This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

Interpretable Transformer Model for Capturing Regime Switching Effects of Real-Time Electricity Prices

Jérémie Bottieau, *Student Member, IEEE,* Yi Wang, *Member, IEEE,* Zacharie De Grève, *Member, IEEE,* François Vallée, *Member, IEEE* and Jean-François Toubeau, *Member, IEEE,*

***Abstract*—Real-time electricity prices are economic signals incentivizing market players to support real-time system balancing. These price signals typically switch between low- and high-price regimes depending on whether the power system is in surplus or shortage of generation, which is hard to capture. In this context, we propose a new Transformer-based model to assist the short-term trading strategies of market players. The proposed model offers high-performance probabilistic forecasts of real-time prices while providing insights into its inner decision-making process. Transformers rely on attention mechanisms solely computed via feed-forward networks to explicitly learn temporal patterns, which allows them to capture complex dependencies such as regime switching. Here, we augment Transformers with subnetworks dedicated to endogenously quantify the relative importance of each input feature. Hence, the resulting forecaster intrinsically provides the temporal attribution of each input feature, which foster trust and transparency for subsequent decision makers. Our case study on real-world market data of the Belgian power system demonstrates the ability of the proposed model to outperform state-of-the-art forecasting techniques, while shedding light on its most important drivers.**

***Index Terms*—Attention mechanism, deep learning, imbalance price, explainable AI, multi-horizon forecasting, real-time electricity markets.**

I. Introduction

PROVIDING correct price signals to market participants is fundamental in modern competitive electricity markets for achieving an affordable, reliable, and sustainable electric power system [1]. Electricity Price Forecasting (abbreviated EPF hereafter) contributes to that objective by supporting the trading strategies of the various system actors, such as the optimal bidding of wind power [2], the optimal self-scheduling of generation companies [3], or arbitraging opportunities for energy storage systems [4], [5]. Overall, electricity prices in the day-ahead market have been well studied by the EPF community, with efficient approaches ranging from econometric to machine learning methods – see for instance [6], [7] and references therein. However, the increasing share of intermittent, renewable-based energy sources in power systems tends to increase the close-to-real-time balancing needs [8], so that the real-time trading of electricity, and as a consequence, the forecasting of real-time electricity prices, are currently gaining strong momentum among the power systems community [9], [10].

The work is supported via the energy transition funds project 'EPOC 2030-2050' organized by the FPS economy, S.M.E.s, Self-employed and Energy.

The trading of real-time electricity differs according to the market implementation [11], [12]. In US-styled pools, the real-time electricity prices are defined using a locational marginal pricing market, wherein energy deviations and operating reserves are settled at a unique price for each electrical node. While in European markets, real-time electricity prices are defined according to a zonal model and may refer to either imbalance or balancing prices, which arises from the intrinsic segmentation between energy and balancing markets. Balancing prices remunerate the "Balancing Service Providers" for the actual activation of balancing reserves (e.g., the automatic frequency restoration reserve), whereas imbalance prices penalize any real-time energy deviations of "Balance Responsible Parties" (abbreviated BRPs hereafter) from their position in energy markets. Both prices are connected since imbalance tariffs are based on the activation fees in the balancing stage. Regardless of the market design, accurate estimations of real-time electricity prices in a multi-horizon setting are crucial for market players for either optimally exploiting real-time arbitrage opportunities or reducing their exposure to imbalance costs [13], [14]. The challenging nature of this task is exacerbated by two fundamental causes: i) the signal exhibits a regime-switching behavior, where it flips from low- and high-price regimes depending on whether the power system is in surplus or shortage of generation [15], and ii) price spikes occur more frequently due to the market's small size and vulnerability to unexpected changes in operating conditions, e.g., outages or congestion of transmission lines [16]. As both characteristics are common across the real-time price signals, our case study considers the multi-horizon prediction of imbalance prices in European markets without loss of generality; see more details in Section II-A.

Despite these two challenges, the literature is still scarce concerning the prediction of real-time electricity prices compared to their day-ahead counterparts. Markov regime-switching models are considered in references [15], [17], [18] for capturing the real-time prices. Such models nest several linear forecasters, each representing a specific regime of the real-time electricity prices, for which transition probabilities are computed based on a latent state variable (e.g., a lagged observation of real-time electricity price or system imbalance). Although limited by linear dependencies, the coefficients of the forecasters can be used for attributing an importance value for each corresponding input feature [15]. Besides, these models are also limited by the well-known Markov property, which

