\boldsymbol{X}_{\text{trn}}$, $\boldsymbol{X}_{\text{val}}$
    employ model($\boldsymbol{h}$) on $\boldsymbol{X}_{\text{tst}}$

    $W_{\text{tst}} \leftarrow [\ W_{\text{tst}}[0] + N_{\text{tst}},\ W_{\text{tst}}[-1] + N_{\text{tst}}\ )$    ▷ Update testing window
    $W_{\text{val}} \leftarrow [\ W_{\text{val}}[0] + N_{\text{tst}},\ W_{\text{val}}[-1] + N_{\text{tst}}\ )$    ▷ Update validating window
    $W_{\text{trn}} \leftarrow [\ W_{\text{trn}}[0] + N_{\text{tst}},\ W_{\text{trn}}[-1] + N_{\text{tst}}\ )$    ▷ Update training window

    $t \leftarrow t + 1$                                                                            ▷ Update observation

end while
```

Table G.1: Pseudocode walkthrough of the exploitation procedure

H Hyperparameters of considered models

Regularization parameter λ	Uniform $[0, 10]$

Table H.1: Hyperparameters of REG.LASSO

Regularization parameter C	Uniform $[0, 20]$
Standard deviation of the Gaussian kernel σ	Uniform $[0, 20]$

Table H.2: Hyperparameters of REG.SVR

Number of hidden layers	Choice $\{1, 2\}$
Number of neurons (1 layer)	Choice $\{25, 50, 75, 100, 150, 200\}$
Number of neurons (2 layers)	Choice $\{25, 50, 75, 100, 150, 200\}$
Dropout rate	Choice $\{0, 0.1, 0.2\}$
Activation function	Choice $\{\text{tanh}, \text{relu}\}$

Table H.3: Hyperparameters of ANN.MLP

Number of hidden layers	Choice $\{1, 2\}$
Number of neurons (1 layer)	Choice $\{25, 50, 75, 100, 150, 200\}$
Number of neurons (2 layers)	Choice $\{25, 50, 75, 100, 150, 200\}$
Dropout rate	Choice $\{0, 0.1, 0.2\}$
Activation function	Choice $\{\text{tanh}, \text{relu}\}$

Table H.4: Hyperparameters of ANN.GRU

I ANN models

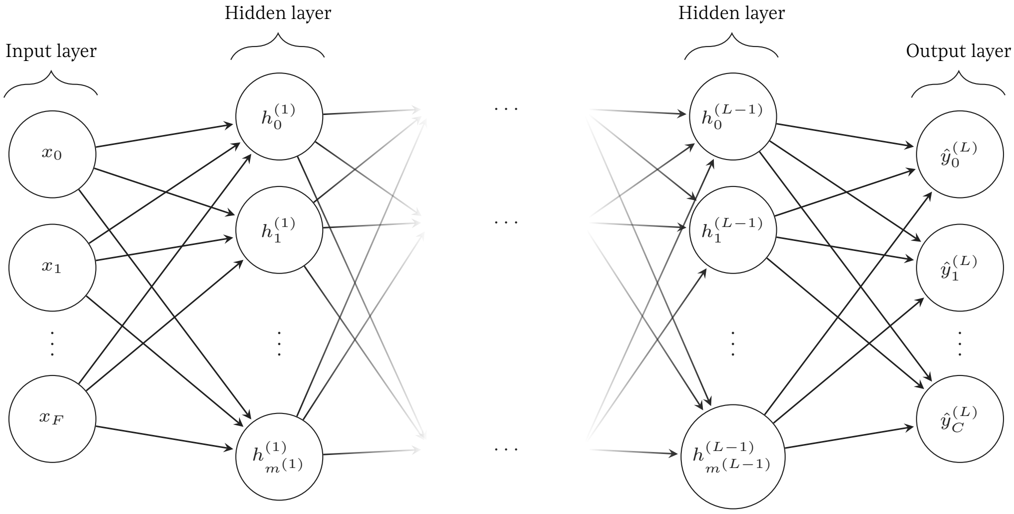


Figure I.1: Architecture of an MLP. *For F inputs, C outputs, and $m^{(l)}$ neurons per hidden layer.*

Number of epochs	Choice $[50, 100, 200]$
Batch size	Choice $[16, 32, 64, 128, 256]$
Learning rate	0.01
Early stopping min_delta	0.1
