Energy Economics 119 (2023) 106521

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneeco

Foreseeing the worst: Forecasting electricity DART spikes[☆]

Rémi Galarneau-Vincent[a], Geneviève Gauthier[b,c], Frédéric Godin[d,e,*]

[a] *HEC Montréal, Department of Decision Sciences, Montreal, Canada*
[b] *HEC Montréal, Department of Decision Sciences and GERAD, Montreal, Canada*
[c] *Oxford-Man Institute of Quantitative Finance, Oxford, UK*
[d] *Concordia University, Department of Mathematics and Statistics, Montreal, Canada*
[e] *Quantact Laboratory, Centre de Recherches Mathématiques, Montreal, Canada*

ARTICLE INFO

JEL classification:
C53
L94
N72

Keywords:
Power markets
Spikes prediction
DART spreads
NYISO
Predictive analytics
Statistical learning

ABSTRACT

Statistical learning models are proposed for the prediction of the probability of a spike in the electricity DART (day-ahead minus real-time price) spread. Assessing the likelihood of DART spikes is of paramount importance for virtual bidders, among others. The model's performance is evaluated on historical data for the Long Island zone of the New York Independent System Operator (NYISO). A tailored feature set encompassing novel engineered features is designed. Such a set of features makes it possible to achieve excellent predictive performance and discriminatory power. Results are shown to be robust to the choice of the predictive algorithm. Lastly, the benefits of forecasting the spikes are illustrated through a trading exercise, confirming that trading strategies employing the model predicted probabilities as a signal generate consistent profits.

1. Introduction

Electricity generation and consumption must happen simultaneously, which makes the electricity market particularly volatile. The slow reaction time of large producers of inexpensive electricity, combined with bottlenecks and failures in the transmission grid, or sudden increases (decreases) in demand, give rise to a phenomenon known as price spikes, where electricity trades at extremely high (or low negative) prices.

The New York Independent System Operator (NYISO) administers electricity flow operations for a large area in New York State. Electricity transactions are performed through two main markets: the day-ahead (DA) market and the real-time (RT) market. The DA market allows for the scheduling of power production and consumption one day in advance and contains the bulk of traded electricity volumes. Conversely, the RT market acts as a balancing market, correcting for the real-time departure of electricity volumes previously booked in the DA market. The NYISO is responsible for calculating DA and RT prices, which are decided through an auction system matching supply and demand while preserving the integrity of the power system. The DA market closes at 5:00 the day before the generation and distribution of electricity take place, and DA prices are published by NYISO at 11:00 on the same day. Market participants, thus, learn about DA prices only after the DA market closes. On the DA market, the scheduling of electricity is performed on an hourly basis, which allows participants to submit different bids for each hour. Conversely, the RT prices are updated every five minutes, and the hourly RT prices are obtained by aggregating all 5-minutes RT prices published during the hour of interest.

On each transmission grid node, hourly DA and RT prices are determined through a locational-based marginal pricing (LBMP) approach, reflecting the marginal cost of consumption of an additional MWh of electricity on the node for that hour. When reported by the NYISO, the LBMP is further decomposed into three components: (1) energy cost, (2) congestion cost, and (3) losses. Congestion costs occur when the grid's electrical transmission capacity is exceeded under the most economical dispatching scenario. The NYISO is then constrained to dispatch more-expensive power generation units from local power plants, leading to substantial price increases for the associated nodes.

[☆] Galarneau-Vincent is supported by the TMX Group, Hydro-Québec, Letko-Brosseau, Manulife, IVADO, and MITACS. Gauthier is supported by NSERC (RGPIN-2019-04029), and by a professorship funded by HEC Montréal and the HEC Montréal Foundation. Godin is supported by NSERC (RGPIN-2017-06837). We also wish to thank Plant-E Corp (https://www.plantecorp.com/) and its team, in particular Pierre Plante, Vincent Marchal, Shahab Einabadi and Maedeh Mehranfar, for providing access to their data and for their deep insight into the power markets which was decisive in the production of this paper.

[*] Corresponding author at: Concordia University, Department of Mathematics and Statistics, Montreal, Canada.
E-mail addresses: remi.galarneau-vincent@hec.ca (R. Galarneau-Vincent), genevieve.gauthier@hec.ca (G. Gauthier), frederic.godin@concordia.ca (F. Godin).

https://doi.org/10.1016/j.eneco.2023.106521
Received 20 June 2022; Received in revised form 5 January 2023; Accepted 9 January 2023
Available online 20 January 2023

The present study is concerned with a quantity referred to as the *DART spread*, which is the difference between the DA and RT prices of power for a given grid node and hour. Since DA prices encompass market participants' expectations about the next-day RT prices, the DART spread could loosely be thought of as the market's price forecast error, up to a risk premium typically embedded in DA prices (Longstaff and Wang, 2004).

A thorough understanding of DART spread dynamics is essential for several market participants. For instance, virtual bidders who do not possess production or supply capacity and who must therefore reverse DA commitments in the RT markets are exposed to DART spreads instead of standalone prices from the DA or RT markets. A long position on the DA market puts the virtual bidders at risk when the DART is negative. DART spread dynamics also have implications for production facility and retailer risk managers who must decide on the volumes to be locked in ahead of time on the DA market to optimize risk-reward trade-offs faced by their institution.

Spike events are strong sources of risk for electricity market participants. Most of the literature is concerned with spikes in the electricity prices because generators, retailers, and large electricity consumers are exposed to extreme price levels. This has led several authors to explore price spike forecasting, e.g., Christensen et al. (2009, 2012), Eichler et al. (2014), and Sandhu et al. (2016).

However, the present study instead considers the perspective of virtual bidders who are concerned with spikes in DART spreads rather than in prices. Therefore, this study considers the problem of forecasting these extreme DART events, or more precisely, the probability that a DART spread spike occurs in a given hour based on available information. Such a problem is expressed as a supervised learning problem which is tackled with four machine learning algorithms: (1) logistic regression, (2) random forests, (3) gradient boosting trees, and (4) deep neural networks (DNN).

To illustrate the developed approach, this study focuses on the Long Island zone as it is well known for being susceptible to DART spikes. This phenomenon results from Long Island's geographic location – it is a peninsula – which entails a smaller capacity to carry electricity from inexpensive power plants situated outside of the zone. This reduced grid capacity creates frequent bottlenecks, thereby raising the congestion price component for Long Island.

From an economic standpoint, the added value of the model-predicted spike probabilities is assessed through trading backtests involving trading strategies that rely on these as signals. Such strategies are compared to a base-case strategy that systematically holds long positions on the DART spread. Such an approach seeks to collect the DART premium (the average positive DART spread), which rewards investors for exposing themselves to negative DART spread spikes caused by high RT prices. However, the flip side of this strategy is constant exposure to sudden significant losses stemming from such spikes. The signal generated by the predictive models makes it possible to modify the base-case strategy to develop novel strategies that avoid long positions when the likelihood of a price spike is too high. Such strategies are shown to lead to significantly better profitability and lower risk than the base-case strategy, outlining the contribution of the spike probability signals generated by the models. Results show that trading strategy performance is robust to the choice of the predictive model.

In summary, this paper offers two main contributions. The first is the comparison of multiple statistical and machine learning models producing predictions of DART spread spike occurrence. Since the bulk of the literature is concerned with price spikes, considering DART spreads instead of prices is a key differentiating feature of our study. The second contribution consists in showcasing the usefulness of the informational content embedded in spike probability forecasts by integrating such signals into trading strategies, which improves trading performance.

The paper is subdivided as follows. Section 2 provides a review of the raw data, describes the spike labeling methodology, and discusses the engineered feature set that is considered. In Section 3, four predictive algorithms are trained on the data, and their performance is assessed. Furthermore, individual features' contribution to predictive performance is assessed. Section 4 proposes simple trading strategies integrating the model-generated signals to determine investment positions, with their profitability and risk assessed during the conduction of an out-of-sample backtest. Section 5 concludes.

2. Data description

This section discusses the raw data and their transformation for subsequent predictive analysis with supervised learning algorithms. In particular, the construction of labels and features for each observation is outlined.

2.1. Raw data

This research project focuses on the Long Island zone overseen by the NYISO. The NYISO provides historical data on day-ahead and real-time electricity prices and loads for its various zones, including Long Island, with hourly granularity, as well as the grid transfer capacity of the multiple interfaces supplying Long Island.[1] The electricity price (LBMP) and load data considered extend from January 1, 2015, to October 31, 2021. The DART spread for hour t is calculated by subtracting the hour-t real-time price RT_t from the corresponding day-ahead price DA_t, that is

$$\mathrm{DART}_t = \mathrm{DA}_t - \mathrm{RT}_t.$$

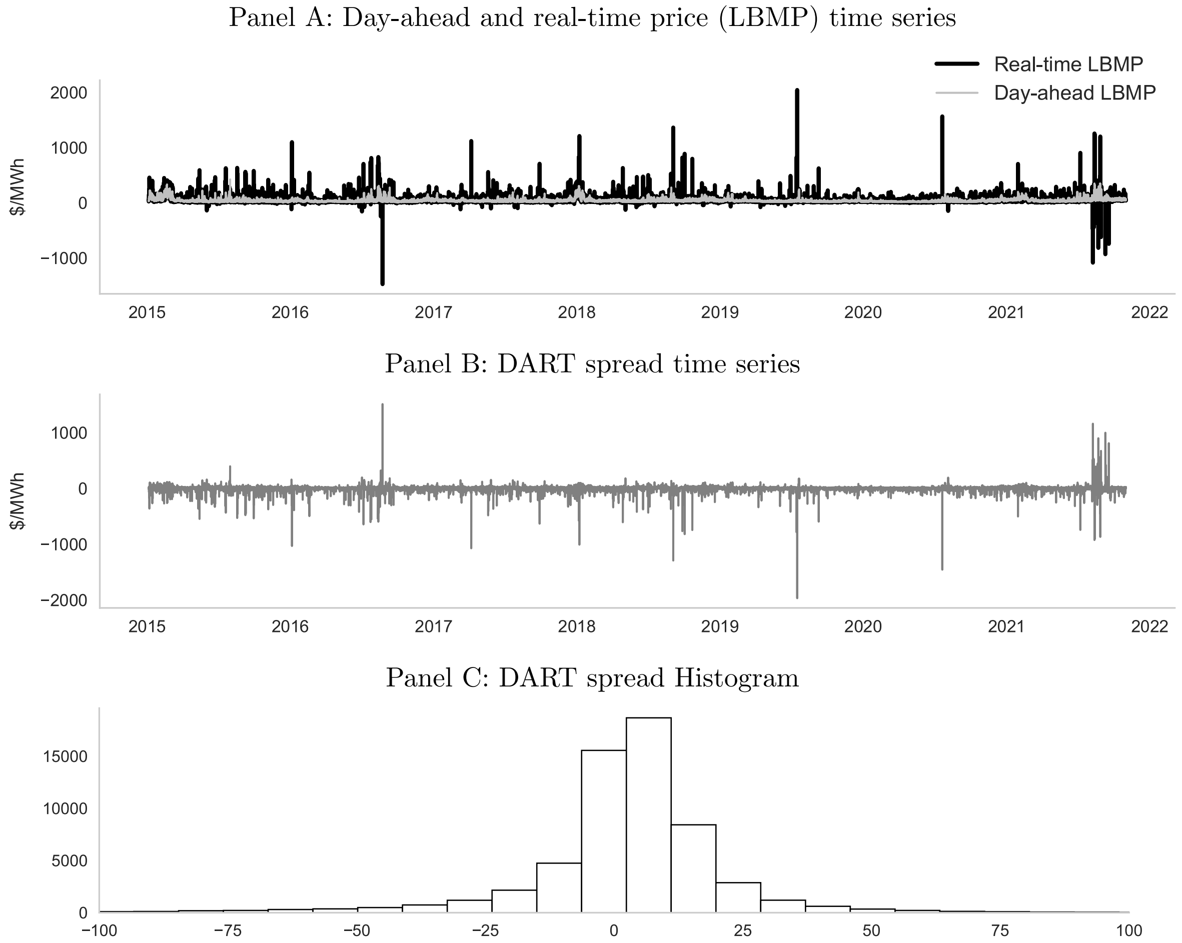
Time series of the day-ahead and real-time prices are reported in Panel A of Fig. 1, whereas Panel B provides the associated DART spread time series. Panel C exhibits the historical distribution of DART spreads through a histogram. As expected, real-time prices are much more volatile than day-ahead prices. The sample average of the DART spread is \$0.45/MWh, implying a positive DART premium compensating for aversion to spike risk.

Various weather-related variables are obtained from the data provider *Openweathermap*. First, hourly realized temperature (in degrees Celcius) is collected for Long Island between January 1, 2015, and October 31, 2021.

Second, temperature forecast data are included, which consist of daily weather forecasts as of 18:00 with horizons ranging from 30 to 54 h in three-hour increments. More details about temperature forecasts are provided in Appendix A.