states that the expected future regime state of the process only depends on the current observation of the state variable. More specifically, Olsson and Söder present a Markov-switching seasonal auto-regressive moving average model in [17], while they investigate the introduction of exogenous variables using non-linear time series models in [19]. In the same vein, Dimoulkas et al. apply a hidden Markov model for modeling Nordic balancing prices [18], while Bunn et al. analyze the predictability of British balancing prices using Markov switching dynamic regression models [15]. Following a similar reasoning, a seasonal auto-regressive moving average model based on the activated balancing volume is proposed in [20], and a Holt-Winters model conditioned by the sign of the net imbalance volume is developed in [21]. The importance of embedding balancing state information (e.g., lagged volumes of activated balancing energy) in the forecasting models tends to be confirmed by the benchmark analysis in [22], where models without such information provide larger interval forecasts. This observation is further highlighted in [23], which shows that the net system imbalance volume has the highest explanatory power when used with tree-based ensemble methods. Tree-based ensemble methods are powerful forecasters, but their extension towards multi-horizon forecasting commonly requires a novel model at each prediction step. Such a strategy prevents the learning of time dependencies between outputs, which may consequently produce completely unrelated forecasts over the prediction horizon [24], [25]. In complement, references [26], [27] focus solely on the net system imbalance volume, which is then used to compute imbalance prices based on merit order estimates of the balancing energy market. This approach has recently gained interest since Transmission System Operators (abbreviated TSOs hereafter) have made these merit order estimates publicly available within the trend of improving market transparency [28]. However, although the forecaster is relieved from capturing the complex regime-switching behavior of the real-time electricity prices, the simplifying market hypotheses adopted for constructing the merit order estimates are inevitably limiting the accuracy of the predicted real-time prices.

Overall, the literature shows that capturing both the price-regime switching behavior and spikes of real-time electricity prices is a non-trivial task [29], [30]. Furthermore, even an accurate predictive model may face barriers in terms of acceptability among the users' community, especially if it behaves as a black box and generates non-interpretable outcomes. Neural networks are particularly prone to that phenomenon, where the underlying reasoning is more complex to extract than simpler, readily interpretable models – see e.g., [31]. In this line, combining the predictive power of deep neural models with interpretable features has attracted a high-level interest within the machine learning community [32], [33]. Following [34], interpretability can be defined as the ability of a model to provide explanations in understandable terms to a human. The scope can be global, i.e., interpreting the average behavior of the model over the whole dataset, or local, i.e., explaining a case-specific outcome of the model. Interpretability can be integrated via two approaches: i) a post-hoc approach, which consist in analyzing an *already trained* (black-box) model by, e.g., interpretable local surrogates, gradient-based or perturbation-based methods [35], and ii) an intrinsic approach, in which the architecture of the model is directly designed with interpretable components [36], [37]. For instance, the tree-based ensemble methods used in [23] are able to provide the global importance of each input feature, but they need to be complemented with post-hoc methods for providing local interpretations [38]. However, such post-hoc methods are limited when using on multivariate time series as they do not consider the temporal dependencies between input features [39]. On the other hand, an interpretable attention-based recurrent neural architecture is used in [40] for forecasting the grid imbalances. Although the model has proven successful in a multi-horizon setting, its interpretability does not allow to capture the relative importance between between past observed and future known inputs. In this direction, this work also develops a neural network model with intrinsic interpretable components, but with the ability to identify this key interaction between both past and future horizons. The interpretability aspect of the proposed model can be analysed both globally and locally. The global analysis allows visualizing the most influential input features and the most persistent temporal patterns, while the local analysis shows the behavior of the model for case-specific outcomes such as during a regime-switching event.

The proposed model is based on the Transformer neural architecture, which shows an improved capture of long-range dependencies between the elements of an input sequence. Hence, Transformers tend to become the novel state-of-the-art neural model in various tasks such as natural language processing applications [41], [42]. By relying on attention mechanisms solely computed via feed-forward neural networks, our model is able to capture distinct temporal patterns of the input signal depending on the predicted price regime. In addition, the model is augmented with subnetworks able to provide direct insights on the relative importance of each individual input feature [37]. Finally, as deep learning models are known to be difficult to optimize and require careful tuning of hyper-parameters, we leverage a normalization block to improve the convergence and performance of the proposed model [43]. The contributions of this paper are summarized below:

- We present for the first time a Transformer-based model for the multi-horizon probabilistic forecasting of real-time electricity prices. More particularly, the model is trained to predict a set of price quantiles, which are free of any distributional assumption. We observe that the Transformer-based attention mechanism enables the learning of distinct temporal patterns of the input signal depending on the predicted price regime. The effectiveness of our approach is illustrated in a detailed case study using data from a real-life power system.
- The proposed model is able to provide a global and local interpretability analysis of the temporal attributions of each input feature. To achieve that, the architecture of the Transformer-based model is enriched with dedicated sub-networks, which aim at computing the relative importance between input features at each time step. The interpretability

of our model serve a dual purpose: i) selecting endogenously the most informative features without resorting to separate data preprocessing steps, and ii) providing direct insights to the user on which are the most important forecasting drivers and how they are used temporally.
- Each non-linear transformation performed by the proposed model is used in conjunction with a normalization block, which reduces the vanishing gradient problem and improves the information flow in deep neural models. This block is composed of a layer normalization, a gated linear unit and a residual connection. Results demonstrate that this added block plays an important role in the improved performance of the proposed model.

The remaining parts of the paper are organized as follows. The real-time market framework and the forecasting model are described in Section II. In Section III, performance measures and benchmark methods are presented. The case study and the evaluation of the proposed forecasting strategy, both in terms of performance and interpretability perspectives, are discussed in Section IV. Finally, Section V concludes the paper.

II. Methodology

In this section, the targeted real-time market framework is firstly introduced. Then, we present an overview of the proposed forecasting model, where the most dominant layers of the neural architecture are depicted. Last, each neural network layer and its associated functionality are detailed.

A. Real-Time Market Framework

We consider the favoured European real-time market segment dedicated to monetizing deviations from the energy markets, i.e., the single price imbalance settlement. In this market framework, the BRPs endorse the financial responsibility of their real-time energy deviations, which are computed based on their positions in energy markets. Hence, a unique imbalance price is applied to all positive and negative imbalance positions of BRPs at each imbalance settlement period (typically of 15 minutes duration), which reflects the operational balancing costs. More particularly, depending on the net system imbalance state, the single imbalance price is calculated whether as a function of the upward or downward real-time balancing prices. This regime-switching behavior is illustrated in Fig. 1 considering the Belgian power system. The two regimes of the single imbalance price are considered: i) a high-price regime (typically higher than the day-ahead electricity prices), which may describe a shortage of production at the system level, and ii) a low-price regime (generally lower than the day-ahead electricity prices), characterizing an excess of production. In the rest of this paper, the term "real-time prices" refers to imbalance prices.

B. Forecasting Model

The model is designed for generating multi-horizon probabilistic forecasts of the real-time price λ^{RT} for each Imbalance Settlement Period (abbreviated ISP hereafter) :

$$p\left(\lambda^{\text{RT}}_{t_0+1},...,\lambda^{\text{RT}}_{t_0+\tau_{max}}|\mathbf{x}^h_{t_0-l_{\max}},...,\mathbf{x}^h_{t_0},\mathbf{x}^f_{t_0+1},...,\mathbf{x}^f_{t_0+\tau_{max}}\right) \quad (1)$$

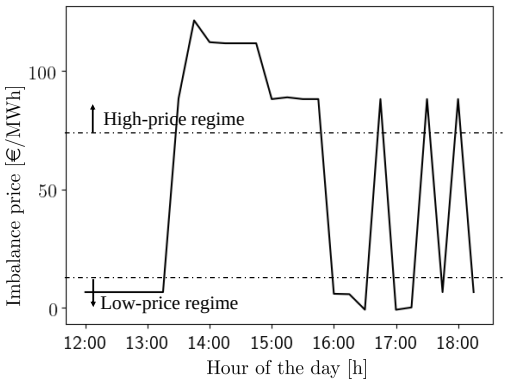


Fig. 1. The regime-switching behavior of the Belgian imbalance price on the 1st January 2019.

where t_0 is the forecast creation time, $l_{\max}, \tau_{\max}$ are respectively indices determining the number of look-back and look-ahead ISPs, $\mathbf{x}^h_{\cdot} \in \mathbb{R}^{m_h}$ are time series observed, and $\mathbf{x}^f_{\cdot} \in \mathbb{R}^{m_f}$ are future information, e.g., the prices cleared at the day-ahead stage or calendar information, already known over the prediction horizon.

Many architectural variations of neural models were developed to process efficiently such a set of heterogeneous inputs [44], [45]. Three major trends can be identified: i) deeper architectures, when adequately designed, increase the ability of the network to extract meaningful representation from the raw data, ii) convolutional neural networks or recurrent neural networks – e.g., the Long Short-Term Memory (abbreviated LSTM hereafter) – are efficient in learning local spatio-temporal relationships, and iii) attention mechanisms, which grant the model direct access to information on specific time steps, enable an improved representation of long-term dependencies.