Early stopping patience	25

Table I.1: Parameters of training.

J Forecast visualization

Single-step-ahead

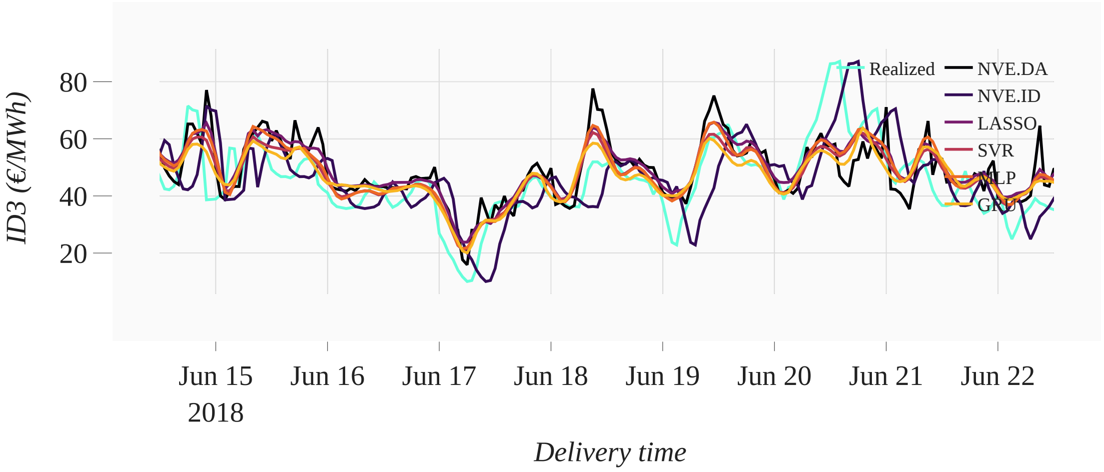


Figure J.1: Visualization of a slice of the forecasted and realized series. *Single-step-ahead point forecasting. Exploitation procedure.*

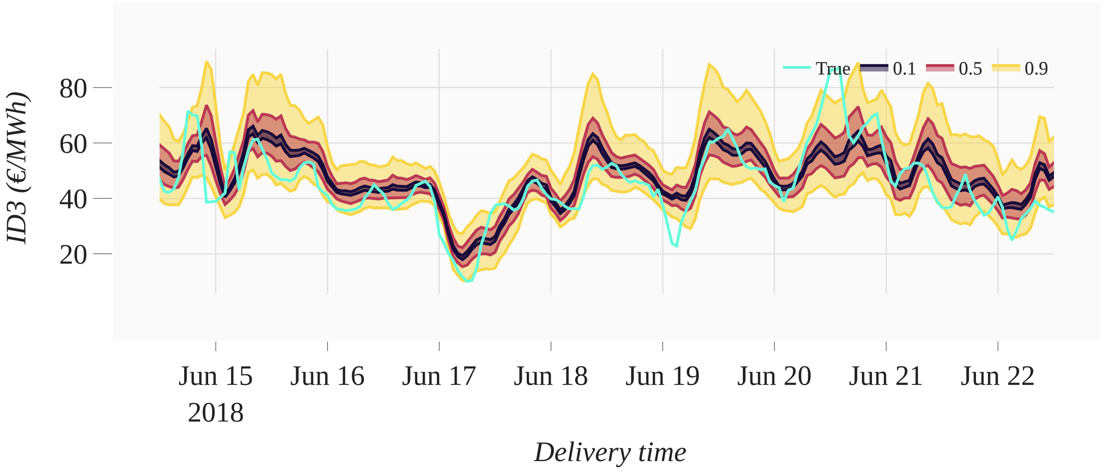


Figure J.2: Visualization of a slice of the forecasted and realized series. *ANN.MLP.Q. Single-step-ahead interval forecasting. Exploitation procedure.*

Multi-step-ahead

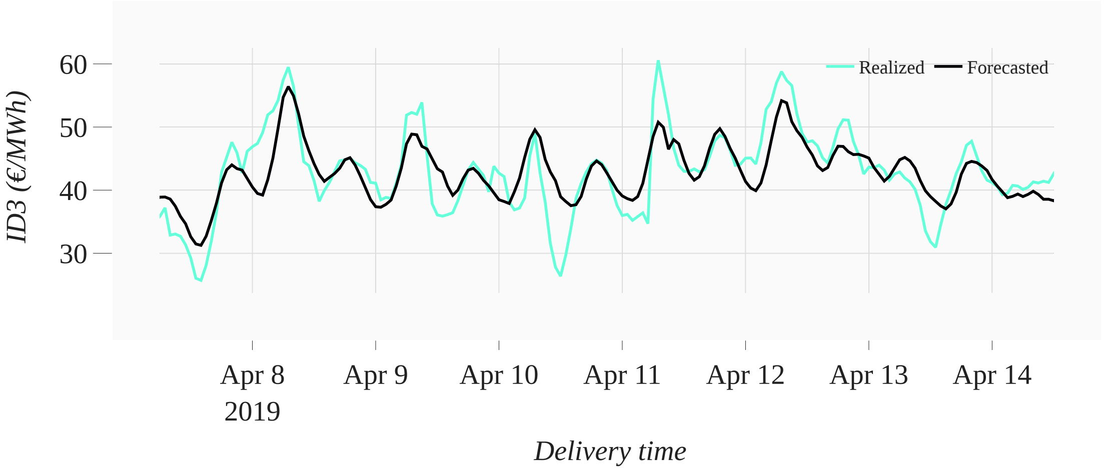


Figure J.3: Visualization of a slice of the forecasted and realized series. *REG.LASSO. Multi-step-ahead point forecasting. Look-ahead-time of 0 hours. Exploitation procedure.*

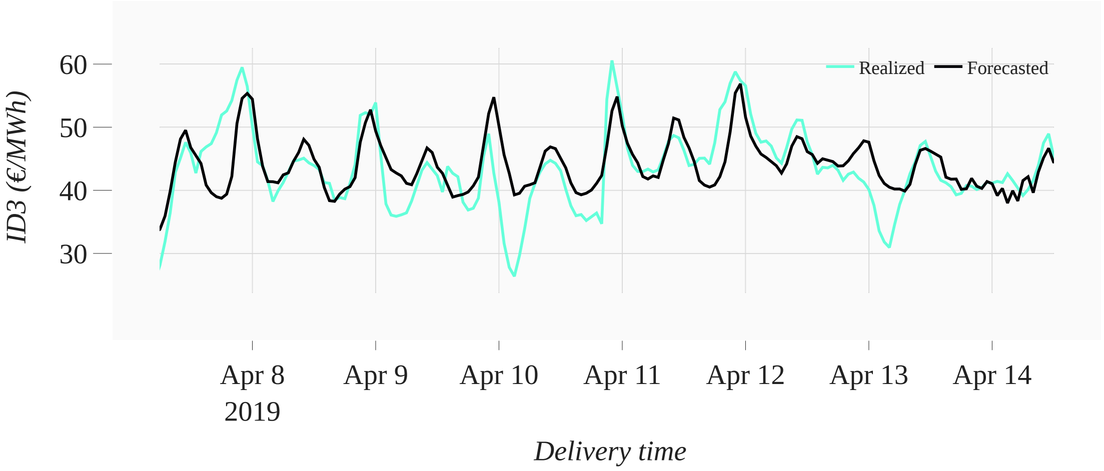


Figure J.4: Visualization of a slice of the forecasted and realized series. *REG.LASSO. Multi-step-ahead point forecasting. Look-ahead-time of 9 hours. Exploitation procedure.*

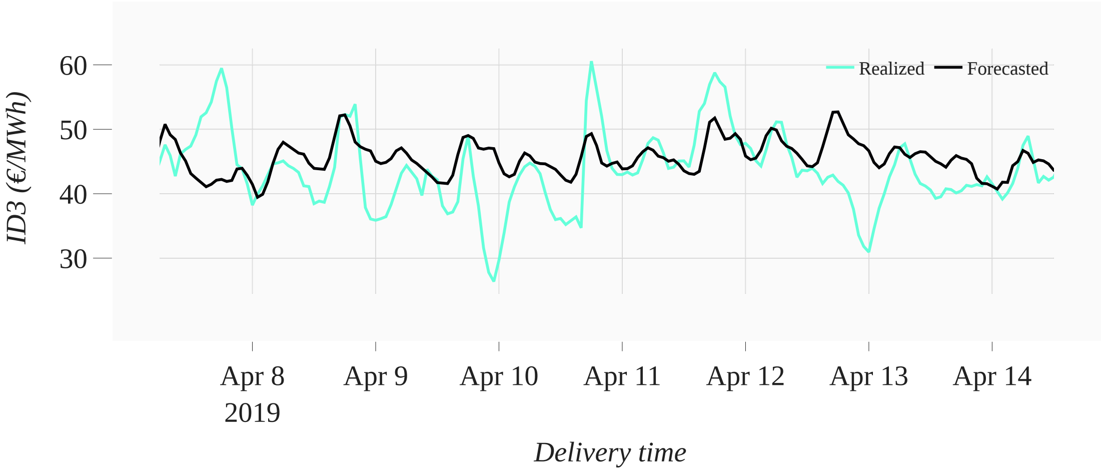


Figure J.5: Visualization of a slice of the forecasted and realized series. *REG.LASSO. Multi-step-ahead point forecasting. Look-ahead-time of 13 hours. Exploitation procedure.*

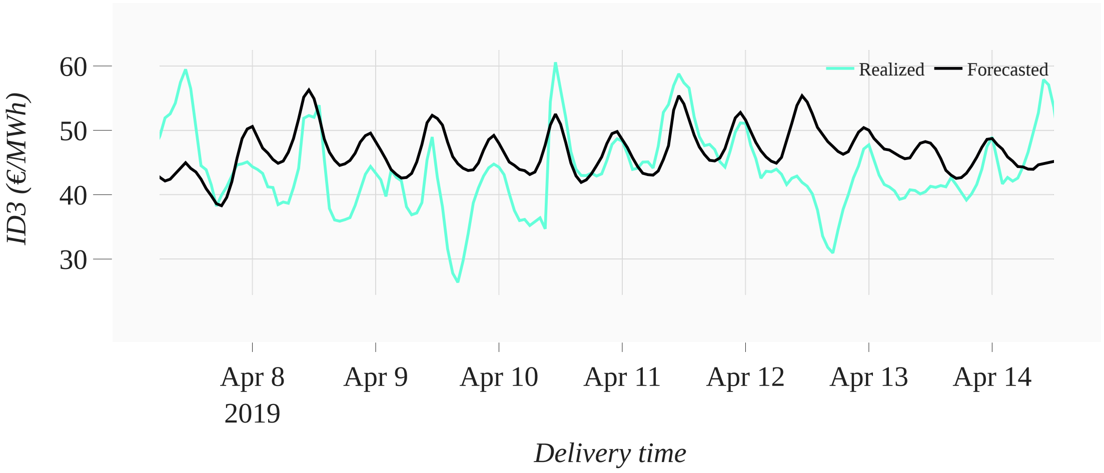


Figure J.6: Visualization of a slice of the forecasted and realized series. *REG.LASSO. Multi-step-ahead point forecasting. Look-ahead-time of 20 hours. Exploitation procedure.*

K Accuracy

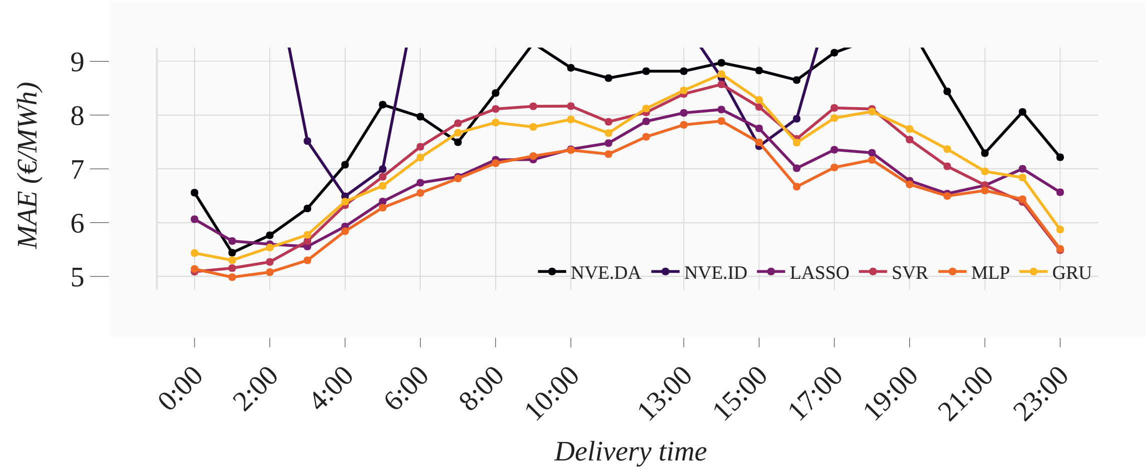


Figure K.1: Accuracy in terms of MAE as function of deliver hour. *Single-step-ahead point forecasting. Exploitation procedure. 2018.*

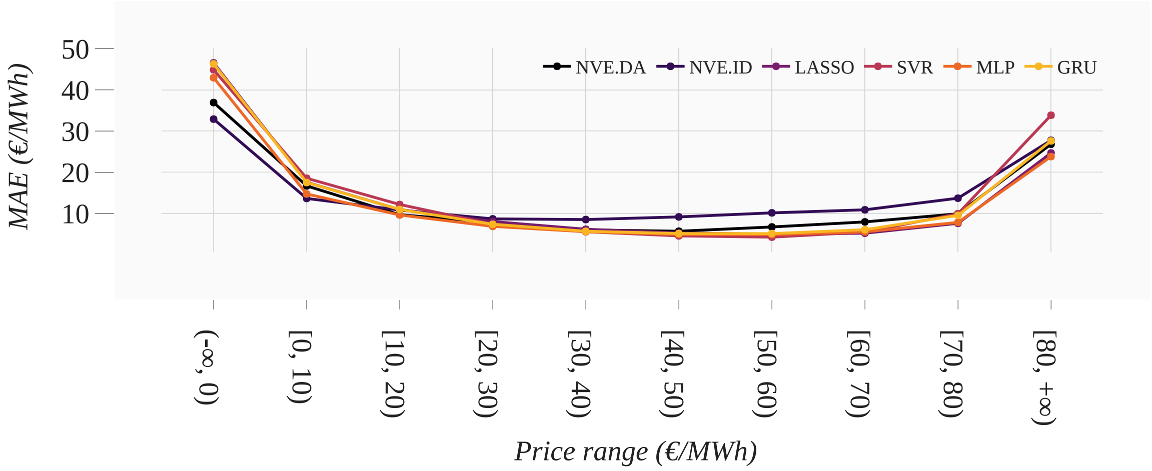


Figure K.2: Accuracy in terms of MAE as function of price range. *Single-step-ahead point forecasting. Exploitation procedure. 2018.*

L Generation profiles

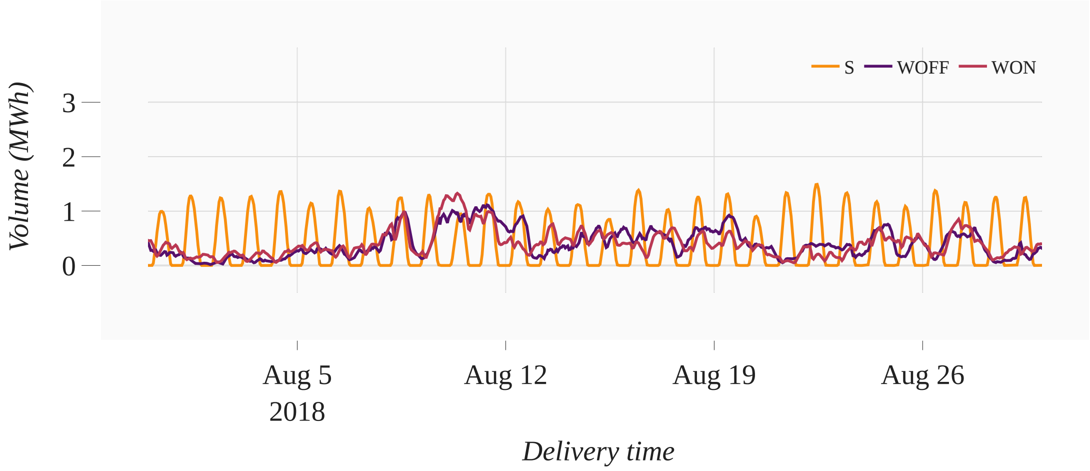


Figure L.1: Averaged and normalized generation profiles of onshore wind, offshore wind, and solar generation. *Summer*.