Historical temperature forecast data are only available as of October 7, 2017. Thus, to complete the sample and make up for missing temperature forecasts between January 1, 2015, and October 6, 2017, synthetic forecasts are generated by adding statistical noise on realized temperature data. Appendix A.2 explains this procedure in detail. Throughout the study, synthetic temperature forecasts are never included in test samples used for performance assessment, but only in training sets. This prevents spurious performance assessment due to information leakage from training to test sets where information related to realized temperature, rather than genuine forecasts, would unduly be provided to the predictive model.

[1] The Long Island zone's grid transfer capacity consists of the total transfer capacity from the Con ED-LIPA, NPX-1385, NPX-CSC, SprainBrooke-Dunwoodie South lines Y50 and Y49, and PJM-NEPTUNE interfaces as described in NYISO (2016). In this work, the grid transfer capacity considered is the sum of reported capacities for all such interfaces.



Panel A displays the hourly real-time and day-ahead price (LBMP) time series for the Long Island zone of the NYISO from January 1, 2015, to February 15, 2021. Panel B displays the corresponding DART spreads (difference between the day-ahead and real-time prices) for the same dates. Panel C displays a histogram characterizing the DART spread distribution on the same period.

Fig. 1. Historical data for the real-time price, day-ahead price and DART spread.

2.2. *Identifying electricity price spikes*

This study aims to perform a daily computation of probabilities of a DART spread spike occurring in each hour of a given day. The task is approached as a supervised learning exercise. The response variable has a binary format: either “1” for hours in which a spike occurs or “0” otherwise. Such labels are not readily observable, and the first step is to determine which observations are considered to be spikes.

2.2.1. *Spike identification criteria*

While there is consensus in the literature on the notion that price spikes are extreme price events, no single, objective definition of “price spike” has emerged. Weron (2007) and Janczura et al. (2013) highlight this lack of consensus in the literature and explain that such a definition is a subjective matter. Notwithstanding disparities among the various possible definitions, many authors such as Sandhu et al. (2016) or Janczura et al. (2013) characterize price spikes as extreme high prices that exceed a certain threshold and are short-lived.

Several approaches to identify electricity price spikes have been considered in the literature, among which the following three have proven quite popular:

- *Fixed price threshold*: Occurrences exceeding some selected fixed price threshold are classified as spikes. See for instance Klüppelberg et al. (2010), Amjady and Keynia (2010), Christensen et al. (2012), Herrera and González (2014), Eichler et al. (2014), Clements et al. (2015) and Manner et al. (2016), He and Chen (2016).
- *Variable price threshold*: Occurrences exceeding a given sample quantile of observed values are flagged as spikes. For example, the highest (lowest) 5% of sample prices are considered spikes. A non-exhaustive list of works applying this criterion includes Trueck et al. (2007) and Sandhu et al. (2016).
- *Statistical filtering of spikes*: A stochastic process capturing prices dynamics is selected and fitted to the data, and statistical filtering methods such as Sequential Monte-Carlo algorithms are applied to disentangle the portion of prices caused by spikes from that caused by normal price movements. See for instance Benth et al. (2007) and Gudkov and Ignatieva (2021).

The aforementioned studies apply the threshold to electricity prices. However, other studies, such as Cartea and Figueroa (2005) and Weron and Misiorek (2008), identify spikes through price variations rather than through the level itself. For a more in-depth review of spike identification methodologies, see Janczura et al. (2013).

The first two approaches (fixed and variable price thresholds) are conceptually similar as both set pre-determined thresholds and directly assign a spike label to any prices exceeding such thresholds. The beauty of such methods lies in their simplicity. The third approach based on statistical filtering differs vastly from the first two. An a priori

Table 1
Summary statistics for spikes.

	DA	RT	DART	Spikes		
				$\gamma^- = -30$	$\gamma^- = -45$	$\gamma^- = -60$
Count				3534	2294	1605
Proportion				0.06	0.04	0.03
Mean	39.41	38.96	0.45	−84.44	−110.34	−135.44
Standard Deviation	27.81	46.98	37.58	98.91	114.68	129.19
Median	32.49	27.91	3.54	−55.62	−76.16	−135.44
Min	2.57	−1476.07	−1971.57	−1971.57	−1971.57	−1971.57
Max	424.00	2045.79	1506.74	−30.01	−45.01	−60.02
10%-level quantile	18.40	14.31	−16.88	−157.75	−191.38	−229.92
25%-level quantile	24.04	19.86	−3.28	−90.08	−119.79	−146.09
75%-level quantile	44.00	41.97	10.56	−39.34	−57.06	−73.95
90%-level quantile	65.28	72.15	19.29	−33.34	−49.04	−64.57
Skewness	3.42	7.01	−7.62	−6.82	−6.09	−5.54

Various summary statistics for non-null values of (S^-) labeled spikes. All numbers are expressed in \$/MWh. DA, RT and DART refer to the day-ahead price, the real-time price and the DART spread, respectively.

stochastic generative model for prices embedding the spike-generating mechanism must be specified, and statistical inference, i.e., filtering, methods are applied to estimate the spike component of prices based on its posterior distribution given observed prices. The use of sophisticated statistical filtering methods and the requirement to design a stochastic process matching the complex stylized facts of electricity prices are associated with higher inherent complexity.

This study uses a fixed threshold approach to identify spikes; it is a common choice made in the literature and by practitioners, which makes it possible to avoid the technical complexities that stem from the statistical filtering approach.

2.2.2. Results for DART spread spike identification

In the literature, the target variable used for spikes labeling is often the real-time price. However, for certain market participants, such as virtual bidders who take positions on the day-ahead market and revert them on the real-time market, the payoff is the DART spread. In this study, fixed thresholds are thus applied to DART spreads. The negative spikes are obtained through

$$S_t^- = \begin{cases} \text{DART}_t & \text{if DART}_t < \gamma^-, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for some fixed threshold $\gamma^- < 0$. Thus, observations are labeled as spikes, i.e., $S_t^- \neq 0$, when the DART spread is smaller than the specified threshold for negative spikes. For the remainder of this study, only negative spikes are considered. The size of a spike, when one occurs, is considered to be the DART spread itself; the DART spread is not subdivided into regular and spike components during an occurrence of a spike.

DART spikes thus embed an element of surprise associated with a sudden change of circumstances within a one-day horizon. This element entails that DART spikes are most likely to last less than one day, although definition (1) does not explicitly enforce short-livedness.

The first three columns of Table 1 exhibit summary statistics for the DA prices, RT prices and DART spreads. The standard deviation of RT prices (46.98) is larger than that of DA prices (27.81), highlighting the more volatile nature of RT prices. The large negative DART spread skewness (−7.62) stresses the significant risks to which the virtual bidders are exposed when taking long positions on the DART. As seen in the Long Island electricity price and DART spread time series exhibited in Fig. 1, numerous extreme price events of various sizes are displayed. The following threshold values are considered to capture several spike magnitudes: $\gamma^- = -30, -45, -60$.[2] The last three columns of Table 1 display summary statistics of the spikes, i.e., non-null values of S_t^-, for each threshold. The proportion of labeled spikes varies between 6% and 3%, indicating that only a minority of observations are labeled as spikes.

Fig. 2 displays the autocorrelations of the DART spikes time series $\{S_t^-\}$ for lags extending from 1 to 72 h. Results show that for the three considered thresholds, the autocorrelations are high and statistically significant at lags 24, 48, and 72, indicating the presence of spike clusters lasting multiple days.

Strong seasonal effects are detected in spike occurrences. Indeed, Fig. 3 depicts the proportion of observed spikes across the various months, days of the week, or times of day. Panel A indicates more frequent spikes in either summer or winter months but fewer in fall and spring. Panel C shows more frequent spikes during the late afternoon and fewer at night and in the early morning hours. Surprisingly, the week-versus-weekend effect is not striking, as seen in Panel B.

2.3. Features used for prediction

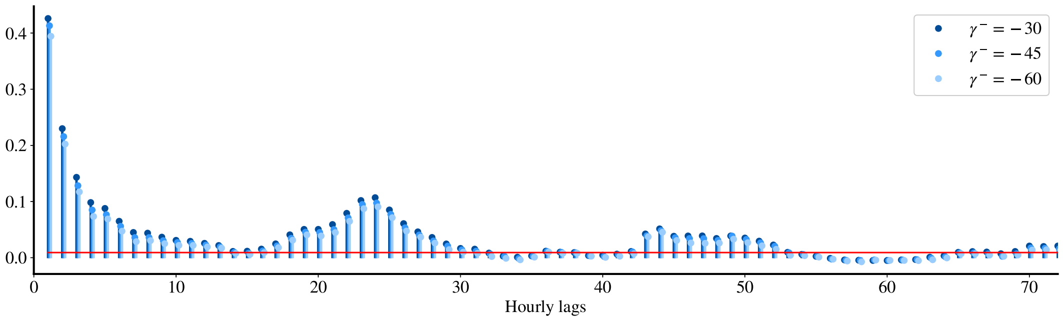
This section discusses and defines the various features, i.e. explanatory variables, considered in the spike prediction analyses. While some features are directly extracted from the raw data, others are obtained through data transformation. These engineered features aim to complement the information set provided by the conventional sources of information available. The steps involved in constructing such features are outlined.

2.3.1. Prediction generation timeline

Features should be included as predictive variables only if they are available at the time when predictions are being generated. The perspective of a market participant placing bids on the day-ahead market is considered herein. The timeline that determines availability or the predictors is now explained.

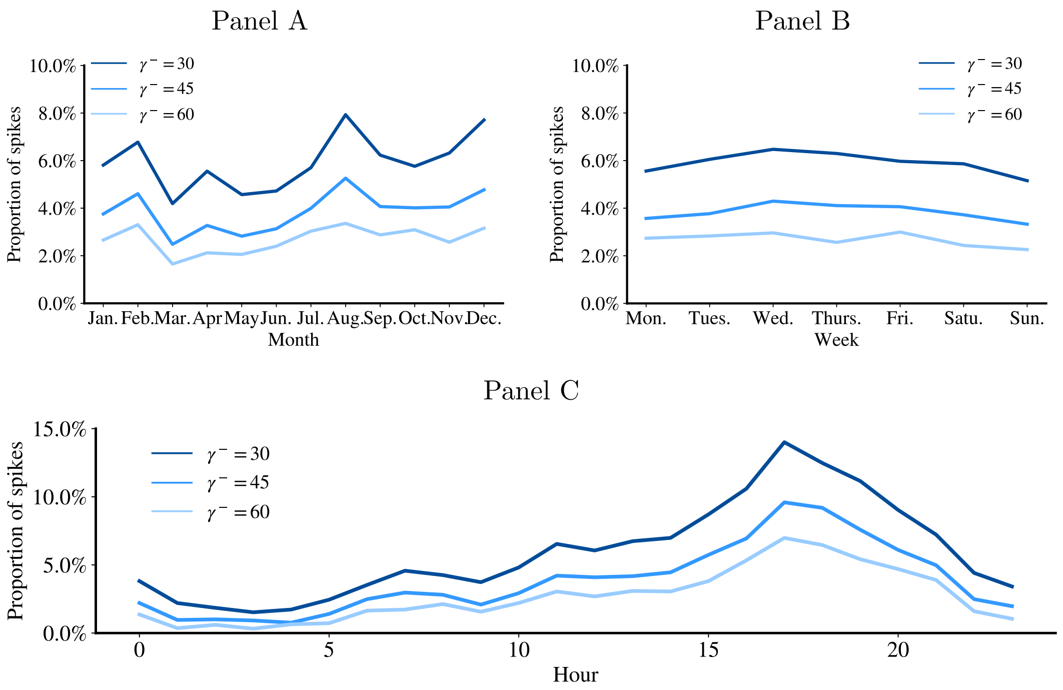
Each daily round of predictions being performed is associated with a window of three consecutive days. The third and last day, referred to as the *target day*, is the day on which all hourly predictions apply. Predictions are performed on the first day, coined as the *prediction day*. The second day (*trading day*) is where day-ahead bids are placed for all hours of the target day. The reason to include a delay between the time at which predictions are performed and the moment at which day-ahead bids are placed is to reflect that participants would typically

[2] The selection of the −\$30/MWh, −\$45/MWh, and −\$60/MWh values is driven by (i) discussions with industrial partners, and (ii) statistical considerations. The first threshold is set to −\$30/MWh since it is the largest negative DART spread considered as an extreme economic event. The smallest threshold is set to −\$60/MWh since it is among the lowest threshold values providing enough spike observations to train the statistical learning models adequately. Indeed, Table 1 highlights that a threshold of −\$60/MWh allows capturing 3% of the observations (1605 data points), indicating events that are sufficiently rare to be considered spikes, but frequent enough to retain sufficient training data.



Autocorrelations of the DART spikes time series $\{S_t^-\}$ for the three considered thresholds. The considered lags extend from 1 hour to 72 hours. The autocorrelations are computed over the whole sample period (2015-2021). The red line exhibits the upper bound of the 95% confidence intervals.

Fig. 2. Autocorrelation of the DART spikes time series.



Proportion of spikes (number of spikes divided by the number of observations in the corresponding hourly/daily/monthly bucket). Fixed thresholds considered are $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$. The data sample extends from January 1, 2015, to October 31, 2021.

Fig. 3. Proportion of spikes per month, day of the week and hour.

require some time to process their predictive analysis outputs and determine their bids. Key elements of the timeline are now presented.