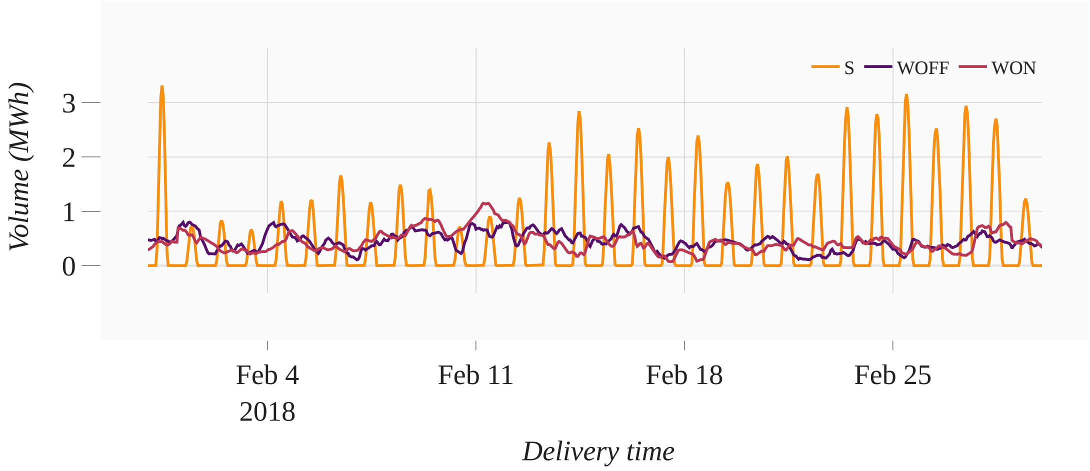


Figure L.2: Averaged and normalized generation profiles of onshore wind, offshore wind, and solar generation. *Winter*.

M Dispatch schedules

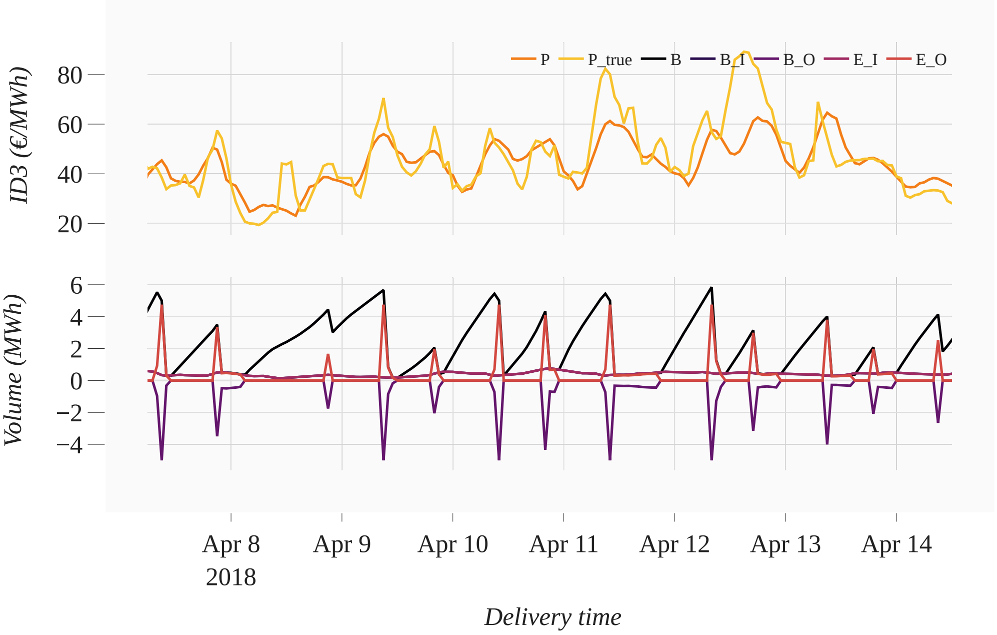


Figure M.1: System variables as function of time. *REG.LASSO. Wind plant.*

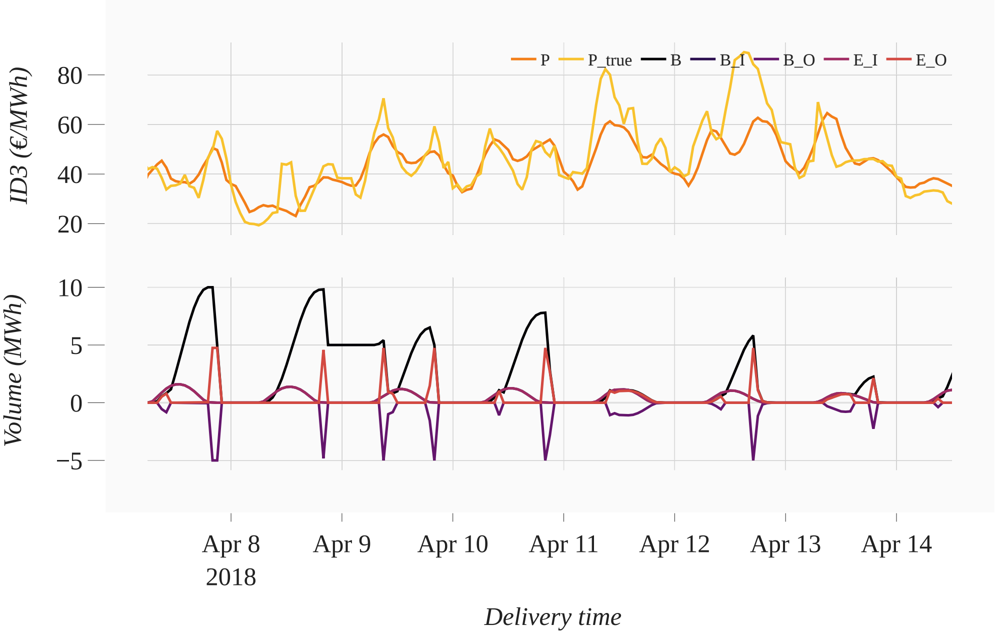


Figure M.2: System variables as function of time. *REG.LASSO. Solar plant.*

N Manuscript

ELSEVIER

Manuscript

Journal 00 (2021) 1–10

Point and interval forecasting of short-term electricity price with machine learning: A theoretical and practical evaluation of benchmark accuracies for the Dutch intraday market

Timo P. Vijn[a]

[a] *Delft University of Technology*

Abstract

This research provides benchmark accuracies for forecasting of an aggregated price of the Dutch intraday market. While point forecasts in a single-step-ahead horizon for that unresearched market provide novel insights already, the scope of this research also includes interval forecasts in a multi-step-ahead horizon. A forecasting procedure is established that organizes several stages of in-sample and out-of-sample testing so that the number of arbitrary choices regarding features and hyperparameters is kept as low as possible. It is concluded on the basis of accuracies attained by naive, regression, and artificial neural network models that the models that are capable to incorporate linear and nonlinear relationships are able to infer to some degree what drives intraday and day-ahead prices, and obtain superior forecasts, especially for the period of 2018 and its many extreme observations. Furthermore, it is addressed whether superiority in terms of accuracy coincides with what is deemed as superior in practice. A simulation of a generic system, which consists of a battery and a wind turbine located in the Netherlands, smartly dispatches stored energy according to a schedule optimized with model predictive control based on point forecasts of intraday price. It is concluded that, in general, higher profits are obtained with more accurate point forecasts and that different point forecasts lead to very different dispatch schedules that vary more than 10% in terms of dispatch frequency.

Electricity price forecasting, Dutch intraday market, Machine learning, Artificial neural networks, Quantiles, Intervals, Model predictive control, Dispatch strategy

1. Introduction

In the Netherlands, the ongoing growth of renewables in the energy mix has its origin around the same time as market liberalization. Initiatives that support the move to a more carbon-neutral energy mix, now heavily incentivized by political agendas and financial stimuli, flow from the the premise that it plays a vital role in mitigating dependence on importing energy, and in reducing the effects of global warming [16]. Generation of variable renewable energy sources, e.g. wind and solar, generally deviates more from generation schedules than that of traditional energy sources. The increased level of uncertainty that imposes to market participants leads to higher needs to balance portfolios just before delivery on the intraday market, which allows for continuous trading throughout the day, and up to 5 minutes before delivery. A steep rise of Dutch intraday trading volumes is the result, with repeated year-on-year growth figures between 30 and 70% during the last five years [8, 9, 10, 11, 12]. The endeavour of forecasting the price of energy contracts has drawn broad academic and commercial interest. Gaining insight into possible future price realisations, and investigating how to improve models and modeling approaches to reduce forecast uncertainties has the potential to improve decision-making of market participants, which in turn might lead to a more stable grid and increased profitability [19]. The price of short-term energy contracts is subject to many complex, nonlinear, and unobservable factors and exhibits challenging characteristics such as high volatility, negative prices, spikes, and jumps, that are generally much more pronounced than in other commodity or stock markets [6].

1.1. Problem description

The numerous models and modeling approaches that have been proposed to push performance beyond that of benchmarks, and the comprehensive frameworks that offer guidelines and best practices [3, 25], demonstrate that electricity price forecasting is an advanced field of research. Nevertheless, the attention for intraday markets is limited, despite their growing relevance. Many of the methodologies and recommendations published in the more extensive body of research on day-ahead markets can be utilized for the intraday market, but differences between the two require novel approaches. For one, the continuous trading of most intraday contracts versus the auction-based trading of day-ahead contracts lead to fundamentally different circumstances with regard to data availability. Inward looking on intraday price forecasting shows that research scopes are often limited to (single-step-ahead) point forecasts, although recent recommendations note that interval forecasts are an obvious avenue for future work [30, 28]. By giving an indication of forecast uncertainty, interval forecasts offer a richer and truer representation of reality. Ideally, superiority is evaluated in situations where forecasts will actually be used [2]. In many cases, that is not or not all a possibility. Therefore, some reluctance to base conclusions upon traditional measures only might be appropriate; these measures might not fully coincide with financial measures that are used in an operational context and thus might fail to address problems that are important to practitioners. This matter has been identified in literature and alternative approaches that consider optimal participation of power plants have been brought forward [33, 27]. Notwithstanding, a lot of research does not acknowledge any theoretical-practical mismatch at all, and utilizes traditional measures only.

1.2. Research objective

In consideration of the limited research and increasing intraday trading volumes on intraday markets in general, there is the opportunity to investigate components that constitute a systematic forecasting procedure. Furthermore, due to the lack of research on the Dutch intraday market, there are opportunities to establish benchmark accuracies. The objective of this research is to evaluate forecasts of an aggregated price for the Dutch intraday market from naive, regression, and ANN models. To push the boundaries of the current body of research, not only point forecasting but also interval forecasting is investigated, in single- and multi-step-ahead horizons. A systematic forecasting procedure ensures that there are a limited number of arbitrary choices and that “what is considered optimal” is reconsidered repeatedly. Evaluation of point forecasts is based on widely-used absolute scores, and on relative (e.g. rMAE) scores that offer a more general indication of accuracy. It is furthermore investigated whether forecasts capture the dynamics of intraday price adequately based on residual error analysis. Evaluation of interval forecasts is based on an aggregated scoring metric for quantile forecasts (CRPS), and the 0.1, 0.5, and 0.9 intervals are assessed in terms of reliability and sharpness. To assess whether theoretical superiority coincides with practical superiority, forecasts are evaluated on the basis of accuracy as well as in an operational context. This research establishes a simulation of a generic system that consists of a battery and a wind turbine located in the Netherlands, that smartly dispatches stored energy according to a schedule optimized with model predictive control and future information on intraday price and generation. It is investigated whether the theoretically superior point forecast leads to a higher profit, or to a different dispatch schedule altogether.

2. Dutch intraday market

In the Netherlands, there is a strong growth of variable renewable energy in consumption and production in an otherwise mostly constant energy mix. That growth is expected to persevere. Uncertain schedules of variable renewable energy sources introduce uncertainty on energy markets, and thus higher volumes are traded on the intraday market than before. Energy trading is thus shifting closer to delivery and the difference between volumes on day-ahead markets and intraday markets is shrinking every year. Dutch intraday price has complex characteristics such as seasonality, spikes, and negative values, which leads to a distribution that has a high degree of skewness and kurtosis.

Given a set $\mathbb{T}_t = {}^{-3.25}_{-3}\mathbb{T}_t(t_d)$, that contains all transactions for an intraday product with time of delivery t_d in the time window from 3 hours to 15 minutes before t_d. The ID3 index represents the volume-weighted average price of all transactions $k \in \mathbb{T}_t$, and is calculated by

$$\mathrm{ID3}_{d,h} = \left(\sum_{k \in \mathbb{T}_t} v_k \right)^{-1} \sum_{k \in \mathbb{T}_t} v_k p_k \qquad (1)$$

where v_k is the volume and p_k is the price of transaction k. This research is concerned with forecasting the price of the ID3 index (*alias* ID3 price), although it is noted that the price of individual transactions may vary considerably [21].

Prices of the more often studied German intraday market have similar statistical properties although Dutch intraday prices are slightly more extreme. The main differences during the period from 2015 through 2020 are

that the German ID3 index has more occurrences of negative price;

that the Dutch ID3 index has more occurrences of extreme prices;

...

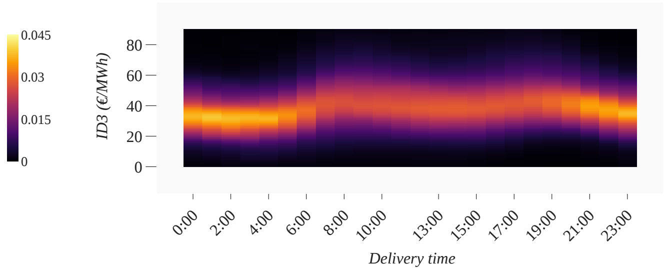


Figure 1: Kernel density estimates of the Dutch ID3 index as function of delivery time. *2015–2020. Data from EEX [13].*

A challenging empirical distribution is the result that is inconsistent from year to year, shown in Figure 2. Particularly 2018 and 2019 are important for later stages of this research, as forecasts are evaluated from an out-of-sample test of that period. While intraday price seems to be rather temperate for 2019, it is far more extreme for 2018. That year saw an increase in electricity prices across Europe that was especially severe for the Netherlands, which is explained mostly due to increases of carbon emission allowances [10]. In 2020, the social distancing rules that were introduced in reaction to the COVID-19 pandemic led to a deeply rooted imbalance between supply and demand due to an abrupt global decline in power consumption [12].

3. Data and feature engineering

An increasing amount of high-resolution data regarding energy markets and environmental conditions is available to practitioners. The European Network of Transmission System Operators for Electricity (ENTSO-E), for instance, sustains the ‘Transparency Platform’, which is part of an ambition to advance the transparent sharing of data with market participants. The platform holds data on the energy market including actual and forecasted load and generation that is hourly or intra-hourly, and

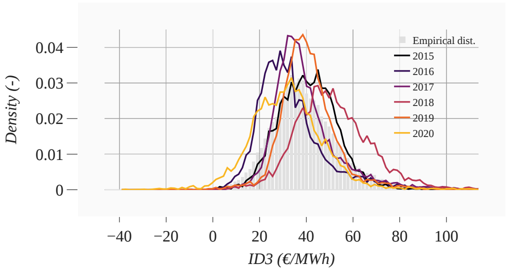


Figure 2: Empirical distributions of the Dutch ID3 index for different years. *Data from EEX [13].*

updated daily. In addition, there are parties that offer data commercially. Among those is EEX, that offers subscription access to price data of all the European energy markets that they (EPEX SPOT) operate. For this research, ENTSO-E and EEX are the main sources of raw data. Table 1 shows details. All data is retrieved for the period from 2015.07 through 2021.07.

Ind.	Type	Resolution	NL	DE	Source	Public
R_{p1}	Transactions ID	Cont.	•		[13]	
R_{p2}	Price ID (ID3)	Hourly		•	[14]	•
R_{p3}	Price DA (MCP)	Hourly	•	•	[13]	
R_{g1}	Generation* actual	Hourly	•	•	[15]	•
R_{g2}	Generation forecast DA	Hourly	•	•	[15]	•
R_{g3}	Generation forecast ID	Hourly		•	[15]	•
R_{l1}	Load actual	Hourly	•	•	[15]	•
R_{l2}	Load forecast DA	Hourly	•	•	[15]	•

Table 1: Raw data

Motivated by literature, various features are included in the full feature set $\subseteq \mathcal{F}_F$. The set, shown in Table 2, contains features that originate from 28 unique time-series and are assigned into one of five categories. Most features are the result of basic computations on the raw data of Table 1. Whereas raw data can come in any resolution and form, a feature must have a single value per pre-defined time-step, which is equal to one hour in this research.

Two concepts that this research employs to establish the full feature set are lags/leads and market integration. Utilizing lags/leads in time-series analysis amounts to creating many features from a single time-series, where instances are shifted forward/backward in time. Market integration in time-series analysis amounts to using features from a different market than the target market. This has been shown to improve forecasting performance when dynamics of the considered markets are related [24, 39]. Many features in the full feature set are lags/leads, and many have both Dutch (NL) and German (DE) variants.

4. Point forecasting

As a representation of future outcomes that are inherently uncertain, single-value (*alias* point) forecasts are always a simplified representation of reality. Nevertheless, point forecasts underlie decision-making in many practices and within many fields, and are not simply put aside by more information-laden representations, whenever they are available [17].

Assumed is a relationship between the target variable $\boldsymbol{y}$ and the explanatory variables $\boldsymbol{X}$, represented by $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$. Approximation of that relationship is a problem addressed in literature, where fundamental models are proposed that, for instance, simulate the market clearing [29]. However, just like the majority of literature, the research is not troubled by naivety in that respect. The intention is rather that given explanatory variables $\boldsymbol{X}$, a forecasted series $\hat{\boldsymbol{y}}$ is calculated that is as close as possible to the corresponding realized series $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{X}) \approx \boldsymbol{y}$.

4.1. Scoring metrics

A scoring function S is a formalization of what an accurate forecast is. It is not so much that for every problem, there exists a scoring function that is clearly more suitable than all others; practitioners should instead be aware of the biases that scoring functions might impose on the solution of a problem, and should motivate choices accordingly. For the field of EPF, where published results and drawn conclusions are often based on absolute scoring functions, [25] suggests that it should become the norm to employ the relative mean absolute error (rMAE) (*alias* mean relative absolute error (MRAE)), which normalizes the MAE score to what a reference forecast $\hat{\boldsymbol{y}}^*$ achieves, i.e.

$$\text{rMAE}(\boldsymbol{y},\ \hat{\boldsymbol{y}},\ \hat{\boldsymbol{y}}^*) := \frac{1}{T}\sum_{i=t}^{T}\frac{|y_t - \hat{y}_t|}{|y_t - \hat{y}_t^*|} \tag{2}$$

When the rMAE is at least included in evaluation, accuracy can be more easily compared across problems. To the best knowledge of the author, intraday price forecasting literature has not (yet) employed the rMAE scoring function. Therefore, the research employs the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) to accompany the rMAE. The MAE and RMSE are used widely and often exclusively in EPF and intraday price forecasting [30, 21, 28]. Like the rMAE, the MAPE normalizes the score, such that evaluation of accuracy can be more easily compared across problems.

4.2. Residuals

The error between the forecasted and realized values (*alias* residuals) $\boldsymbol{\epsilon} = \hat{\boldsymbol{y}} - \boldsymbol{y}$ provides insight in the adequacy of a forecast to capture the characteristics of the target variable. An indication that the vector of forecasted values $\hat{\boldsymbol{y}} = \{\hat{y}_1,\ \ldots,\ \hat{y}_T\}$ captures the information contained in the realized values $\boldsymbol{y}$ adequately, is when the residuals vector $\boldsymbol{\epsilon}$ resembles white noise. It can thus be validated whether $\boldsymbol{\epsilon}$ has an expected value of zero; whether its variance is constant over time; whether its values are uncorrelated in time; and whether its values are normally distributed.

The characteristics of residuals are often overlooked, despite the fact that they can indicate when forecasts are unable to capture the characteristics of the original time series fully, and thus inherit characteristics of the original time series [1]. Analysis of residuals is mostly ignored in EPF, although there are examples where it is utilized to evaluate price forecasts of the day-ahead market [22, 40]. This research provides the results of four concise statistical tests to assess whether the four desired characteristics are attained by point forecasting.

Ind.	Time-series	Lags/Leads	Raw	NL	DE	No.
F_{p1}	Price ID (ID3)	$\{t_d - 168,\ t_d - 48\}$ & $[t_d - 24,\ t_f]$	R_{p1}, R_{p2}	•	•	23
$\cdots$	$\cdots$	$[t_d - 24,\ t_f]$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
F_{p2}	Price ID (ID3) avg. (last week)		R_{p1}, R_{p2}	•	•	1
F_{p3}	Price ID (ID3) avg. $[t_f - 24,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p4}	Price ID (ID3) avg. $[t_f - 168,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p5}	Price ID (ID3) avg. $[t_f - 5040,\ t_f]$		R_{p1}, R_{p2}	•	•	1
F_{p2}	Price DA (MCP)	$\{t_d - 168,\ t_d - 48\}$ & $[t_d - 24,\ t_f + 12]$*	R_{p3}	•	•	36
$\cdots$	$\cdots$	$[t_d - 24,\ t_f + 12]$*	$\cdots$	$\cdots$	$\cdots$	$\cdots$
F_{p7}	Price DA (MCP) avg. (last week)		R_{p3}	•	•	1
F_{p8}	Price DA (MCP) avg. $[t_f - 24,\ t_f - 1]$		R_{p3}	•	•	1
F_{p9}	Price DA (MCP) avg. $[t_f - 168,\ t_f - 1]$		R_{p3}	•	•	1
F_{p10}	Price DA (MCP) avg. $[t_f - 5040,\ t_f - 1]$		R_{p3}	•	•	1
F_{c1}	Hour of day			•		2
F_{c2}	Day of week			•		2
F_{c3}	Month of year			•		2
F_{c4}	Day of year			•		2
F_{g1}	Generation actual	$[t_d - 24,\ t_f - 1]$	R_{g1}	•	•	21∗3
F_{g2}	Generation forecast DA	$[t_d - 24,\ t_f + 5]$	R_{g2}	•	•	37∗3
F_{g3}	Generation forecast ID	$[t_d - 24,\ t_f]$	R_{g3}		•	37∗3
F_{g4}	Generation error DA	$[t_d - 24,\ t_f - 1]$	R_{g1}, R_{g2}	•	•	21∗3
F_{g5}	Generation error ID	$[t_d - 24,\ t_f - 1]$	R_{g1}, R_{g3}		•	21∗3
F_{l1}	Load actual	$[t_d - 24,\ t_f]$	R_{l1}	•	•	21
F_{l2}	Load forecast DA	$[t_d - 24,\ t_f + 12]$	R_{l2}	•	•	37
F_{l3}	Load error DA	$[t_d - 24,\ t_f - 1]$	R_{l1}, R_{l2}	•	•	21

Table 2: Full feature set $\mathcal{F}_F$

4.3. Naive models

In essence, naive models represent a minimum-effort solution, and set the bar that forecasts from more sophisticated models should be able to *at least* outperform.