- Prediction day (day 1): At 11:00, the NYISO publishes hourly load forecasts for each hour of the target day. At 18:00, the temperature forecasts for the target day are published. All such information is combined with other available features to produce hourly spike probability predictions at 18:00.
- Trading day (day 2): Bids for the day-ahead participants are placed by 5:00. At 11:00, the NYISO publishes day-ahead prices for the target day.
- Target day (day 3): Real-time prices are revealed throughout the day, allowing for the computation of realized DART spreads.

In summary, all features entering the predictions applying to the target day must be available by 18:00 of the prediction day, i.e., two days in advance. Fig. 4 provides an illustration of the timeline.

2.3.2. The list of features and their construction

The electricity literature identifies multiple features which are known to embed informational content that is useful to forecast prices and price spikes, see for instance Lago et al. (2021). Even though DART spread spike forecasting is a different exercise than price and price spike forecasting, we nevertheless consider features similar to these proposed in such literature.

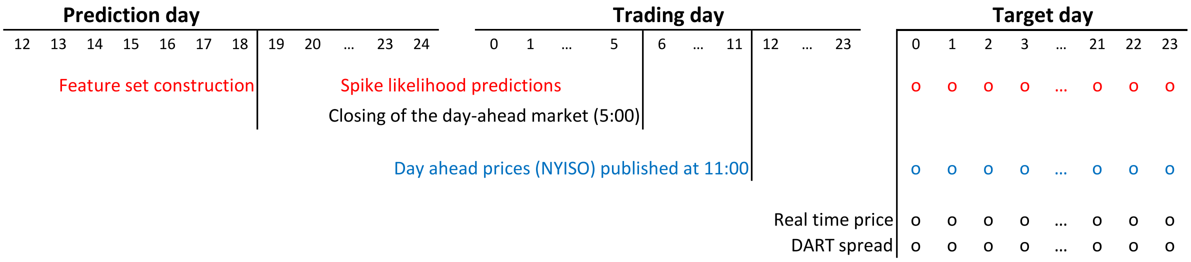


Fig. 4. Timeline for spike predictions and subsequent DART spread realization.

Table 2
Feature variables used for spike prediction.

Forward-looking	Seasonal	Backward-looking
HDD forecast	Hour	Past spikes
CDD forecast	Month	Past day-ahead price error
Load/Grid	Week-end/Holidays	Past day-ahead load error

The set of all selected features, which are listed in Table 2, can be divided into three categories: (1) forward-looking features, (2) seasonal features, and (3) backward-looking features.[3]

The forward-looking features encompass information related to market participants' expectations about the future realization of various variables. For any observation, i.e., target hour, such features include the 48h-ahead load forecast to grid capacity ratio (see below for details), and the time-18:00 prediction day's temperature forecast associated with the corresponding target hour, i.e., the 30h- to 54h-ahead forecast depending on the target hour (see Appendix A for details). As explained below, two non-linear transformations of the latter temperature forecast are considered.

The NYISO (see Itron, 2008) as well as the literature (see Fan et al., 2019, Zahedi et al., 2013 or Yi-Ling et al., 2014) consider non-linear transformations of temperature metrics. This makes it possible to reflect the non-linear relationship between electricity consumption and temperature. Indeed, more electricity is consumed when temperatures are either very low (as heaters are turned on) or very warm (as air conditioners are turned on). A popular methodology described in the literature and adopted by the NYISO (see Itron, 2008) consists in transforming the temperature feature into *heating degree day* (HDD) and *cooling degree day* (CDD) features. The HDD and CDD thus reflect the electricity demand for heating and cooling and are expressed as

$$\text{HDD}_t = \max\left(\text{BP} - T_t, 0\right), \quad \text{CDD}_t = \max\left(T_t - \text{BP}, 0\right),$$

where T_t is the hour-t temperature measurement and $\text{BP} = 18.3°\text{C}$ (65° F) is the breakpoint considered by the NYISO, which is also used herein. When used in conjunction with predictive algorithms that do handle automatically non-linear relationships, the HDD and CDD transformed variables are most likely more appropriate than the original temperature forecast as a predictive feature.

The last forward-looking feature, *load/grid*, corresponds to the ratio of the 48-h-ahead load forecast over the grid transfer capacity supplying Long Island. Indeed, the load-to-capacity ratio has been suggested by Anderson and Davison (2008) as a driver of spike likelihood, although the latter paper considers generation capacity instead of transmission capacity. The interface transfer capacity is used in the denominator to reflect that, for the same amount of load, a curtailed transmission capacity is associated with higher spike risk due to an increase in the likelihood of bottlenecks.

The second class of features, namely the seasonal features, aim to capture well-known seasonal patterns in electricity markets. They include dummy variables for each hour of the day and month of the year, and another dummy variable indicating (additionally) if the day of the target hour is either a weekend day or a holiday.

The last category, the backward-looking features, are engineered features that consist of metrics computed from historical observations. Such features are meant to capture market conditions of the recent past. The three backward-looking features are calculated once at the prediction time, and each of them has identical values for all 24 h of the target date. The first backward-looking feature, *past spikes*, is the number of observed spikes in the 24 h leading up to the prediction. This reflects the tendency of spikes to occur in clusters, as highlighted for instance in Klüppelberg et al. (2010), Christensen et al. (2012), Herrera and González (2014), He and Chen (2016), and Manner et al. (2016). Thus, the presence of many recent spikes indicates a higher likelihood of observing spikes in the near future.

The second backward-looking feature, *past day-ahead load error*, aims to indicate periods where the estimation of near-term future load consumption by the market proves more difficult. Such a feature is helpful because sudden and unexpected surges or drops in load increase the likelihood of observing a spike. The *past day-ahead load error* is computed by summing the hourly squared load forecast errors (real-time load minus the day-ahead load) over the 24-h period leading up to the prediction.

The construction of the third backward-looking feature, *past day-ahead price error*, is analogous to that of the past day-ahead load error; it is also calculated by summing the last 24 hourly observed squared price forecast error (real-time price minus the day-ahead price) in the period prior to the prediction. It is meant to capture periods with higher price volatility.

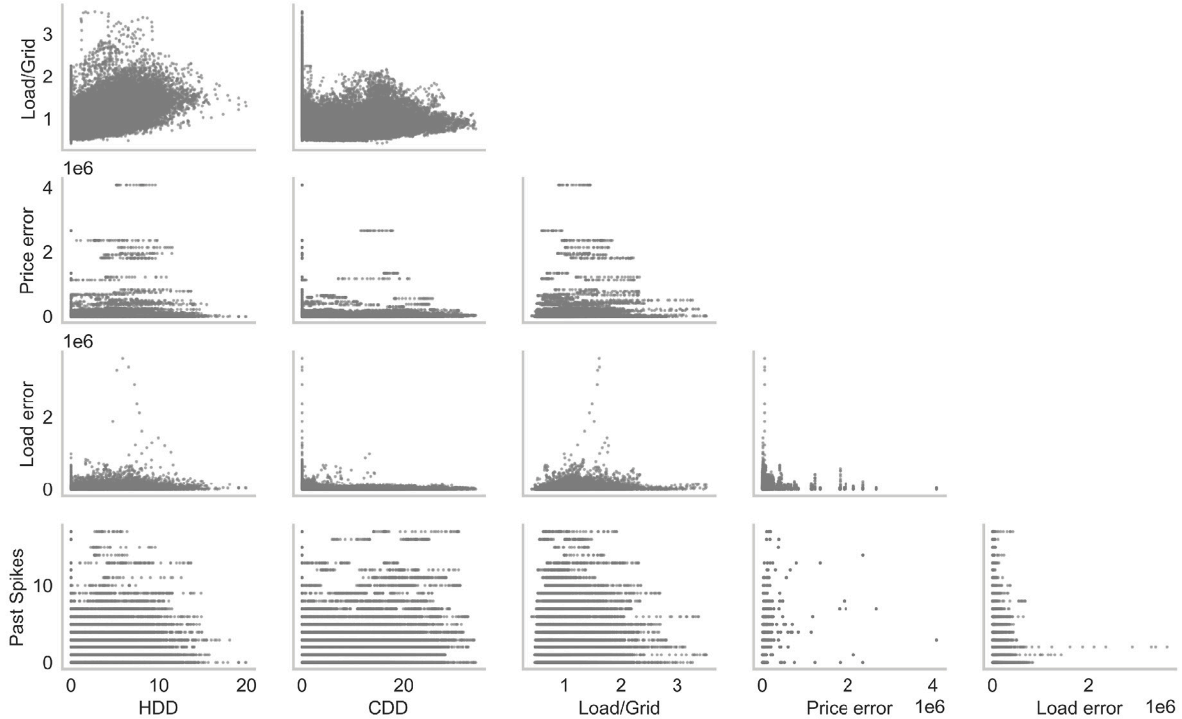
Fig. 5 illustrates the scatterplots of realized values for all considered features in each of the hourly observations, thereby illustrating the relationship between the features. As expected, the HDD and CDD features are positively correlated with *load/grid*. The relationship between the other features does not display any clear dependence structure.

Fig. 6 illustrates kernel density estimates of feature distributions conditional on either the presence or the absence of a DART spike. All panels indicate that larger feature values are more likely to occur when spikes are observed.

3. Spike prediction model

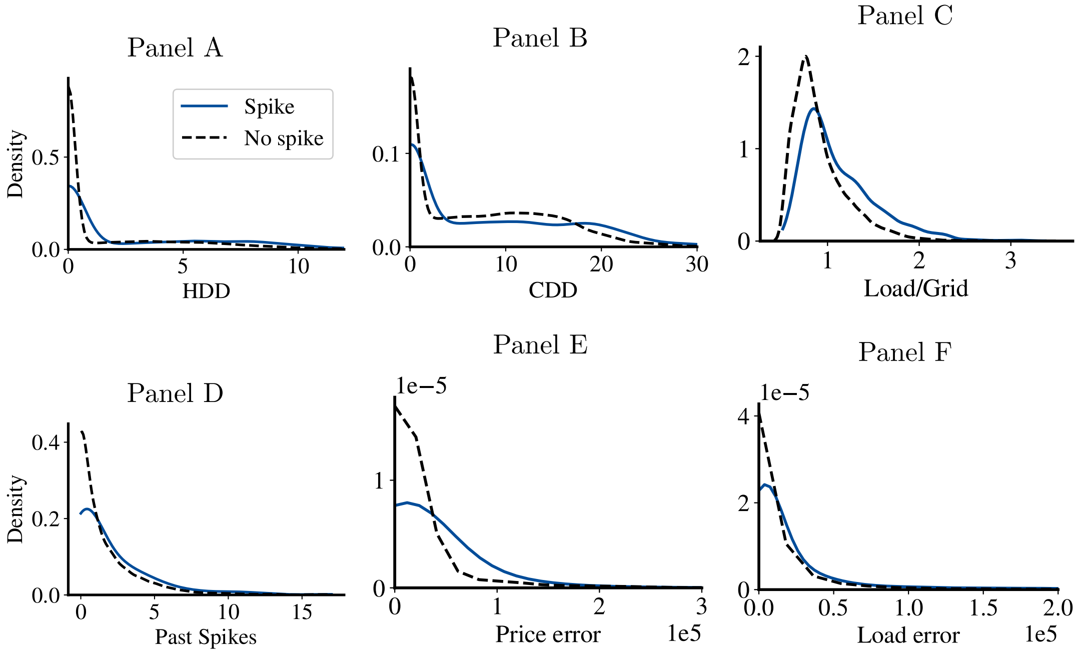
This section illustrates the prediction of DART spread negative spike probabilities based on available information. Four predictive algorithms are applied to the data, and their performance is assessed through conventional statistical metrics. A feature importance assessment evaluating the contribution of each feature to predictive performance is also provided. We hereby focus solely on the spike occurrence probability and leave the challenging task of predicting the spike magnitude to a future study.

[3] An experiment reported in Section D.1 of the Online appendix integrates an additional feature called the *cumulative temperature and humidity index* (CTHI), which is used as predictor by the NYISO to forecast the load. Such inclusion does not improve the general performance of the models. Furthermore, a graphical exploration exercise reveals that the CTHI feature exhibits strong dependence with other features (HDD, CDD and Load/Grid). Therefore, its inclusion in the feature set increases the likelihood of multicollinearity-related issues. For such reasons, the CTHI is not included in the set of features.



Scatterplots of realized values for model features. The presented features are *heating degree days* (HDD), *cooling degree days* (CDD), *Load/Grid* representing the load forecast to grid transfer capacity ratio, *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Fig. 5. Scatterplots of model features.



Each panel illustrates the kernel density estimate of a feature distribution conditional on either the presence (blue continuous line) or the absence (black dashed line) of a DART spike. The presented features are *heating degree days* (HDD), *cooling degree days* (CDD), *Load/Grid* representing the load forecast to grid transfer capacity ratio, *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Fig. 6. Kernel density estimates of feature distributions.

3.1. The predictive models

The four predictive algorithms considered are (1) logistic regressions, (2) random forests, (3) gradient boosting trees, and (4) feed-forward deep neural networks (DNN).[4] Details about their implementation are presented in Appendix B.

The logistic regression is a conventional base case for any binary classification problem. It expresses the logit of the probability of a spike as a linear function of predictors, which makes the model easy to interpret and straightforward to estimate with conventional regression tools. However, logistic regression is not necessarily well suited to handling non-linear relationships with the target variable and interactions between features. Since electricity market price data might be fraught with such complex relationships, it is desirable to contemplate alternative predictive models. Therefore, random forests, boosting trees, and neural networks are also considered as they can automatically represent complex and non-linear interactions.

Except for logistic regression, the models considered here cannot be trained out-of-the-box and require the user to select a set of hyperparameters. The hyperparameter tuning methodology is described in greater detail in Appendix B.2.

3.2. Model performance

Model performance is assessed through two statistical metrics: the area under the receiver operating curve (AUC) and the average log-likelihood. The former is meant to measure the discriminatory power of the models, whereas the latter characterizes the precision of the predictive model. The receiver operating curve (ROC) provides the set of all possible trade-offs between false positive and false negative error rates obtained across the possible choices of probability thresholds for classification (see, for instance, James et al., 2013). The value of the AUC must lie between 0 and 1, with a higher value indicating a higher ability to distinguish between the two classes. An AUC under or equal to 0.5 indicates that the model has no predictive power. The second performance metric, the average log-likelihood, is computed by comparing spike labels and the predicted probabilities:

$$\ell = \frac{1}{\tau}\sum_{t=1}^{\tau} \log(p_t)\mathbb{1}_{\{S_t^- < 0\}} + \log(1-p_t)(1-\mathbb{1}_{\{S_t^- < 0\}}) \tag{2}$$

where p_t corresponds to the model generated probability of observing a spike in hour t, τ is the sample size and S_t^- is zero if and only if no negative spike occurs in hour t.

The performance assessment relies on an expanding window approach consisting in iteratively training the model over an expanding training set for each testing iteration. During the first iteration, the model is trained over the first three years of the dataset. The out-of-sample performance metrics are computed over the following year, i.e., the fourth year. One year is added to the training dataset for the subsequent iteration while generating predictions for the following year. Performance metrics (AUC and average log-likelihood) are computed for training and test set observations.

Panel A of Table 3 displays the AUC for negative spikes. The results presented are for the training sets (in-sample) and test sets (out-of-sample). The last row of each panel displays the aggregated out-of-sample results, i.e., the computed AUC over the merged out-of-sample sets. A larger AUC indicates that the model is more powerful at discriminating between the binary classes. All Panel A entries show an AUC considerably above 0.5 (more precisely, always above 0.65), indicating that the four models exhibit material discriminatory power for every threshold considered. The in-sample AUC is only slightly higher than the out-of-sample AUC, implying that models are not plagued with over-fitting issues. The gradient boosting trees displays the highest aggregated out-of-sample AUC for each threshold ($\gamma^- = -\$30$/MWh (0.722), $\gamma^- = -\$45$/MWh (0.755), and $\gamma^- = \$60$/MWh (0.769)).

However, all models display quite similar out-of-sample AUC across all thresholds. This result is confirmed in Fig. 7, which illustrates aggregated out-of-sample ROC curves for all models and threshold γ^-. The displayed ROC curves are similar in shape and height for all panels, except for the DNN, which is slightly lower than the others. Thus, there is no apparent domination of one model over the others.

Panel B of Table 3 displays the average log-likelihood for each model and every threshold considered (γ^-) and confirms previous findings. Indeed, the gap between the in-sample and out-of-sample model performance is quite small. Furthermore, the models demonstrate similar performance for each threshold, although the gradient boosting trees display a slightly higher aggregated log-likelihood. This showcases that prediction performance is robust to the choice of predictive models. To assess which of the models are the best-performing ones from a statistical standpoint, the Hansen et al. (2011) model confidence set approach is considered.[5] Stars in Panel B's last row identify the best model(s) associated with each threshold. The model confidence set approach indicates that for each threshold, the gradient boosting tree model significantly outperforms the other models, with an exception for the $\gamma^- = -\$60$/MWh threshold where the gradient boosting tree statistically outperforms the random forest and the DNN models but not the logistic regression.

Fig. 8 illustrates the scatterplot matrix of model-generated out-of-sample probabilities for $\gamma^- = -\$60$/MWh. Most of the time, spike likelihoods are moderate. Therefore, one should not expect to predict spikes with very a high degree of certainty. This result raises the question of what level of spike likelihood could be considered substantial enough to become actionable. This issue is investigated in Section 4. Despite the very close performance of the models, material dissimilarities between the individual predicted probabilities are observed, especially for the DNN model. For instance, unlike the other models, the DNN rarely outputs probabilities of observing a spike of over 20%. This result implies that although they all exhibit similar statistical performance, the models might not be considered fully interchangeable. This is further investigated in the next section, where trading strategies based on each model are examined.

Fig. 9 reports the relationship between the proportion of observed spikes and the spike probabilities generated by each model over the out-of-sample period (2018 to 2021). To obtain such figure, the observations are regrouped into buckets based on their model generated spike probability.[6] Each bar indicates the proportion of observed spikes within each bucket. The precision of the model is deemed adequate if the proportions are close to the associated probabilities for each bucket, i.e. if the bars closely follow the identity function (the 45-degree diagonal). This relationship seems to hold reasonably well for buckets associated with low probabilities that contain large numbers of observations. For high-probability buckets, unreported statistical tests highlight that the variability can be attributed to the small number of observations. Indeed, the green dots indicate that only a few dozen observations are associated with high spike probabilities.

[4] A stacked classifier combining the predictions of the four models (logistic regression, random forests, gradient boosting trees, and DNN) is considered in Section D.2 of the Online appendix. Stacking the models does not improve the out-of-sample predictive performance metrics (i.e. the log-likelihood and AUC) in comparison to the standalone models.

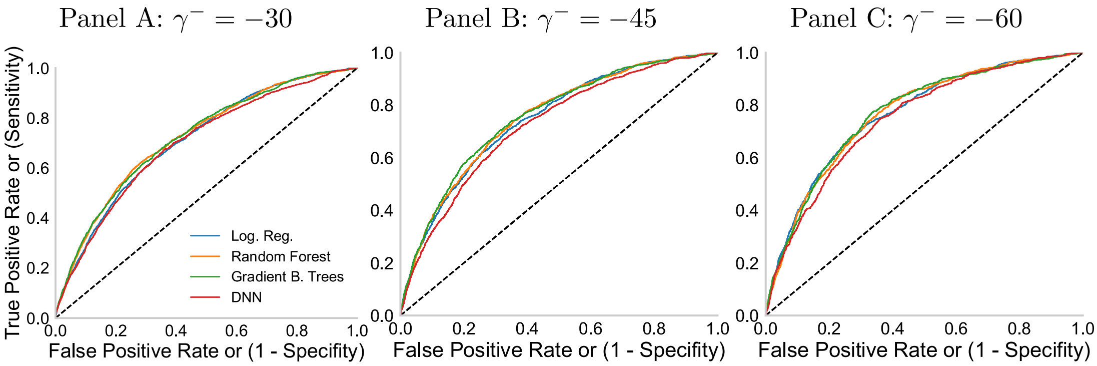
[5] The implementation of the Hansen et al. (2011) model confidence approach is described in detail in Appendix C.

[6] The non-overlapping bucket intervals are of size 2%. The first and last bucket intervals are [0% − 2%[and [48% − 50%] respectively.

Table 3
In-sample and out-of-sample performance metrics.

	γ^-	Logistic regression			Random forest			Gradient boosting trees			DNN		
		−30	−45	−60	−30	−45	−60	−30	−45	−60	−30	−45	−60
A	AUC												
In-sample	2015–2017	0.712	0.730	0.742	0.758	0.784	0.818	0.768	0.783	0.796	0.722	0.699	0.715
	2015–2018	0.704	0.727	0.743	0.743	0.769	0.793	0.757	0.774	0.791	0.725	0.751	0.767
	2015–2019	0.710	0.739	0.755	0.747	0.775	0.798	0.752	0.783	0.806	0.727	0.715	0.734
	2015–2020	0.713	0.740	0.756	0.749	0.776	0.800	0.752	0.791	0.794	0.725	0.746	0.766
Out-of-sample	2018	0.669	0.708	0.729	0.672	0.710	0.729	0.680	0.719	0.726	0.657	0.686	0.718
	2019	0.717	0.771	0.793	0.755	0.789	0.820	0.740	0.785	0.807	0.738	0.745	0.765
	2020	0.740	0.762	0.786	0.741	0.746	0.761	0.740	0.756	0.780	0.705	0.736	0.757
	2021	0.701	0.730	0.756	0.707	0.737	0.740	0.706	0.747	0.756	0.684	0.717	0.741
	Aggregated	0.710	0.745	0.765	0.722	0.751	0.766	**0.722**	**0.755**	**0.769**	0.700	0.723	0.748
B	Average log-likelihood												
In-sample	2015–2017	−0.210	−0.154	−0.121	−0.202	−0.147	−0.113	−0.201	−0.147	−0.115	−0.209	−0.158	−0.124
	2015–2018	−0.213	−0.154	−0.120	−0.207	−0.149	−0.115	−0.203	−0.148	−0.116	−0.209	−0.151	−0.118
	2015–2019	−0.203	−0.145	−0.112	−0.197	−0.140	−0.107	−0.196	−0.138	−0.106	−0.200	−0.148	−0.114
	2015–2020	−0.197	−0.141	−0.107	−0.191	−0.136	−0.103	−0.190	−0.133	−0.103	−0.195	−0.140	−0.106
Out-of-sample	2018	−0.224	−0.155	−0.120	−0.223	−0.156	−0.121	−0.222	−0.155	−0.121	−0.226	−0.157	−0.120
	2019	−0.163	−0.109	−0.079	−0.162	−0.111	−0.079	−0.159	−0.109	−0.079	−0.162	−0.111	−0.080
	2020	−0.172	−0.122	−0.086	−0.169	−0.121	−0.085	−0.167	−0.119	−0.084	−0.172	−0.124	−0.088
	2021	−0.279	−0.196	−0.140	−0.278	−0.195	−0.142	−0.278	−0.194	−0.140	−0.284	−0.201	−0.143
	Aggregated	−0.206	−0.143	−0.105*	−0.205	−0.143	−0.105	**−0.203***	**−0.142***	**−0.104***	−0.208	−0.146	−0.106

The four models are the logistic regression, the random forest, gradient boosting trees, and the deep neural network (DNN). Panel A's performance metric is the area under the curve (AUC), while Panel B is the average log-likelihood. The models generate out-of-sample predictions for 2018 to 2021. The models are trained on the previous years' observations for each out-of-sample forecast. For example, to generate out-of-sample forecasts for 2019, the models are trained on the observations from 2015 to 2018. For each threshold, the Hansen et al. (2011) confidence set approach is applied to the aggregated out-of-sample log-likelihood to identify the set of models whose performance cannot be distinguished from that with the highest performance. The best models remaining in the model confidence set at a level of significance 5% are identified with a star in the table. The testing procedure is described in Appendix C. Boldface numbers highlight models with the highest performance over the aggregated out-of-sample data.



Each panel displays the ROC curves for the four predictive models at a specific threshold (γ^-). The ROC curve illustrates the attained true positive rate on the y-axis against the corresponding false-positive rate on the x-axis. The formula for the true positive rate and false positive rate is True positives/(True positives + False negatives), and 1 − True negatives/(True negatives + False positives) respectively. The ROC curves are computed over the aggregated out-of-sample set (2018 to 2021).

Fig. 7. ROC curves.

3.3. Feature importance assessment

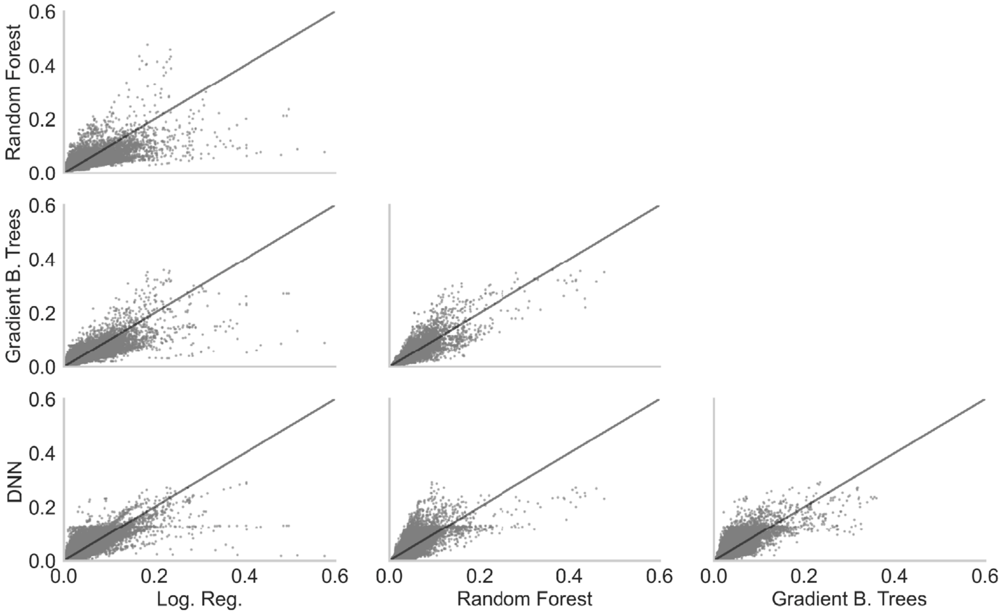
Spike prediction probabilities are constructed by combining the informational content of several features. This section aims to provide information about how each feature contributes to the overall model predictive performance, thereby making it possible to rank features in terms of their absolute and relative importance.