Day-ahead price The naive model referred to as NVE.DA obtains a forecast of the ID3 price by assuming that it matches the corresponding MCP value. The forecast for delivery on day d and hour h is formally defined as $\hat{y}_t = \mathrm{MCP}_t$. That naive model is also employed in literature [28, 30].

Intraday price A naive approach to obtain a forecasted ID3 value from intraday prices is to simply utilize the most recent ID3 price that is available at the time of forecasting. The naive model referred to as NVE.ID obtains a forecast of the ID3 price by assuming that it matches the volume-weighted average transaction price in the three hours before the time of forecasting. The forecast is formally defined as $\hat{y}_t = {}^{-7}_{-4}\mathrm{ID}_t$.

As most intraday trading is incited by conditions that are unforeseen during day-ahead trading, the day-ahead market clearing price largely defines the intraday ID3 price, which is reflected in high correlations. In the search of forecasts that are able to capture the dynamics of intraday price, forecasts must thus be capable to learn to some extent what sets them apart. The forecast from NVE.DA thus represents the forecast to at least outperform, and this research utilizes the forecasted series of NVE.DA to calculate the rMAE score of other forecasts, i.e. NVE.DA attains an rMAE of 1.00.

4.4. Regression models

Least absolute shrinkage and selection operator (LASSO) regression The second type of considered models base their forecasts on explanatory variables $x_{f,t}$, and estimate how a set of independent variables relates to the target variable of intraday price.

Given a target variable $\boldsymbol{y} = \{y_1,\ \ldots,\ y_T\}$ and explanatory variables $\boldsymbol{X} = \{\boldsymbol{x}_1^\top,\ \ldots,\ \boldsymbol{x}_F^\top\}$ where $\boldsymbol{x}_f = \{x_{f,1},\ \ldots,\ x_{f,T}\}$. A LASSO regression introduces a regularisation term to the cost function of an ordinary least squares regression, i.e.

$$\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = ||\boldsymbol{y} - \boldsymbol{\beta X}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{3}$$

Minimizing the cost function for $\boldsymbol{\beta}$ will find the estimated coefficients, as a function of the regularisation parameter λ, i.e.

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^F}{\arg\min}\left\{\sum_{t=1}^{T}\left(y_t - \sum_{f=1}^{F}\beta_f \cdot x_{f,t}\right)^2 + \lambda\sum_{f=1}^{F}|\beta_f|\right\} \tag{4}$$

Because of the added 'penalty' term, the problem is now penalized on the sum of the magnitudes of the coefficients. Effectively, this constrains the magnitude of the coefficients to a certain range of values, and a larger regularisation parameter λ leads to tighter constraints on the coefficients. The model based on a LASSO regression is referred to as REG.LASSO. That model employs the features in the input feature set $\mathcal{F}_I$ as explanatory variables, i.e. $\boldsymbol{X} \equiv \mathcal{F}_I$, and thus has as many regression coefficients as the number of features in $\mathcal{F}_I$. The forecast is formally defined as

$$\hat{y}_t = \sum_{f=1}^{F}\hat{\beta}_f \cdot x_{f,t} \tag{5}$$

The hyperparameters of REG.LASSO are only the regularization parameter λ.

Support vector regression (SVR) An SVR (*alias* support vector machine regression) utilizes the ϵ-insensitive loss function, which does not tolerate errors that exceed ϵ. A transformation function $\phi(\cdot)$ is employed that transforms data points from the original 'input space' to a higher dimensional 'feature space'. It is attempted to construct a linear model in the feature space, that translates to a non-linear model in the input space. If it is not able to achieve this task within the constraints, the inputs are mapped into an ever higher dimensional feature space. Given a target variable $\boldsymbol{y}$ and explanatory variables $\boldsymbol{X}$, for T observations and F features. The linear function approximation can be denoted as

$$\boldsymbol{y} = f(\boldsymbol{X}) = \sum_{t=1}^{T} w_t \phi_t(\boldsymbol{x}_t) + \gamma = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{X}) + \gamma \tag{6}$$

where $\boldsymbol{\phi}(\cdot)$ is the nonlinear mapping function, $\boldsymbol{w} = \{w_1, \ldots, w_T\}$ is the weight vector, and γ is a bias term. The goal is to find a function, so that for all observations t, the absolute error between the value of the target variable and value of the estimated variable does not exceed ϵ.

This research employs the widely used Gaussian radial basis function kernel

$$K(\boldsymbol{x}_t, \boldsymbol{x}_s) = \exp \frac{-||\boldsymbol{x}_t, \boldsymbol{x}_s||^2}{2\sigma^2} \tag{7}$$

as it can produce complex decision boundaries and adds only a single hyperparameter, the standard deviation σ, to the SVR [26].

The model based on an SVR is referred to as REG.SVR. The hyperparameters of REG.SVR are the regularization parameters C, the width of the ϵ-insensitive tube ϵ, and the standard deviation of the Gaussian kernel σ.

4.5. Artificial neural network models

Feedforward neural networks Multilayer perceptron (MLP) neural networks, that are often regarded as prototypical ANNs, have been demonstrated to perform well in the area of EPF [23]. In contrast to recurrent neural networks, these feedforward neural networks have no cycles. In between input and output layers are one or many hidden layers that contain a given number of identical units. Each unit is connected to all units in the next layer and as such constitute a 'fully connected' network.

The units of an MLP neural network (*alias* neurons) are linear threshold units that take an input vector $\boldsymbol{x} \in \Re^X$ and weight matrix $\boldsymbol{w} \in \Re^{N \times X}$, and utilize the weighted sum and an activation function $\phi(\cdot)$ to calculate their state. For layer $l \in \{0, \ldots, L\}$ the states of all units are given as

$$\boldsymbol{h}^{(l)} = \phi^{(l)}(\boldsymbol{w}^{(l)} \boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}) \tag{8}$$

where the first state vector is the input vector, i.e. $\boldsymbol{h}^{(0)} \Leftrightarrow \boldsymbol{x}$, and the last state vector is the output vector, i.e. $\boldsymbol{h}^{(L)} \Leftrightarrow \hat{\boldsymbol{y}}$.

The model based on an MLP neural network is referred to as ANN.MLP. The hyperparameters of ANN.MLP are the number of hidden layers, the number of neurons in each layer, the activation function, and the dropout rate.

Recurrent neural networks Feedforward neural networks can be limited by the fact that the incorporation of temporal dependencies requires additional features. As the number of input features increases, the size of the network will increase rapidly because of the fully connected nature, which might lead to a considerable increase of computional cost. Recurrent neural networks (RNNs) address that limitation as they are of a chain-like nature. Essential to the successes of RNNs are networks with two particular kinds of gated units, namely the long short-term memory (LSTM) unit and the gated recurrent unit (GRU). These gated units ease optimization and reduce learning degeneracies that more traditional RNNs suffer from.

An LSTM unit employs three gates that control the flow of information. Generally, the activation of a gate is approximated by a linear combination of the previous hidden state $h^{(t-1)}$ and the current input $x^{(t)}$, and a nonlinear activation function $\phi(\cdot)$, i.e.

$$\boldsymbol{g}^{(t)} = \phi(\boldsymbol{W} \cdot [\boldsymbol{h}^{(t-1)}, \ \boldsymbol{x}^{(t)}] + \boldsymbol{b}) \tag{9}$$

A variation and simplification of the LSTM unit is the GRU. Unlike an LSTM unit, a GRU does not have a separate cell state and, thereby, no controlled exposure of its memory content. The hidden state is exposed without any control.

The model based on a GRU network is referred to as ANN.GRU. Similarly to ANN.MLP, the hyperparameters of ANN.GRU are the number of hidden layers, the number of neurons in each layer, the activation function, and the dropout rate.

5. Interval forecasting

On many levels of decision-making, the incorporation of uncertainty into forecasts is invaluable. Some applications require the full probability of the target variable, for instance, while others require estimation of the target variable in the tail of the distribution [36].

Widely used for the bounds of intervals are quantile forecasts. For an interval forecast with a nominal coverage rate of $(1 - \tau)$, the bounds then correspond to the α-quantile forecasts where $\alpha \in \{\tau/2, \ 1 - \tau/2\}$, i.e.

$$\hat{I}_t = \left[\hat{\zeta}_t^{(\tau/2)}, \ \hat{\zeta}_t^{(1-\tau/2)}\right] \tag{10}$$

Given is a series of T realized values $\boldsymbol{y} = \{y_1, \ldots, y_T\}$. In a quantile regression

$$y_t = \zeta_t^{(\alpha)} + \epsilon_t^{(\alpha)} \tag{11}$$

it is assumed that the α-quantile of the residuals is zero, i.e. $P(\epsilon_t^{(\alpha)} \leq 0) = \alpha$, which is unlike in a linear regression, where it is assumed that the mean of residuals is zero, i.e. $E(\epsilon_t) = 0$.

5.1. Reliability and sharpness

The two main concepts that formalize the desires for interval forecasts are reliability (*alias* calibration) and sharpness (*alias* resolution).

Reliability is concerned with the desire that, in essence, realized values should be indistinguishable from random draws of the estimated probability distribution, or else a systematic bias is introduced. Given a series of T forecasted quantile values $\hat{\boldsymbol{y}}^{(\alpha)} = \{\hat{y}_1^{(\alpha)}, \ldots, \hat{y}_T^{(\alpha)}\}$ and a series of realized values $\boldsymbol{y} = \{y_1, \ldots, y_T\}$. It is calculated whether the realized value was smaller than the forecasted quantile value, i.e. whether it was "covered"

Sharpness is concerned with the desire that the more concentrated an interval is, at constant reliability, the better. Intuitively, sharpness is thus concerned with the size of distribution. The size of the interval between the two quantile values with nominal coverage rate $(1 - \alpha)$ can be defined as

$$\delta_t^{(\alpha)} = \hat{\zeta}_t^{(1-\alpha/2)} - \hat{\zeta}_t^{(\alpha/2)} \tag{12}$$

A measure of overall sharpness is the expected value $E(\boldsymbol{\delta}^{(\alpha)})$. It is noted that in real world applications, conditional heteroscedasticity may lead to substantial variability in the width of intervals; simply employing the average width might then be an inadequate characterization of sharpness [18].