Each feature's importance is assessed through two approaches: (1) Shapley decompositions and (2) marginal performance loss through feature removal.

The Shapley (2016) decomposition has recently been integrated into the machine learning literature through algorithms referred to as SHAP (Lundberg and Lee, 2017) or SAGE (Covert et al., 2020). They make it possible to decompose individual predictions (in SHAP) or their total predictive performance (in SAGE) into a sum of contributions from the various features, thereby making it possible to evaluate their respective importance. This study focuses on the SHAP algorithm, in which the feature i contribution to the spike probability predictions made for hour t is defined as

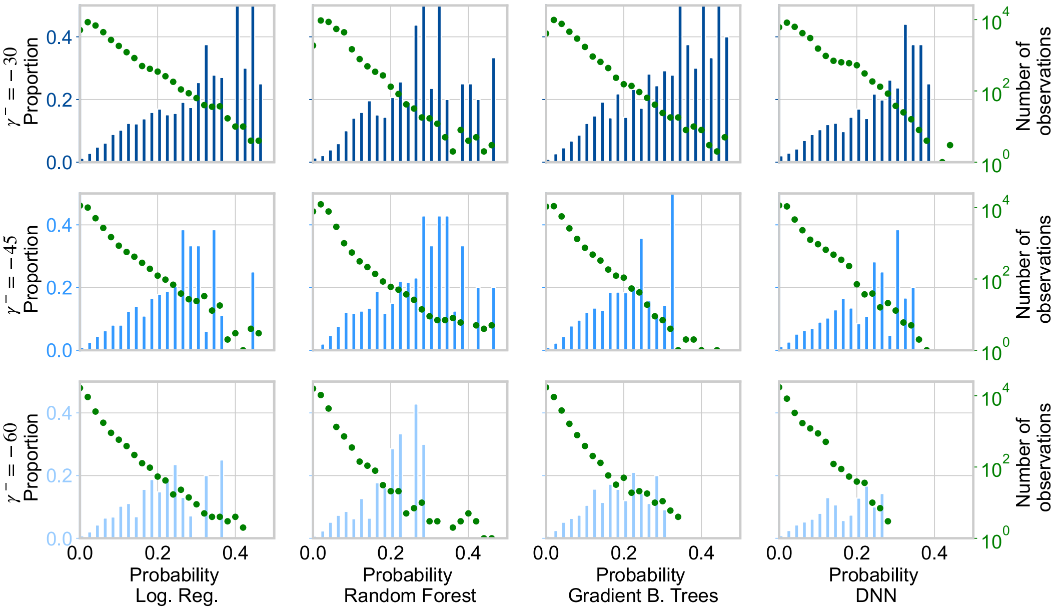
$$\phi_{i,t} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{t,S\cup\{i\}}) - f_S(x_{t,S}) \right]$$

where F is the set of all predictors, $|\cdot|$ denotes the cardinality of a set, $x_{t,S}$ is the hour-t features values for the subset of features S and $f_S(x_{t,S})$ is the spike probability generated by the model trained exclusively with predictors S. It quantifies adjustments to predictions when the subsets of features are incremented with predictor i. The Shapley decomposition has the favorable property of explaining each



The panels display the scatterplots for all pairs of predicted spike probabilities belonging to the aggregated out-of-sample set (2018 to 2021). The threshold is $\gamma^- = -60$. The identity curve is also displayed in each panel.

Fig. 8. Scatterplots of predicted spike probabilities across models for $\gamma^- = -60$.



Each panel's x-axis reports spike probability intervals, while the y-axis displays the proportion of spikes observed relative to the number of observations inside the interval. The figures are obtained using the out-of-sample data period (2018 to 2021). The green dots display the number of observations in each interval with a log-scale y-axis.

Fig. 9. Proportion of spikes vs predicted probability.

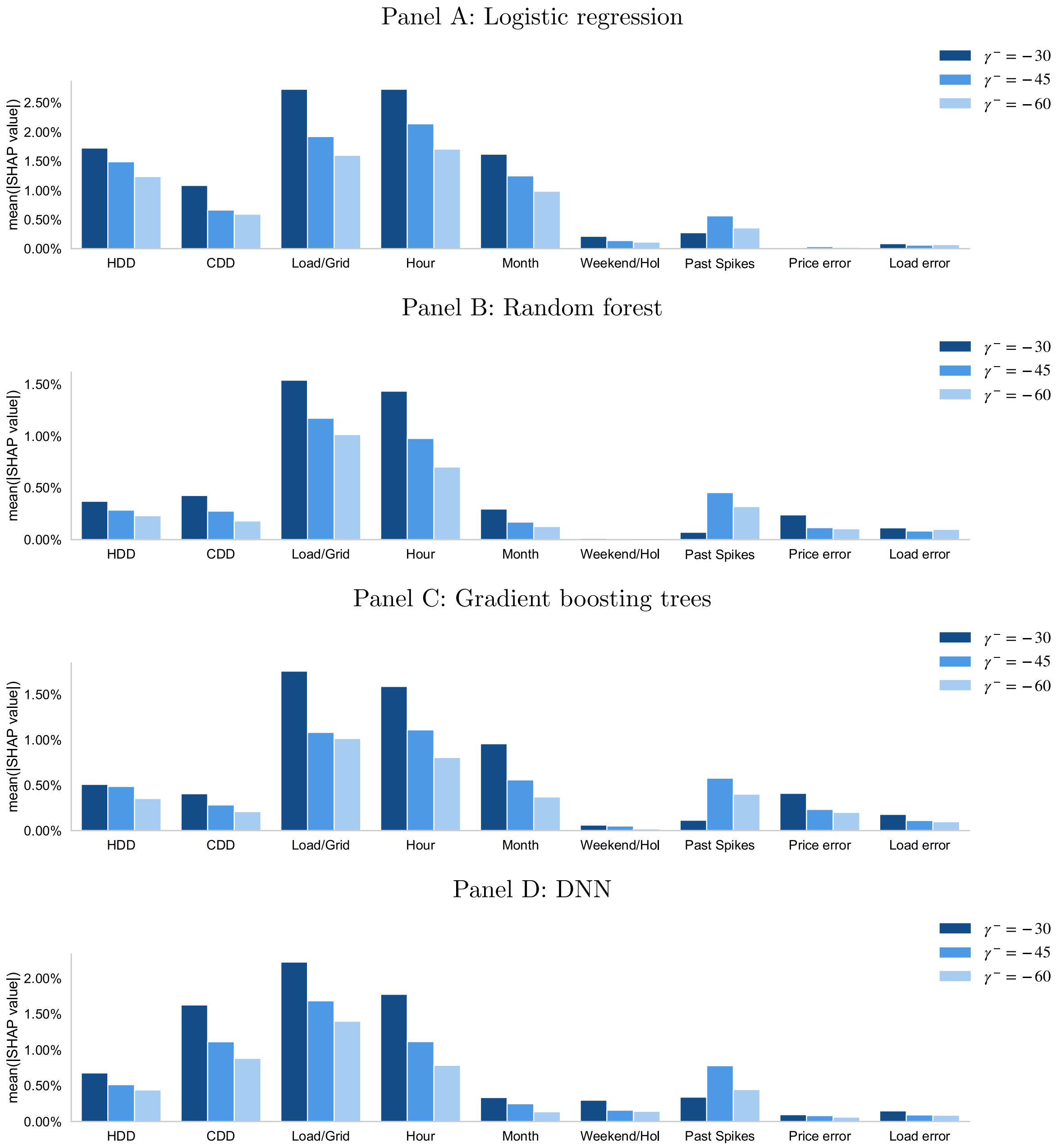
prediction as the sum of its contributions:

$$f_F(x_{t,F}) = \phi_{\emptyset,t} + \sum_{i\in F} \phi_{i,t}.$$

To measure the importance of each respective feature, the average absolute feature contributions are presented:

$$\psi_i = \frac{1}{\tau}\sum_{t=1}^{\tau} |\phi_{i,t}|, \tag{3}$$

with larger values of ψ_i relative to other features meaning that feature i is more impactful when making predictions. SHAP values are computed using the `Python` package `shap`.



Panel A: Logistic regression

Panel B: Random forest

Panel C: Gradient boosting trees

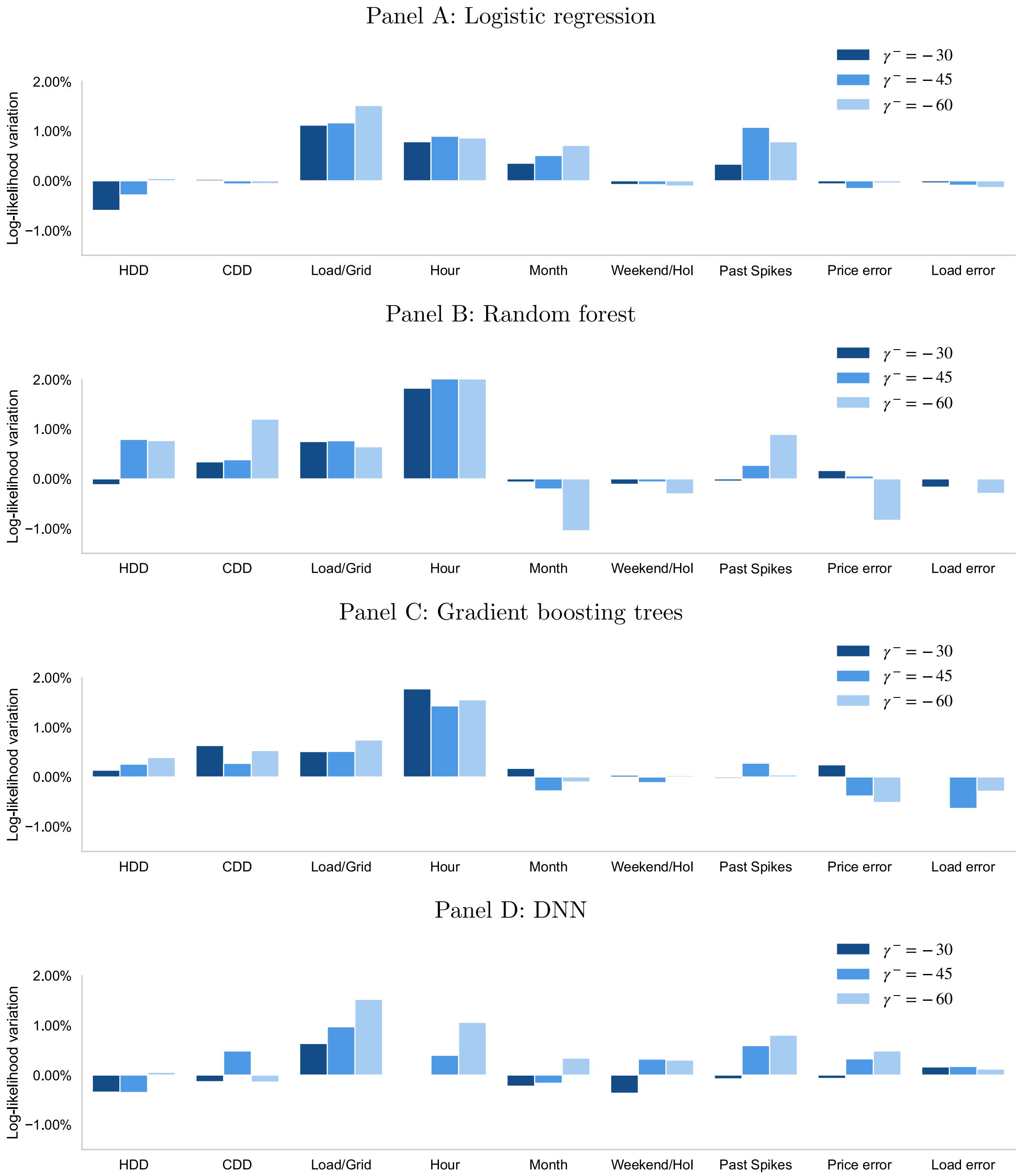
Panel D: DNN

Each panel reports, for the three thresholds considered, the features' mean absolute SHAP contributions over the out-of-sample period (2018 to 2021). The features are *heating degree days* (HDD), *cooling degree days* CDD, *hour* indicators, *month* indicators, *weekend/holidays* indicators (Weekend/Hol.), *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error). The SHAP values for the categorical features *month* and *hour*, which are divided into buckets for the logistic regression, are computed by summing the SHAP values of each category.

Fig. 10. Shapley additive explanation values.

Fig. 10 reports the mean absolute Shapley values (3) computed over the out-of-sample period for every feature, model and threshold considered. Results indicate that the load forecast over transfer capacity ratio (*load/grid*), the hourly and monthly indicators (*hour* and *month*), the CDD/HDD and *past spikes* features contribute the most to the predictions. Interestingly, for all predictive models, the *past spikes* feature offers a much higher contribution when the threshold γ^- is large than when it is small. Other features such as *weekend/holidays* and *past day-ahead load error* exhibit generally low contributions to the predictions.

To complement the information provided by SHAP, this study quantifies the marginal performance loss through the decrease in out-of-sample average log-likelihood observed when any of the features are omitted from the set during training. A drop in performance implies that the feature does bring useful information, while minuscule improvements up to degradation in performance suggest that the feature



Panel A: Logistic regression

Panel B: Random forest

Panel C: Gradient boosting trees

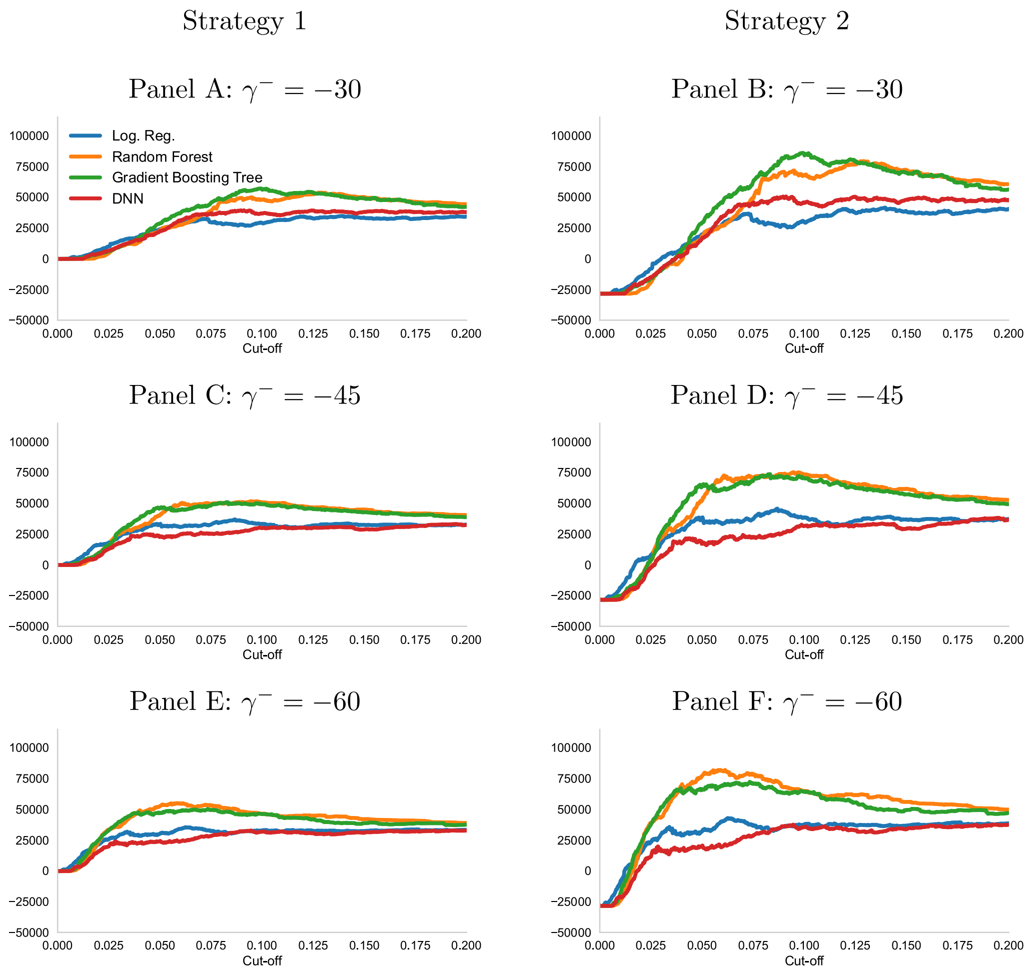
Panel D: DNN

Each panel reports, for the three considered thresholds, the percentage decrease in the out-of-sample log-likelihood when the feature is excluded from the feature set. More precisely the model is re-trained with a reduced feature set where only the targeted feature is removed. The table reports the ratio of the difference between the out-of-sample log-likelihood from both model (full model minus reduced model) over the log-likelihood of the full model. The features are *heating degree days* (HDD), *cooling degree days* CDD, *hour* indicators, *month* indicators, *weekend/holidays* indicators (Weekend/Hol.), *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Fig. 11. Decrease in average log-likelihood when removing a single predictor.

conveys little to no information. Fig. 11 exhibits such percentage increase/drops in the average log-likelihood. Features with the highest contribution are the load forecast to transfer capacity ratio, hourly indicators, and the number of spikes in the previous day. Conversely, features *past day-ahead price error*, *past day-ahead load error* and *weekend/holidays* once again convey little to no predictive power relative to the other features for every model. Such findings are mostly consistent with those provided by the SHAP algorithm.[7]

[7] A model performance assessment with a revised feature set selected based on the results of the present section is reported in Section D.3 of the Online



Panel A: $\gamma^- = -30$

Panel B: $\gamma^- = -30$

Panel C: $\gamma^- = -45$

Panel D: $\gamma^- = -45$

Panel E: $\gamma^- = -60$

Panel F: $\gamma^- = -60$

Each panel displays each model's total P&L as a function of the cut-off value over the training set spanning from 2015 to 2017. Strategy 1 (left panels) and 2 (right panels) take a long position in the DART spread if the predicted spike probability is below the cut-off point. While Strategy 1 takes no position when the spike probability is above the cut-off point, Strategy 2 takes a short position in such case. The volume of the positions taken for both strategies is always 1 MWh.

Fig. 12. Total P&L for a continuum of cut-off values.

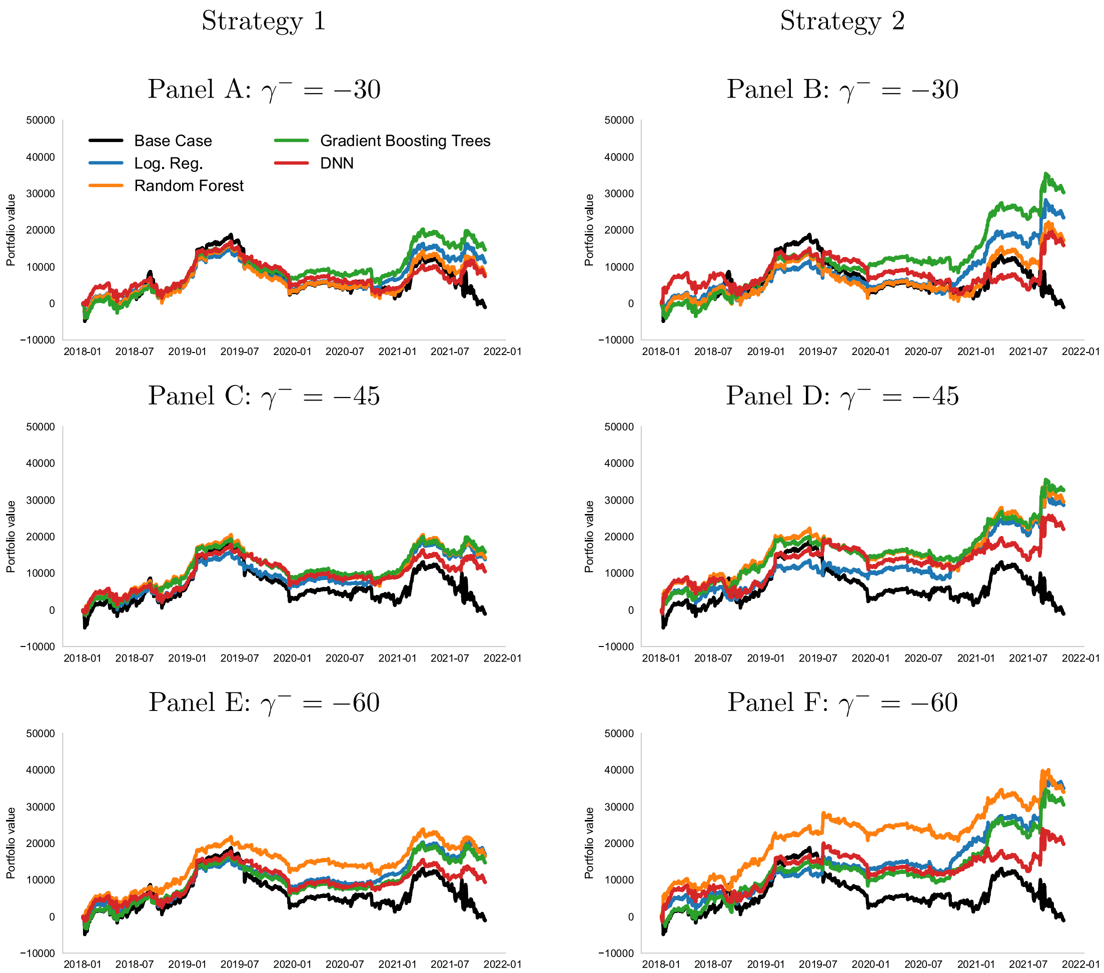
The Shapley (2016) decomposition and the marginal performance loss provide different information about the features contribution. While the Shapley decomposition quantifies the extent to which the model relies on each respective feature, the marginal loss assesses the incremental performance gain/loss when the feature is included into the model. The Shapley (2016) decomposition indicates that the models strongly rely on some of the variables that lead to low marginal performance gains, or even to a performance loss. This phenomenon is the result of strong dependence between certain features. For example, the *month* feature is extensively used by the models, but its associated marginal loss is mainly negative (i.e. dropping such variable improves out-of-sample performance). This result can be attributable to the fact that other features, such as *load/grid*, *heating degree days*, and *cooling degree days*, exhibit strong dependence with the *month* feature and already intrinsically capture the information provided by the latter quantity related to spike likelihood prediction.

4. Trading strategies performance

This section aims at evaluating the performance of the four models from an economical (rather than statistical) perspective. A large strand of the electricity literature integrates price forecasting methods to devise the trading strategies (see Conejo et al., 2005, Zhang et al., 2012, Ziel et al., 2015, Lago et al., 2018, and many others). This study concentrates only on forecasting DART spikes as a complement to such methods.

In a first exercise, two trading strategies are implemented over the in-sample set, i.e. from January 2015 to December 2017. Both strategies take a long position on the DART when the model-predicted probability is lower than a predetermined cut-off point. The first strategy takes no position otherwise, i.e., when the probability exceeds the cut-off point, while the second strategy takes a short position in the DART. The first strategy reflects the situation of a participant trying to collect the DART premium when the risk of a spike is not too considerable, while in the second strategy, the participant tries to also benefit from the occurrence of a spike. The volume of any position taken in the two strategies is always 1 MWh.

appendix. Results are not reported in the body of the article because the model performance with the revised feature set is improved artificially due to data leakage; the decision to remove some features from the feature set described in Section 2 is based on performance results computed over the out-of-sample period, thus leaking information from the out-of-sample period into the feature selection procedure.



Each panel illustrates the time series of a portfolio value starting at $0 following strategies 1 and 2 over the out-of-sample period (2018 to 2021). Strategy 1 (left panels) and 2 (right panels) take a long position in the DART spread if the predicted spike probability is below the cut-off point. While Strategy 1 takes no position when the spike probability is above the cut-off point, Strategy 2 takes a short position in such a case. The cut-off point for threshold $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$ are respectively 12%, 7% and 5%. The volume of the positions taken for both strategies is always 1 MWh. The base case strategy consists in taking a long position each period.

Fig. 13. Portfolio value over time.

This first exercise is completely in-sample and aims to (1) better understand the effect of the cut-off point on the strategies' performance and (2) select a cut-off point for the subsequent out-of-sample exercise. The left (respectively right) panels of Fig. 12 display the cumulative hourly profits and losses (P&L) of the first (respectively second) strategy as a function of the cut-off point over the in-sample horizon.

Every panel of Fig. 12 unanimously indicates that profits are maximized for relatively small cut-off points ranging between 5% and 12%, depending on the threshold γ^-. This result is explained by the asymmetry of the loss function, where gains associated with the successful prediction of a spike far exceed the losses incurred when falsely predicting a spike. Such asymmetry encourages the participant to short the DART spread as soon as a small potential spike signal is detected. Comparing the two strategies, Strategy 2 outperforms the first one, pointing toward the added value of short DART positions when the spike probability is high.

The two same trading strategies are implemented in a second exercise, this time over the out-of-sample set (from January 2018 to October 2021). To avoid information leakage, the considered cut-off points are selected based on the aforementioned first in-sample exercise exhibited in Fig. 12 and are chosen as respectively 12%, 7% and 5% for $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$.[8] To further assess the added value of the spike probability forecast as a trading signal, the two strategies are compared against a third one, a base case strategy consisting in always taking a long position in the DART.