In general, reliability is regarded as the most important desire; an interval forecast with high reliability and low sharpness may have wide intervals, it is still more dependable than the unjustly concentrated interval forecast with low reliability and high sharpness. Of two forecasts with similar reliability and similar sharpness, the one with higher resolution is favored.

Just like the scores for point forecasts, there exist scores for quantile forecasts. The widely used quantile scoring function (*alias* pinball loss function)

$$\mathrm{QS}(\hat{\zeta}_t^{(\alpha)}, y_t, \alpha) = \begin{cases} (\hat{\zeta}_t^{(\alpha)} - y_t) \cdot (1-\alpha) & \text{if} \quad y_t \leq \hat{\zeta}_t^{(\alpha)} \\ (y_t - \hat{\zeta}_t^{(\alpha)}) \cdot \alpha & \text{if} \quad \textit{otherwise} \end{cases} \tag{13}$$

is asymmetric in the sense that it penalizes differently when the realized value is smaller or greater than the forecasted quantile value. For quantiles below the middle-quantiles, a higher penalization is for realized values that are smaller than the forecasted quantile, and vice versa for quantiles above the middle-quantile.

Unlike the quantile score, that considers a specific point in the distribution, the continuous ranked probability score (CRPS) considers the distribution as a whole, and requires no predefined classes such as quantiles. The CRPS is a quadratic metric for the difference between a theoretical cumulative distribution and an empirical cumulative distribution, and is thus formally defined for a distribution function $F(x)$

$$\mathrm{CRPS}(F, y_t) = \int_{\Re} (F(x) - \mathbf{1}\{x > y_t\})^2 \, dx \tag{14}$$

which can be equivalently denoted as a scaled integral of the quantile score over all quantiles

$$\mathrm{CRPS}(\hat{\zeta}_t^{(\alpha)}, y_t) = 2 \cdot \int_0^1 \mathrm{QS}(\hat{\zeta}_t^{(\alpha)}, y_t, \alpha) \, d\alpha \tag{15}$$

An approximation of the score can be given if one has quantile forecasts for a finite grid of quantiles $\{\alpha_i, \ \ldots, \alpha_I\}$ that are equidistant [5]. Given forecasted quantile values for a total of H equidistant quantiles $\hat{\zeta}_t^{(\alpha)} = \{\hat{\zeta}_t^{(\alpha_1)}, \ \ldots, \ \hat{\zeta}_t^{(\alpha_H)}\}$

$$\mathrm{CRPS}(\hat{\zeta}_t^{(\alpha)}, y_t) \approx \frac{2}{I} \cdot \sum_{i=1}^{I} \mathrm{QS}(\hat{\zeta}_t^{(\alpha_i)}, y_t, \alpha_i) \tag{16}$$

The CRPS is thus intrinsically an aggregate metric for all target quantiles, while the pinball loss is calculated for each target quantile individually before non-compulsory averaging [20].

5.2. Naive models

In the extension of NVE.DA is a naive model for quantile forecasts, referred to as NVE.Q.DA. The residual errors $\boldsymbol{\epsilon}$ are calculated between the forecasted series $\hat{\boldsymbol{p}}$ and realized series $\boldsymbol{p}$ of the in-sample period, i.e. $\boldsymbol{\epsilon} = |\hat{\boldsymbol{p}} - \boldsymbol{p}|$. From this series of T residual errors, the quantile value is calculated, i.e. $\boldsymbol{Q}^{(\alpha)}(\boldsymbol{\epsilon})$, that is by definition larger than $\alpha \cdot T$ residual error values. Because it is the absolute error, the interval forecast of nominal coverage τ is then obtained by adding (and subtracting) the quantile value of the residual errors to (and from) the out-of-sample forecasted values. The forecast is formally defined as $\hat{q}_t^{(\alpha)} = \hat{p}_t + \mathrm{sign}(\tau - 0.5)\boldsymbol{Q}^{(\tau)}(\boldsymbol{\epsilon})$.

A second naive model, referred to as NVE.Q.PNT employs the exact same approach as NVE.Q.DA, but bases itself on the forecast of the point forecasting model that attains the highest accuracy. This model can thus not be regarded as truly naive when a sophisticated point forecast must be provided. Under the assumption that such point forecasts are available, NVE.Q.PNT can be regarded as a naive interval forecasting model.

5.3. Regression models

Quantile regression appeals because it can describe the relationship between the price and independent variables not only on the mean, but also on the tails of the conditional price. Because a set of coefficients for each quantile is obtained, asymmetric effects of the independent variables can be investigated, which might bring insight on whether features affect the price differently at different price levels. In quantile regression, that is reflected in the cost function, which is minimized as

$$\hat{\boldsymbol{\beta}}^{\tau}(\lambda) = \underset{\boldsymbol{\beta}^{\tau} \in \mathbb{R}^F}{\arg\min} \left\{ \sum_{t=1}^{T} \Big(\tau - [p_t \leq \boldsymbol{\beta}^{\tau}\mathbf{x}_t]\Big)\Big(p_t - \boldsymbol{\beta}^{\tau}\boldsymbol{x}_t\Big) \right\} \tag{17}$$

Contained in the Iverson bracket $[p_t \leq \boldsymbol{\beta}^{\tau}\mathbf{x}_t]$ is the statement whether the true price value is lower than or equal to the estimated price value, i.e. whether there is a positive or negative error. Changing τ changes the ratio of the penalty terms τ and $\tau - 1$, and thus changes how severe over-predictions are penalized compared to under-predictions. It then easily becomes clear that for $\tau = 0.5$, the quantile regression is equivalent to the linear regression, for $|\tau|/|\tau - 1| = 1$. The model based on quantile regression is referred to as REG.Q.LIN.

A second regression model for interval forecasting utilizes quantile regression averaging, as introduced in [31]. Unlike in regular quantile regression, not the features $\boldsymbol{x}_t = [x_t^1, \ldots, x_t^F]$ but forecasts $\boldsymbol{x}_t = [\hat{p}_t^1, \ldots, \hat{p}_t^M]$ are utilized as explanatory variables to arrive at a forecast. Selecting the more valuable point forecasts by eliminating those that are redundant by L^1-norm regularization can improve accuracy [38].

However, this research only utilizes a small number of point forecasts such that regularized quantile regression averaging is not required. Therefore, this research considers only a model based on quantile regression averaging model without regularization, referred to as REG.Q.QRA.

5.4. Artificial neural network models

The model ANN.MLP for point forecasting is employed similarly for interval forecasting. To that end, the same architecture is utilized, but now with an output layer that contains as many nodes as quantiles, i.e. $\boldsymbol{h}^{(L)} = \hat{\boldsymbol{y}} = \{\hat{q}^{(\alpha)}\}$. That network is then trained with the quantile loss function of Equation 13, where each output node receives a unique α_I. The model is referred to as ANN.Q.MLP.

6. Forecasting procedure

This research establishes a forecasting procedure to systematically execute an out-of-sample test with an ‘initialization’ procedure, a ‘calibration’ procedure, and an ‘exploitation’ procedure. For reliable evaluation of accuracies it is vital that the procedure does not interact with observations used in exploitation before exploitation. Hence, this research divides the four-year period from 2016 through 2019 into two two-year periods, so that the period from 2016 through 2017 can be used unhesitatingly for initialization, and the period from 2018 through 2019 can be used for exploitation.

6.1. Initialization

The initialization procedure commences with a stage that reduces the full feature set. A RF regression is fit repeatedly to the initialization period, and after every iteration, the least important features are eliminated. It eliminates 50 features per iteration to come to 500 features, then eliminates 10 features per iteration to come to 100 features, and then eliminates 1 feature per iteration to come to 50 features. The stage of recursive feature elimination is executed in a rolling cross-validation.

This research evaluates all models with initial hyperparameters repeatedly on the initialization period by means of rolling window cross-validation, where every iteration employs a different size for the training window, i.e. $\boldsymbol{N}_{\text{trn}} = \{2\cdot(24\cdot30),\ 6\cdot(24\cdot30),\ 12\cdot(24\cdot30)\}$ (approx. 2 months, 6 months, and 12 months respectively), as well as different sizes of the testing window, i.e. $\boldsymbol{N}_{\text{tst}} = \{24,\ 24\cdot7,\ 1\cdot(24\cdot30)\}$ (approx. 1 day, 1 week, and 1 month).

6.2. Calibration

The calibration procedure commences with a stage that reduces the candidate feature set $\mathcal{F}_C$, which is very similar to the stage of reducing the full feature set $\mathcal{F}_F$ in the initialization procedure. The candidate feature set is reduced from 50 features to 15 features by eliminating 1 feature per iteration, to obtain the input feature set $\mathcal{F}_I \subseteq \mathcal{F}_C$. This stage is performed with $N_{\text{trn}} = 12\cdot(30\cdot24)$ (approx. 12 months) and $N_{\text{tst}} = 1\cdot(30\cdot24)$ (approx. 1 month).

The calibration procedure continues with a stage that finds optimal hyperparameters. A model is fit to the calibration set, and every iteration employs a different set of hyperparameters. This stage is performed with $N_{\text{trn}} = 12\cdot(30\cdot24)$ (approx. 12 months) and $N_{\text{tst}} = 1\cdot(30\cdot24)$ (approx. 1 month).

Grid-search is among the algorithms that search exhaustively and in an unguided way, and is effectively applied in literature [7], although the computational cost can be substantial for a fine-grain search space. Enter informed search algorithms, that transcend the naivety of grid searches or random searches by passing on past results through the optimization procedure. In Bayesian search algorithms, a surrogate function is responsible for updating the prior probability $P(S(\boldsymbol{h}))$ with a sample $\boldsymbol{h}$ and its score $y = S(\boldsymbol{h})$ to get a better posterior probability $P(y|\boldsymbol{h})$. An acquisition function is responsible for guiding the sampling process to where likeliness of finding the optimal solution is highest.

At the core of a tree-structured parzen estimator are two kernel distribution functions $l(\boldsymbol{h})$ and $g(\boldsymbol{h})$, that are often chosen as Gaussian [4], and that are based on the samples that yield a score above or below a pre-defined threshold value y^*, respectively

$$P(\boldsymbol{h}|y) = \begin{cases} l(\boldsymbol{h}) & \text{if} \quad y < y^* \\ g(\boldsymbol{h}) & \text{if} \quad y \geq y^* \end{cases} \tag{18}$$

The acquisition function in a TPE is based on expected improvement

$$EI_{y^*}(\boldsymbol{h}) = \int_{\infty}^{y^*} (y^* - y)P(y|\boldsymbol{h})dy \tag{19}$$

and following the derivations that are explicitly stated in [4], proportionality is found as

$$EI_{y^*}(\boldsymbol{h}) \propto \left(\gamma + \frac{g(\boldsymbol{h})}{\ell(\boldsymbol{h})}(1-\gamma)\right)^{-1} \tag{20}$$

where $\gamma = P(y < y^*)$. Thus, for the expected improvement to grow, the term $g(\boldsymbol{h})/l(\boldsymbol{x})$ must shrink, such that a hyperparameter set $\boldsymbol{h}$ is chosen that has a high probability under $l(\boldsymbol{h})$ and low probability under $g(\boldsymbol{h})$.