The left (respectively right) panels of Fig. 13 illustrate the time evolution of a portfolio value starting at $0 and invested in the first (respectively second) strategy. Looking at Strategy 1, for all thresholds and models, the portfolio value time series appears to closely follow the upward trends of the base case portfolio from January 2018 to June 2019 and January to June 2021 periods while being much less impacted by the downward trends over June 2019 to January 2021 and May to October 2021 periods. Furthermore, Strategy 1 generates substantially higher profits over the out-of-sample period while holding fewer positions than the base case strategy. The second strategy exhibits a different behavior where the portfolio value tends to increase steadily over time, except for a few downward stretches.

[8] The cut-off points are selected as the values maximizing the cumulative in-sample P&L of the gradient boosting trees model.

Table 4
Average and total out-of-sample P&L.

		γ^-	Logistic regression			Random forest			Gradient boosting trees			DNN		
			−30	−45	−60	−30	−45	−60	−30	−45	−60	−30	−45	−60
Strategy 1	2018	Avg.	0.77	1.02	1.07	0.75	1.35	1.55	0.93	1.32	1.15	0.96	0.98	1.04
		Total	6060	7612	7921	6201	10701	12063	7430	10197	8490	7545	7644	7973
	2019	Avg.	−0.29	−0.13	0.01	−0.34	−0.27	0.08	−0.08	−0.18	−0.25	−0.31	−0.03	−0.10
		Total	−2403	−1077	78	−2881	−2266	690	−685	−1449	−2014	−2581	−251	−799
	2020	Avg.	0.40	0.45	0.52	0.04	0.19	0.24	0.28	0.28	0.44	−0.10	0.23	0.26
		Total	2931	3185	3624	364	1522	1887	2246	2138	3280	−760	1628	1810
	2021	Avg.	0.79	0.75	0.99	0.71	0.75	0.33	0.89	0.86	0.85	0.60	0.28	0.08
		Total	4617	4083	5370	4416	4301	1846	5652	4918	4999	3217	1503	415
	Agg.	Avg.	0.38	0.49	0.61	0.26	0.48	0.56	0.48	0.54	0.51	0.25	0.37	0.34
		Total	11205*	13803*	16992*	8099	14258*	16486*	14644*	15804*	14756*	7422	10524*	9399
Strategy 2	2018	Avg.	0.59	0.94	1.01	0.62	1.65	1.96	0.90	1.53	1.14	0.93	0.95	1.02
		Total	5144	8248	8866	5425	14425	17151	7885	13418	10005	8115	8312	8970
	2019	Avg.	−0.10	0.20	0.47	−0.21	−0.07	0.61	0.29	0.12	−0.01	−0.14	0.39	0.27
		Total	−867	1785	4094	−1823	−592	5320	2570	1040	−88	−1223	3437	2342
	2020	Avg.	0.72	0.78	0.88	0.13	0.40	0.48	0.56	0.54	0.80	−0.12	0.42	0.46
		Total	6315	6824	7700	1180	3497	4226	4945	4728	7013	−1067	3709	4072
	2021	Avg.	1.76	1.61	1.97	1.70	1.67	1	2.04	1.84	1.86	1.38	0.91	0.61
		Total	12838	11770	14343	12435	12206	7295	14908	13440	13602	10038	6610	4433
	Agg.	Avg.	0.70	0.85	1.04	0.51	0.88	1.01	0.90	0.97	0.91	0.47	0.66	0.59
		Total	23430*	28626*	35003	17217	29536*	33992*	30308*	32628*	30532*	15863	22068*	19817
			2018	2019	2020	2021	Agg.							
Base case		Avg.	0.79	−0.45	−0.05	−0.49	−0.03							
		Total	6976	−3939	−453	−3603	−1020							

The table presents P&L statistics for the two trading strategies considered and the base case. The statistics are divided by year (2018, 2019, 2020, and 2021) and combined over the out-of-sample period (2018 to 2021) under the rows named “Aggregated". Both strategies take a long position on the DART spread when the model predicted probability remains under the pre-determined cut-off point, i.e., 12%, 7% and 5% for $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$ respectively. The first strategy takes no position otherwise, i.e., when the probability does exceed the cut-off point, while the second strategy takes a short position in the DART otherwise. The summary statistics are the total P&L and the average profit per position. The Hansen et al. (2011) confidence set approach, identifying which of the models have a performance that is statistically indistinguishable from that of the best model, is applied to the total P&L for each threshold. The best models remaining in the model confidence set at a level of significance 5% are identified with a star in the table. The testing procedure is described in Appendix C.

Table 4 illustrates the average P&L per position and the total P&L organized by model, threshold, and year. The results for Strategy 1 indicate that every model generates a positive aggregated P&L, unlike the base case strategy, which generates a negative aggregated P&L. For all predictive models, the cumulative profit is generally more significant for lower thresholds (−$45/MWh and −$60/MWh) than for the higher ones (−$30/MWh). In 2018 and 2021, Strategy 1 generates a strong total P&L, while in 2019 and 2020, it is more modest or even negative. Nonetheless, the four algorithms produce a higher total P&L in most years compared with the base case strategy. As expected, the P&L produced by Strategy 2 is more prominent than that of Strategy 1 since Strategy 2 additionally benefits from the short positions on the DART spread when the spike probability is sufficiently high.

As in Section 3, the Hansen et al. (2011) model confidence set approach is harnessed to identify the best-performing model(s) for each threshold using the P&L as the loss function. The model confidence set approach is not used across strategies, i.e. to compare Strategy 1 and Strategy 2, but rather within each strategy to identify the best associated predictive models. Stars in the last row of each of the Table 4 panels identify models with superior predictive ability. Results show that no single model significantly outperforms all others. The gradient boosting tree model and the logistic regression are always included in the best-performing set, while the random forest is included for the −$45/MWh and −$60/MWh thresholds. However, the DNN model is only included in the best-performing set at the −$45/MWh threshold, indicating that the DNN is, overall, the least-performing model in terms of generated P&L. Sets of best-performing models using the P&L as the loss function do not coincide with these using the log-likelihood that are presented in Section 3. Such disparities are mainly explained by the more volatile nature of the P&L in comparison to log-likelihood scores, which makes the discrimination between models more difficult when using the P&L as the performance metric.

Table 5 reports a risk-adjusted performance measure (the Sortino ratio) and two risk metrics (the semi-deviation and the Value-at-Risk) for both strategies as well as the base case.[9] The Sortino ratio is (1) always positive for the two developed strategies, in opposition to that of the base case, which is negative, and (2) large, i.e., always greater than one and most of the time greater than two. Strategy 2 generally produces Sortino ratios larger than those of Strategy 1, indicating that shorting the DART in periods of high spike probability improves the strategy's risk/reward profile. Both the semi-deviation (Std$^-$) and the Value-at-Risk at confidence level 1% (VaR 1%), are substantially smaller for both Strategy 1 and Strategy 2 than for the base case strategy. Therefore, the results outline that for every model and threshold, Strategy 1 and Strategy 2 lead to significantly larger P&Ls than the base case strategy and embed lesser risk (especially tail risk), thus highlighting the twofold contribution of integrating the model signals to a trading strategy.

Table 6 exhibits the precision and recall metrics as well as the dissected aggregated out-of-sample average P&L of Strategy 2. The precision illustrates the ratio of realized predicted spikes over the total number of predicted spikes. The recall considers the proportion of labeled spikes predicted by the models relative to the sample's total

[9] The Sortino ratio is computed as follows:

$$\frac{\frac{8760}{\tau}\sum_{t=1}^{\tau} P_t}{\sqrt{\frac{8760}{\tau}\sum_{t=1}^{\tau}(P_t)^2 \mathbb{1}_{\{P_t<0\}}}}$$

where P_t is the hour-t P&L, and τ is the number of hours in the sample. The constant $(365 \times 24 = 8760)$ corresponds to the number of hours in a year and is applied for annualization purposes.

Table 5
Risk-adjusted metrics for the out-of-sample hourly P&L.

		Logistic regression			Random forest			Gradient boosting trees			DNN		
	γ^-	−30	−45	−60	−30	−45	−60	−30	−45	−60	−30	−45	−60
Strategy 1	Sortino ratio	1.64	2.32	3.03	1.03	2.20	2.73	2.05	2.55	2.32	1.12	1.68	1.58
	Std⁻	21.80	19.80	18.83	23.42	20.25	19.26	21.80	19.88	20.67	21.11	20.55	20.14
	VaR 1%	−80.77	−75.01	−72.07	−92.08	−82.44	−79.31	−85.28	−78.21	−78.34	−83.41	−80.75	−77.14
Strategy 2	Sortino ratio	2.45	3.08	4.01	1.75	3.16	3.65	3.23	3.59	3.26	1.79	2.38	2.06
	Std⁻	26.69	25.90	24.30	27.47	26.02	25.94	26.11	25.34	26.09	24.75	25.88	26.78
	VaR 1%	−83.48	−77.85	−76.44	−94.72	−85.94	−84.53	−89.09	−82.46	−82.46	−85.24	−82.73	−80.86
Base case	Sortino ratio	−0.09											
	Std-	32.61											
	VaR 1%	−105.63											

The table reports (1) the Sortino ratio, (2) the semi-deviation (Std.⁻), and (3) the Value-at-Risk at confidence level 1% (VaR 1%) for hourly P&L of the considered strategies over the out-of-sample period (2018 to 2021). Both Strategies 1 and 2 take a long position on the DART when the model-predicted probability is lower than the pre-determined cut-off point, i.e. 12%, 7% and 5% for $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh respectively. The first strategy takes no position otherwise, i.e., when the probability exceeds the cut-off point, while the second strategy takes a short position in the DART otherwise.

Table 6
Precision/Recall and Strategy 2 hourly P&L dissected.

	γ^-	Logistic regression			Random forest			Gradient boosting trees			DNN		
		−30	−45	−60	−30	−45	−60	−30	−45	−60	−30	−45	−60
	Precision	0.16	0.12	0.09	0.16	0.13	0.09	0.18	0.13	0.09	0.16	0.12	0.08
	Recall	0.21	0.34	0.39	0.14	0.24	0.29	0.22	0.34	0.37	0.20	0.33	0.38
Average P&L	True pos.	120.29	136.06	169.37	136.38	152.94	184.01	116.53	138.73	170.27	121.21	141.54	171.93
	False pos.	−17.62	−12.82	−10.03	−18.96	−15.45	−11.95	−17.61	−13.42	−11.59	−15.05	−12.04	−9.03
	False neg.	−70.80	−91.96	−112.44	−72.05	−92.67	−114.39	−71.36	−91.03	−113.01	−71.37	−90.05	−111.78
	True neg.	4.06	3.15	2.61	4.29	3.31	2.69	4.11	3.14	2.52	4.28	3.21	2.68
	Avg Pos.	5.03	5.57	6.00	6.19	6.64	6.17	6.10	5.97	5.39	7.15	6.01	5.89
	Avg neg.	0.37	0.60	0.68	0.30	0.44	0.48	0.43	0.61	0.55	0.51	0.66	0.71

The first two rows of the table exhibit the precision (proportion of true spikes among the observations predicted to be spikes) and recall (proportion of detected spikes among all spikes) for each model. Rows 3 to 6 illustrate the average P&L associated with either the true positives (True pos.)–predicted spikes that are effectively spikes; false positives (False pos.)–predicted spikes that did not materialize; false negatives (False neg.)–spikes that were not predicted; and true negatives (True neg.). The last two rows detail the average P&L when the model either predicts a spike (Avg pos.) or no spike (Avg neg.). Results are computed over the out-of-sample period (2018 to 2021). The cut-off point for thresholds $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh are respectively 12%, 7% and 5%.

number of labeled spikes.[10] The precision is relatively low for every threshold and all models, i.e., between 9% and 16%, which is not surprising considering the low cut-off points used for the predictions. For all models, the recall is approximately 20%, 33%, and 37% for thresholds $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh, respectively.

The average P&L associated with the true positives also corroborates this conclusion, where the average P&L increases with the thresholds. However, it is interesting to note that the average P&L for false positives is surprisingly low (between −$10 and −$20). Therefore, high probabilities of observing spikes predicted by the models appear to be associated with periods of high DART spread volatility and uncertainty. Unsurprisingly, the average P&L of false negatives is strongly negative. However, it is worth noting that the average P&L of false negatives is, in absolute terms, much lower than the average P&L of true positives. This result indicates that, on average, the models capture the more significant spikes. The last two rows of Table 6 display the average P&L when the models predict a spike (Avg. pos.) and inversely when the models predict no spikes. Both scenarios generate positive P&L for every model and each threshold. The average P&L is much larger when the model predicts a spike. Nonetheless, the results clearly show the potential of integrating the model-generated signals into any trading strategy.

5. Conclusion

This paper studies the forecasting of DART spread spike probabilities. DART spreads of the Long Island zone of the NYISO market are considered in developing the model. A fixed threshold methodology commonly used in the literature is applied to identify spikes in the data. A tailor-made feature set is proposed to perform predictions.

Four statistical learning approaches are considered in predicting spike occurrences. Results indicate that for every threshold considered, all models produce similar predictive performance, although the gradient boosting trees slightly outperform their counterparts. A feature importance assessment highlights the critical importance of forward-looking features such as the load forecast over transfer capacity ratio, predicted heating degree days and predicted cooling degree days, of seasonal features such as hourly and monthly indicators, and of the backward-looking feature counting the number of spikes in the last 24 hours before the prediction. Conversely, the weekly cycle indicators and some of the backward-looking features measuring near-past load or price prediction errors exhibit lesser importance.