6.3. Exploitation

After the initialization procedure and the first calibration procedure, the first exploitation procedure commences. The exploitation procedure is where a model is exploited to obtain a forecast. It employs the input feature set $\mathcal{F}_I$ and the hyperparameter set decided upon in the most recent calibration procedure.

6.4. Multi-step-ahead forecasting

In contrast to the single-model setup of recursive schemes, direct schemes utilize one trained model $\hat{f}_h$ per each lead time, that are separately trained to provide an h-step ahead forecast, i.e.

$$y_t = f_h(y_{t-4},\ \ldots,\ y_{t-4-d}) + w \tag{21}$$

$h \in \{1,\ \ldots,\ H\}$. To obtain a forecast $\hat{y}_{N+H}$, that is for a look-ahead-time of H hours, a direct scheme utilises the model $\hat{f}_H$

$$\hat{y}_{N+H} = \hat{f}_H(y_{N-4},\ \ldots, y_{N-4-d}) \tag{22}$$

Although a direct scheme does not suffer from the propagation of errors, it cannot account for potential interdependencies that may exist between lead times. Besides that, the computational cost that is associated with training numerous separate models greatly exceeds that of recursive schemes. Especially for models that are computational costly, such as the considered ANN models, that might be problematic.

The naive model NVE.DA can be used directly for multi-step-ahead forecasting, albeit for a finite number of steps restricted by the number of published market clearing prices. The series could be extended by forecasts of the market clearing price, but that lies out of the scope of this research. The considered regression and ANN models can be used for look-ahead-times further than 9 hours, however, and this research employs them in a direct scheme for multi-step-ahead forecasting.

Many small results of the forecasting procedure constitute to the bigger result that a more detailed procedure that avoids overfitting leads to slightly higher accuracies. Besides that, it provides insight into the selection of features and hyperparameters, that vary considerably as time progresses. It is demonstrated that the year of 2018, with high volatility and extreme prices, is more challenging than the year of 2019. Overall, the multilayer perceptron ANN for point forecasting obtains the most accurate forecast with rMAEs of 0.81 and 0.77, while the multilayer perceptron ANN for quantile forecasting is outperformed by the model based on quantile regression averaging that attains CRPSs of 3.53 and 2.24. In a multi-step-ahead horizon, results demonstrate that accuracy remains steady until the point that no more market clearing prices are available, after which it rapidly deteriorates.

7. Point forecasting and model predictive control: Energy plant with storage capacity

Only in a practical scenario does the value of a point forecast become apparent; when it translates into actual profit, for instance. In a mostly flat market, sophisticated price forecasts—no matter how accurate—might not lead to more profitable schedules than naive forecasts, while in a volatile market, even mildly accurate forecasts might lead to more profitable schedules. Therefore, it is beneficial to assess not only accuracy but also the impact of a forecast on decision costs and profit [35].

This research establishes a simulation that investigates an energy plant with storage capacity. The system has access to price forecasts of the intraday market and uses model predictive control to settle on a dispatch schedule. The objective of the simulation is not to offer an as-realistic-as-possible outlook on real-world performance. Rather, it is used to see whether and how a system that is a simulation of reality reacts to different price forecasts.

7.1. Base case

The component of the energy plant in the base case is a hypothetical wind plant, because wind energy is the largest variable renewable energy source in terms of volume in the Dutch energy mix. Because high-resolution data of historic generation from actual wind plants in the Netherlands are not publicly available, this research employs the aggregated generation of all wind farms installed on land [15], i.e. all onshore wind plants. Two transformations are performed on that series. Firstly, it is averaged for each hour of 2018 and 2019, so that it becomes a one-year series. Secondly, it is normalized so that for every month there is a total generation of 300 MWh. Because an averaged and normalized one-year generation profile is utilized, the effects of varying generation volumes from month to month and from year to year are eliminated, and the effect of different price forecasts, and that of varying price characteristics of 2018 and 2019, can be assessed. The transformed generation profile has a nominal power of approx. 1.3 MWh, which coincides with the specifications of the relatively small onshore wind turbine SIEMENS SWT-1.3-62, six of which are operational for the wind farm Beabuorren in Friesland, The Netherlands [34].

The component of energy storage in the base case is a large-scale stationary energy storage product in the form of a rechargeable lithium-ion battery. The base case follows the specifications of the TESLA MEGAPACK [37]. The battery is assumed to have a fixed maximum capacity of 10 MWh, and a fixed minimum capacity of 0 MWh, fixed maximum charge and discharge rates of 5 MWh per hour, and fixed charging and discharging efficiencies of 95%, i.e. a fixed round-trip efficiency of approx. 90%. There is also a cost of usage, that incorporates that the lifetime of the battery is affected when repeatedly charging and discharging the battery, that is assumed to have a fixed value of 3 €/MWh. The subsystem of energy storage is thus reduced to a capacity range, charging and discharging rates and efficiencies, and a cost of usage [32].

7.2. Optimization problem

The objective of the system is to maximize an objective variable, the estimated profit $\hat{Q}$, by controlling the amount of energy supplied to the battery $B_I(k)$ and drawn from the battery $B_O(k)$. At a sampling instant k, $\hat{Q}(k)$ is defined as the amount of energy $E_O(k)$ times the estimated ID3 price $\hat{P}(k)$, minus the costs of the battery, i.e. $C_I \cdot B_I(k) + C_O \cdot B_O(k)$. This research refers to $Q(k)$ and $\hat{Q}(k)$ as profit and estimated profit, although a more delicate approach that takes into account trading activity would be necessary for them to be more representative of real world profits.

Model predictive control (*alias* receding horizon control) is a control scheme that optimizes for a finite number of future states in the prediction horizon, then executes the first step of the found solution, and then re-optimizes on a horizon that is shifted one step into the future. That makes it particularly suitable for the problem at hand, as updated forecasts and system states can be provided continuously. Model predictive control utilizes an objective function and constraints that together formalize the problem for the prediction horizon K. The optimization must satisfy all equality and inequality constraints.

$$\begin{aligned}
\min_{B_I,\, B_O} \quad & \hat{Q} = \sum_{k=1}^{K} -\Big(E_O(k) \cdot \hat{P}(k) - (C_I \cdot B_I(k) + C_O \cdot B_O(k))\Big) \\
\text{s.t.} \quad & B(k+1) = B(k) + B_I(k) \cdot \eta_I + B_O(k) \\
& E_O(k) = E_I(k) - B_I(k) - B_O(k) \cdot \eta_O \\
& B_I(k) = 100\% \cdot E_I(k) \\
& 0 \le B(k) \le 10 \\
& -B_O(k) \le B(k) \\
& 0 \le B_I(k) \le 5 \\
& -5 \le B_O(k) \le 0
\end{aligned} \tag{23}$$

All energy that is generated enters the battery before it is dispatched, i.e. $B_I(k) = 100\% \cdot E_I(k)$. Without that constraint, the system would utilize the battery only when deeming that it increases estimated profit. Given the considerable cost of usage and inefficiencies of the battery, it would then be utilized only a couple of times per day, and the dispatch frequency would thus increases significantly. The resulting schedule might be subject to a high degree of unaccepted offers in reality. Furthermore, there are trading constraints in place on the intraday market that such a schedule might not conform to. Another reason to constrain $B_I(k) = 100\% \cdot E_I(k)$ is that it eliminates the effects of the cost of usage and of inefficiencies of the battery on the schedule. These battery specifications now only affect the absolute profit.

It is assumed that all offers that are hypothetically placed for dispatch are accepted for the realized ID3 price of the product in question, although in reality that would require matching counteroffers. The 4-hour gap between forecasting and delivery provides a broad window wherein offers of the ID3 price can be be accepted. Also, the system never dispatches more than 95% · 2.5 MWh. Offers of such volumes are very common on the Dutch intraday market.

When all realized system outputs, i.e. $B_I(t)$ and $B_O(t)$ for $t = \{1, \ldots, T\}$, are determined from the optimization problem of Equation 23, the profit Q can be calculated, i.e.

$$Q = \sum_{t=1}^{T} E_O(t) \cdot P(t) - \Big(C_I \cdot B_I(t) + C_O \cdot B_O(t)\Big) \tag{24}$$

The schedules are evaluated on the profit Q and on rel. profit, i.e. the profit normalized by that attained by a reference schedule. Furthermore, the schedule is evaluated on the percentage of hours that the system dispatches (*alias* dispatch frequency), and the average volume of dispatch (*alias* dispatch volume). The dispatch frequency and volume are linearly related because all generated volume is dispatched. They thus shine a different light on the same result.

Two reference schedules are added to the evaluation. The first reference schedule, referred to as *Direct feed*, involves no optimization or forecasting. That schedule simply dispatches all energy generated during an hour for the ID3 price of that hour, such that $E_O(t)$ simply equals $E_I(t)$ and the profit is $\sum_{t=1}^{T} E_I(t)\cdot P(t)$. The second reference schedule, referred to as *True price*, is optimized based on the true intraday price. The system thus has full knowledge of price and generation and thus provides an upper limit for schedules that are optimized based on price forecasts. Direct feed is utilized to normalize the profits of other schedules, i.e. the rel. profit of *Direct feed* is 100%.

It is concluded that the system attains higher profits when exploiting a schedule that is based on a more accurate forecast. With a schedule based on the most accurate point forecast, that from ANN.MLP, the system attains a profit that is 94.5% of the reference profit that is based on a practically infeasible direct feed strategy. Even based on the least accurate forecast, that from NVE.DA, the system is able to attain 93.1% of the reference profit. It is concluded that it is not so much the accuracy of the forecast than its shape, that is important to attain profit. In terms of dispatch frequency and volume, the schedules are more divergent, however, and especially for a smaller storage capacity, the schedules based on forecasts from REG.LASSO and ANN.MLP become relatively high frequency and low volume schedules, with more than 8% higher dispatch frequencies than the schedule based on forecasts from NVE.DA. It is concluded that the considerable reduction of point forecast accuracies for further look-ahead-times does not mislead the system and deteriorate profits, as it amasses approx. 3% more profit with schedules based on forecasts for further look-ahead-times.