Finally, a backtest of two trading strategies integrating model-generated spike probabilities as a market signal is implemented. Such strategies are shown to produce significantly higher profits, lesser risk, and thus larger risk-adjusted performance in comparison to a base case

[10] The precision (P) and recall (R) are calculated as follows:

$$P = \frac{Tp}{Tp + Fp}, \qquad R = \frac{Tp}{Tp + Fn},$$

where Tp denotes the True positives, Fp the false positives, and Fn the false negatives. A false positive is a predicted spike that did not materialize.

strategy systematically holding long DART spread positions. Therefore, results highlight the added value of the developed signal from an economic perspective.

Future questions worth examining include: (1) applying the prediction scheme to positive DART spikes, (2) determining if the framework developed works well in other nodes of the grid and other power markets, and (3) designing trading strategies where trade volumes are modulated based on the intensity of the spike probability signal to improve profitability, or where the cut-off point driving the trade direction varies depending on the season.

Appendix A. Weather forecast simulation and interpolation

A.1. Temperature forecast interpolation

The temperature forecast data consist of daily weather forecasts as of 18:00 with horizons ranging from 30 to 54 hours in three-hour increments, i.e., three-hour forecast periods corresponding respectively to 00:00, 3:00, 6:00, …, 21:00. Hourly spike forecasts performed in Section 3 require hourly temperature forecasts as input. To handle the missing temperature forecast for hours falling between the three-hour increments, a simple linear interpolation scheme is implemented. More precisely, interpolated temperature forecasts $\widehat{\text{TF}}$ are computed as follows:

$$\widehat{\text{TF}}_{3t+i} = \frac{(3-i)\text{TF}_{3t} + i\text{TF}_{3(t+1)}}{3}, \quad \text{where } i = 1, 2,$$

where TF is the observed temperature forecast.

A.2. Synthetic temperature forecasts before October 2017

The historical temperature forecast dataset obtained is only available as of October 7, 2017. To fill for missing observations in the dataset that starts in 2015, simulated forecasts are generated by injecting noise in realized value. Temperature forecasts are described as a combination of the realized temperatures and error terms:

$$\text{TF}_t = \text{RT}_t + \epsilon_t,$$

where TF is the temperature forecast, RT is the realized temperature, and ϵ is the forecast error. To simulate temperature forecasts before October 7, 2017, a noise component ϵ_t is added to the realized temperature RT_t. Because the forecast error might be influenced by a multitude of seasonal factors, the noise component is sampled using a simple bootstrapping approach to circumvent this potential complexity.

The first step consists in storing the available forecast errors ϵ_t from October 7, 2017, to December 31, 2021. The second step consists in simulating the temperature forecasts from January 1, 2015, to October 6, 2017. For each hour of the sample, this is achieved by randomly sampling a forecast error among all post-October 6, 2017, observations sharing the same hour and date. For instance, to sample an error for hour 6:00 of May 8, 2017, we would randomly pick the forecast error among these from 6:00 on May 8 of either 2018, 2019, 2020, or 2021.

Appendix B. Predictive models

B.1. Logistic regression

Due to a large number of categorical features, the dummy variables are aggregated into buckets when applying the logistic regression. The categorical features month and hour encompass, respectively, 12 and 24 categories. Due to the similarity between multiple categories and to reduce the dimensionality of the feature vector, the categories are regrouped into bins: [January, February], [March, April, May], [June, July, August, September], [October, November, December] and [23:00 to 5:00], [6:00 to 10:00], [11:00 to 13:00], [14:00 to 16:00], [17:00 to 19:00], [20:00 to 22:00].

Another transformation is applied to the *load/grid* feature. The relation between the target variable (spikes) and the *load/grid* feature is non-linear. To capture the non-linearity with the logistic regression model, a feature corresponding to the squared value of *load/grid* is added to the feature set.

Table 7
DNN depth and size for the various training periods and threshold choices.

Threshold/Period	2017–2018	2017–2019	2017–2020	2017–2021
$\gamma^- = -30\$$	25, 15, 10	45, 20, 15	45, 20, 15	15, 10, 5
$\gamma^- = -45\$$	50	45, 20, 15	100	20, 15, 10
$\gamma^- = -60\$$	50	45, 20, 15	50	25, 15, 10

The table cells display the DNN depth and size selected by grid search for each threshold and training period. Each table cell reports the number of neurons for each respective layer of the DNN, i.e., the ith number indicates the number of neurons for the ith layer.

B.2. Model estimation and hyperparameter tuning

The random forests, gradient-boosting trees, and DNNs all encompass a set of hyperparameters. The choice of hyperparameters is largely related to the model's performance. Unfortunately, hyperparameters cannot be optimized with the regular model parameters during the training step.

Several search methods, such as random search, grid search, or Bayesian optimization, can be used to search over many sets of hyperparameters. Furthermore, the search process must be paired with a performance evaluation technique to discriminate between the sets of hyperparameters tested. The current study implements a grid search algorithm paired with a k-fold cross-validation process. The grid search method takes as input a grid of hyperparameters. The algorithm trains the model and computes the performance for each possible combination in the grid. The k-fold cross-validation is a method that makes it possible to compute each set of hyperparameters' performance objectively. The k-fold cross-validation starts by splitting the dataset into k folds. The algorithm then lists the k possible combinations of the $k-1$ folds while keeping the remaining fold for a performance review. The model is then trained for each of the k combinations, and the performance of the model is assessed over the remaining fold. The overall model performance is computed by aggregating the testing fold results for all k combinations. The hyperparameter set with the greatest performance is selected. It is important to note that the size of the grids and the number of folds impact the numerical load. In this study, five-fold cross-validation is applied.

For the DNN model, a fully connected feed-forward neural network architecture is considered. The number of layers and neurons per layer are treated as hyperparameters. This implies the number of layers and their respective size vary over the various training periods and threshold choices. Table 7 shows the hyperparameters selected through grid search for each considered threshold and training period.[11]

Appendix C. Model confidence set approach

In Sections 3 and 4, the model confidence set approach of Hansen et al. (2011) is implemented to identify the best performing model(s) using the log-likelihood and the trading P&L as the performance measures. More precisely, the Hansen et al. (2011) method selects the best-performing model(s) through an iterative process based on a two-step approach, with steps respectively being called the equivalence test and the elimination step. The equivalence test verifies whether

[11] The DNN model is built with the `Scikit-learn` package in `Python`.

all models in the model confidence set generate performances that are statistically indistinguishable. If such null hypothesis from the equivalence test is rejected, the elimination step identifies which model is to be removed from the confidence set.

Since the number of considered models is low compared to the number of out-of-sample observations, Hansen et al. (2011) indicates that the statistic of the equivalence test can be computed as follows. The forecast performance at time t for each model k in the model confidence set is represented by $L_{t,k}$. Such quantities are collected in the matrix $L = [L_{t,k}]$ where t refers to the row and k to the column. L_t refers to row t of L. The total number of considered models is K. In the current study, X_t is set to $X_t = L_t M$, where the M is a $K \times (K-1)$ matrix such that $M_{i,j} = 1$ when $i = j$, $M_{i,j} = -1$ when $i - 1 = j$ and $M_{i,j} = 0$ otherwise. Under the null hypothesis, X_t has a zero mean.

The test statistic is

$$T = n\bar{X}'\hat{\Sigma}^{-1}\bar{X},$$

where $\bar{X}$ is the arithmetic average of $X_1, \ldots, X_n$, $\hat{\Sigma}$ is a consistent estimator of the covariance matrix of $\bar{X}$, and n is the number of hourly time periods. Under the null hypothesis, $T \to \chi^2(q)$, where $q = \text{rank}(\Sigma)$. In this study, the confidence level considered is 5%. If the equivalence test is rejected, the model with the largest standardized excess loss is removed from the model confidence set. The standardized excess loss for model j is computed as follows:

$$t_k = \frac{\bar{p}_k}{\sqrt{\widehat{\text{var}}(\bar{p}_k)}},$$

where $p_{t,k} = L_{t,k} - \frac{1}{K}\sum_{i=1}^{K} L_{t,i}$ and $\bar{p}_k = \frac{1}{n}\sum_{t=1}^{n} p_{t,k}$. The process iterates through steps 1 and 2 until either (i) the equivalence test is not rejected or (ii) only one model remains inside the model confidence set.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eneco.2023.106521.

References

Amjady, N., Keynia, F., 2010. Electricity market price spike analysis by a hybrid data model and feature selection technique. Electr. Power Syst. Res. 80 (3), 318–327.

Anderson, C., Davison, M., 2008. A hybrid system-econometric model for electricity spot prices: Considering spike sensitivity to forced outage distributions. IEEE Trans. Power Syst. 23 (3), 927–937.

Benth, F.E., Kallsen, J., Meyer-Brandis, T., 2007. A non-Gaussian Ornstein–Uhlenbeck process for electricity spot price modeling and derivatives pricing. Appl. Math. Finance 14 (2), 153–169.

Cartea, A., Figueroa, M.G., 2005. Pricing in electricity markets: a mean reverting jump diffusion model with seasonality. Appl. Math. Finance 12 (4), 313–335.

Christensen, T., Hurn, S., Lindsay, K., 2009. It never rains but it pours: modeling the persistence of spikes in electricity prices. Energy J. 30 (1).

Christensen, T.M., Hurn, A.S., Lindsay, K.A., 2012. Forecasting spikes in electricity prices. Int. J. Forecast. 28 (2), 400–411.

Clements, A., Herrera, R., Hurn, A., 2015. Modelling interregional links in electricity price spikes. Energy Econ. 51, 383–393.

Conejo, A.J., Contreras, J., Espínola, R., Plazas, M.A., 2005. Forecasting electricity prices for a day-ahead pool-based electric energy market. Int. J. Forecast. 21 (3), 435–462.

Covert, I., Lundberg, S.M., Lee, S.-I., 2020. Understanding global feature contributions with additive importance measures. Adv. Neural Inf. Process. Syst. 33, 17212–17223.

Eichler, M., Grothe, O., Manner, H., Tuerk, D., 2014. Models for short-term forecasting of spike occurrences in Australian electricity markets: A comparative study. J. Energy Mark. 7 (1).

Fan, J.-L., Hu, J.-W., Zhang, X., 2019. Impacts of climate change on electricity demand in China: An empirical estimation based on panel data. Energy 170, 880–888.

Gudkov, N., Ignatieva, K., 2021. Electricity price modelling with stochastic volatility and jumps: An empirical investigation. Energy Econ. 98, 105260.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

He, D., Chen, W.-P., 2016. A real-time electricity price forecasting based on the spike clustering analysis. In: 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D). IEEE, pp. 1–5.

Herrera, R., González, N., 2014. The modeling and forecasting of extreme events in electricity spot markets. Int. J. Forecast. 30 (3), 477–490.

Itron, 2008. New York ISO climate change impact study. Technical report, New York Independent System Operator.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. vol. 112, Springer.

Janczura, J., Trück, S., Weron, R., Wolff, R.C., 2013. Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. Energy Econ. 38, 96–110.

Klüppelberg, C., Meyer-Brandis, T., Schmidt, A., 2010. Electricity spot price modelling with a view towards extreme spike risk. Quant. Finance 10 (9), 963–974.

Lago, J., De Ridder, F., De Schutter, B., 2018. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Appl. Energy 221, 386–405.

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

Longstaff, F.A., Wang, A.W., 2004. Electricity forward prices: a high-frequency empirical analysis. J. Finance 59 (4), 1877–1900.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Manner, H., Türk, D., Eichler, M., 2016. Modeling and forecasting multivariate electricity price spikes. Energy Econ. 60, 255–265.

NYISO, 2016. NYISO standard template presentation to market participants. Technical report, New York Independent System Operator.

Sandhu, H.S., Fang, L., Guan, L., 2016. Forecasting day-ahead price spikes for the Ontario electricity market. Electr. Power Syst. Res. 141, 450–459.

Shapley, L.S., 2016. 17. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games (AM-28). vol. II, Princeton University Press, pp. 307–318.

Trueck, S., Weron, R., Wolff, R., 2007. Outlier treatment and robust approaches for modeling electricity spot prices.

Weron, R., 2007. Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. vol. 403, John Wiley & Sons.

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

Yi-Ling, H., Hai-Zhen, M., Guang-Tao, D., Jun, S., 2014. Influences of urban temperature on the electricity consumption of shanghai. Adv. Clim. Change Res. 5 (2), 74–80.

Zahedi, G., Azizi, S., Bahadori, A., Elkamel, A., Alwi, S.R.W., 2013. Electricity demand estimation using an adaptive neuro-fuzzy network: A case study from the Ontario province–Canada. Energy 49, 323–328.

Zhang, J., Tan, Z., Yang, S., 2012. Day-ahead electricity price forecasting by a new hybrid method. Comput. Ind. Eng. 63 (3), 695–701.

Ziel, F., Steinert, R., Husmann, S., 2015. Forecasting day ahead electricity spot prices: The impact of the EXAA to other European electricity markets. Energy Econ. 51, 430–444.