8. Conclusion

The considered artificial neural network model based on a multilayer perceptron, which is capable of incorporating nonlinear relationships, provides a superior point forecast in the out-of-sample test. A similar architecture that is trained on the basis of several quantiles does not outperform a quantile regression averaging of the considered point forecasts, however. Relative and absolute accuracies of all forecasts are shown to vary significantly with the delivery hour and with the price regime. The year of 2018, with attained rMAEs and CRPSs as low as 0.81 and 3.53, is considerably more challenging than the year of 2019, with attained rMAEs and CRPSs as low as 0.77 and 2.24, due to high volatility and many extreme prices. Results of the simulation demonstrate that more accurate forecasts lead to slightly higher profits, although much of the essential information is captured by all forecasts. The schedule itself is more sensitive to the price forecast, however, and dispatch frequency and volume deviate more than 8% when based on different forecasts. Practitioners might take that into account and not gravitate towards point forecasts merely on the basis of accuracy.

References

[1] Muhammad Ardalani-Farsa and Saeed Zolfaghari. "Residual analysis and combination of embedding theorem and artificial intelligence in chaotic time series forecasting". In: *Applied Artificial Intelligence* 25.1 (Jan. 2011), pp. 45–73. DOI: 10.1080/08839514.2011.529263.

[2] J. Scott Armstrong. "Evaluating forecasting methods". In: *International Series in Operations Research & Management Science*. Springer US, 2001, pp. 443–472. URL: http://dx.doi.org/10.1007/978-0-306-47630-3_20.

[3] J. Scott Armstrong and Kesten C. Green. "Forecasting methods and principles: Evidence-based checklists". In: *Journal of Global Scholars of Marketing Science* 28.2 (Mar. 2018), pp. 103–159. DOI: 10.1080/21639159.2018.1441735.

[4] James Bergstra et al. "Algorithms for hyper-parameter optimization". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor J. Shawe-Taylor. Vol. 24. Curran Associates, Inc., 2011, pp. 2546–2554.

[5] Jonathan Berrisch and Florian Ziel. "CRPS Learning". In: 2021. URL: https://arxiv.org/abs/2102.00968.

[6] Álvaro Cartea and Marcelo G. Figueroa. "Pricing in electricity markets: A mean reverting jump diffusion model with seasonality". In: *Applied Mathematical Finance* 12.4 (Dec. 2005), pp. 313–335. DOI: 10.1080/13504860500117503

[7] Sumeyra Demir et al. "Introducing technical indicators to electricity price forecasting: A feature engineering study for linear, ensemble, and deep machine learning models". In: *Applied Sciences* 10.1 (Dec. 2019), p. 255. DOI: 10.3390/app10010255.

[8] EEX. *Annual Report 2016*. 2017.

[9] EEX. *Annual Report 2017*. 2018.

[10] EEX. *Annual Report 2018*. 2019.

[11] EEX. *Annual Report 2019*. 2020.

[12] EEX. *Annual Report 2020*. 2021.

[13] EEX. *Data of EPEX SPOT markets*. Webshop EEX Group. URL: https://webshop.eex-group.com/epex-spot-public-market-data.

[14] Energy Charts. *Data of German energy price*. Energy Charts Dashboard. URL: https://energy-charts.info/index.html?l=en&c=DE.

[15] ENTSO-E. *Data of European energy price, generation, and load*. ENTSO-E Transparency Platform. URL: https://transparency.entsoe.eu/.

[16] Angelica Gianfreda, Francesco Ravazzolo, and Luca Rossini. "Comparing the forecasting performances of linear models for electricity prices with high RES penetration". In: *International Journal of Forecasting* 36.3 (July 2020), pp. 974–986. DOI: 10.1016/j.ijforecast.2019.11.002.

[17] Tilmann Gneiting. "Making and evaluating point forecasts". In: *Journal of the American Statistical Association* 106.494 (June 2011), pp. 746–762. DOI: 10.1198/jasa.2011.r10138.

[18] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (Apr. 2007), pp. 243–268. DOI: 10.1111/j.1467-9868.2007.00587.x.

[19] Shadi Goodarzi, H. Niles Perera, and Derek Bunn. "The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices". In: *Energy Policy* 134 (Nov. 2019), p. 110827. DOI: 10.1016/j.enpol.2019.06.035.

[20] Tao Hong and Shu Fan. "Probabilistic electric load forecasting: A tutorial review". In: *International Journal of Forecasting* 32.3 (July 2016), pp. 914–938. DOI: 10.1016/j.ijforecast.2015.11.011.

[21] Tim Janke and Florian Steinke. "Forecasting the price distribution of continuous intraday electricity trading". In: *Energies* 12.22 (Nov. 2019), p. 4262. DOI: 10.3390/en12224262.

[22] Orhan Karabiber and George Xydis. "Electricity price forecasting in the danish day-ahead market using the TBATS, ANN and ARIMA methods". In: *Energies* 12.5 (Mar. 2019), p. 928. DOI: 10.3390/en12050928.

[23] Jesus Lago, Fjo De Ridder, and Bart De Schutter. "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms". In: *Applied Energy* 221 (July 2018), pp. 386–405. DOI: 10.1016/j.apenergy.2018.02.069.

[24] Jesus Lago et al. "Forecasting day-ahead electricity prices in Europe: The importance of considering market integration". In: *Applied Energy* 211 (Feb. 2018), pp. 890–903. DOI: 10.1016/j.apenergy.2017.11.098.

[25] Jesus Lago et al. "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark". In: *Applied Energy* 293 (July 2021), p. 116983. DOI: 10.1016/j.apenergy.2021.116983.

[26] R. Laref et al. "On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications". In: *Chemometrics and Intelligent Laboratory Systems* 184 (Jan. 2019), pp. 22–27. DOI: 10.1016/j.chemolab.2018.11.011.

[27] Katarzyna Maciejowska, Weronika Nitka, and Tomasz Weron. "Day-Ahead vs. intraday—forecasting the price spread to maximize economic benefits". In: *Energies* 12.4 (Feb. 2019), p. 631. DOI: 10.3390/en12040631.

[28] Grzegorz Marcjasz, Bartosz Uniejewski, and Rafał Weron. "Beating the naïve—combining LASSO with naïve intraday electricity price forecasts". In: *Energies* 13.7 (Apr. 2020), p. 1667. DOI: 10.3390/en13071667.

[29] Rodrigo de Marcos, Antonio Bello, and Javier Reneses. "Short-Term electricity price forecasting with a composite fundamental-econometric hybrid methodology". In: *Energies* 12.6 (Mar. 2019), p. 1067. DOI: 10.3390/en12061067.

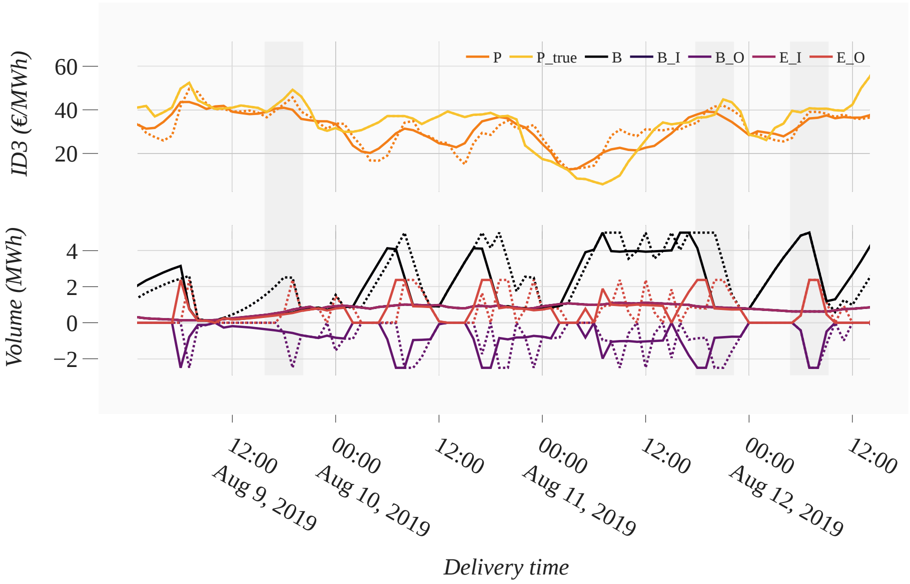


Figure 3: System variables as function of time. *NVE.DA (line) & REG.LASSO (dotted). Wind plant. Storage capacity of 5 MWh.*

[30] Michał Narajewski and Florian Ziel. “Econometric modelling and forecasting of intraday electricity prices”. In: *Journal of Commodity Markets* 19 (Sept. 2020), p. 100107. DOI: 10.1016/j.jcomm.2019.100107.

[31] Jakub Nowotarski and Rafał Weron. “Computing electricity spot price prediction intervals using quantile regression and forecast averaging”. In: *Computational Statistics* 30.3 (Aug. 2014), pp. 791–803. DOI: 10.1007/s00180-014-0523-0.

[32] Amparo Núñez-Reyes et al. “Optimal scheduling of grid-connected PV plants with energy storage for integration in the electricity market”. In: *Solar Energy* 144 (Mar. 2017), pp. 502–516. DOI: 10.1016/j.solener.2016.12.034.

[33] Pierre Pinson, Christophe Chevallier, and George N. Kariniotakis. “Trading wind generation from short-term probabilistic forecasts of wind power”. In: *IEEE Transactions on Power Systems* 22.3 (Aug. 2007), pp. 1148–1156. DOI: 10.1109/tpwrs.2007.901117.

[34] The Wind Power. *Data of wind farm Beabuorren in Friesland, The Netherlands*. Wind farm database. URL: https://www.thewindpower.net/windfarm_en_6264_beabuorren.php.

[35] Akylas Stratigakos, Andrea Michiorri, and Georges Kariniotakis. “A value-oriented price forecasting approach to optimize trading of renewable generation”. In: *2021 IEEE Madrid PowerTech*. IEEE, June 2021. URL: http://dx.doi.org/10.1109/powertech46648.2021.9494832.

[36] James W. Taylor. “Evaluating quantile-bounded and expectile-bounded interval forecasts”. In: *International Journal of Forecasting* 37.2 (Apr. 2021), pp. 800–811. DOI: 10.1016/j.ijforecast.2020.09.007.

[37] Tesla. *Specifications of Tesla Megapack*. Tesla Megapack. URL: https://www.tesla.com/megapack.

[38] Bartosz Uniejewski and Rafał Weron. “Regularized quantile regression averaging for probabilistic electricity price forecasting”. In: *Energy Economics* 95 (Mar. 2021), p. 105121. DOI: 10.1016/j.eneco.2021.105121.

[39] Lennard Visser, Tarek AlSkaif, and Wilfried van Sark. “The importance of predictor variables and feature selection in day-ahead electricity price forecasting”. In: *2020 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, Sept. 2020. URL: http://dx.doi.org/10.1109/sest48500.2020.9203273.

[40] Lei Wu and Mohammad Shahidehpour. “A hybrid model for day-ahead price forecasting”. In: *IEEE Transactions on Power Systems* 25.3 (Aug. 2010), pp. 1519–1530. DOI: 10.1109/tpwrs.2009.2039948.

Intentionally blank