In light of these recent advances, we propose a Transformer-based model, which pursues high-quality probabilistic predictions of real-time electricity price, while attaining interpretable insights. The overall model is depicted in Fig. 2. Note that layers in the same color share the same weights. In addition, the notations FF-NL and FF-L stand for feed-forward networks using respectively non-linear and linear activation functions, while BLSTM refers to a Bi-directional LSTM network. At the early stage, the m_h look-back observed inputs and the m_f look-ahead known inputs are respectively processed by two distinct variable selection subnetworks, which act as an interpretable filter that allows the model to disregard any irrelevant inputs (Subsection II-D). The selected inputs are then handled by BLSTMs, where both backward and forward time correlations are locally captured (Subsection II-E), followed by a FF-NL that computes an additional non-linear mapping if required. For each time step of the prediction horizon, the Transformer-based attention layer selectively identifies the most salient past and future contextual information over the conditioning range $[t_0 - l_{\max},\ t_0 + \tau_{\max}]$ in a single vector representation (Subsection II-F). Finally, based on this condensed representation, a direct multi-horizon strategy is applied, which consists in outputting in one pass the real-time price's q-quantiles $\{\hat{\lambda}^{\text{RT}}_{t_0+\tau,q}, \forall q \in Q\}$ through two successive non-linear and linear mappings. This strategy avoids error

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

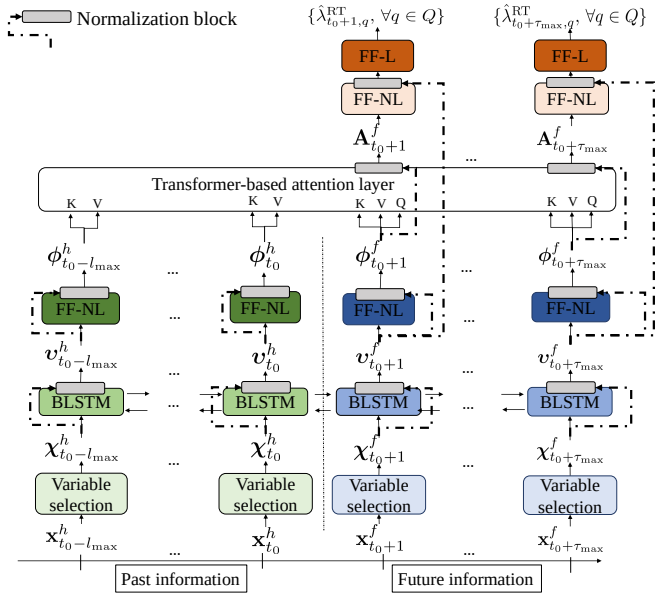


Fig. 2. The Transformer-based model.

accumulation (which is common in fully recurrent models) by alleviating the need of recursively feeding the previously predicted target, while fully making use of the parallel abilities of hardware such as GPUs. In addition, throughout the model, we repeatedly used normalization blocks (Subsection II-G) to control the depth of the model and facilitating its training. All the layers are trained in an end-to-end fashion, i.e., all layers are jointly trained, which guarantees the consistency of the framework.

C. Feed-Forward Networks

Feed-forward networks are used for either transforming a n-dimensional input vector into a d-dimensional vector or applying additional linear and non-linear mappings.

Let $\mathbf{x}^{\text{in}} \in \mathbb{R}^n$ be the input vector. The linear mapping of an FF-L layer is defined as:

$$\mathbf{x}^{\text{out}} = \mathbf{x}^{\text{in}}\mathbf{W}_1 + \mathbf{b}_1 \tag{2}$$

where $\mathbf{x}^{\text{out}} \in \mathbb{R}^d$ is the d-dimensional output vector, and $\mathbf{W}_1 \in \mathbb{R}^{n\times d}$ and $\mathbf{b}_1 \in \mathbb{R}^d$ are parameters to be trained.

An FF-NL layer consists of two linear transformations, with a non-linear activation function in between:

$$\mathbf{x}^{\text{out}} = f^{\text{elu}}(\mathbf{x}^{\text{in}}\mathbf{W}_2 + \mathbf{b}_2)\mathbf{W}_3 + \mathbf{b}_3 \tag{3}$$

where $\mathbf{W}_2 \in \mathbb{R}^{n\times d}$, $\mathbf{W}_3 \in \mathbb{R}^{d\times d}$, $\mathbf{b}_2 \in \mathbb{R}^d$ and $\mathbf{b}_3 \in \mathbb{R}^d$ are parameters to be trained, and f^{elu} is the Exponential Linear Unit activation function, which acts as an identity function for positive values and gets saturated for negative ones [46].

D. Variable Selection Layer

TSOs have the duty to publish a wide range of information for promoting a transparent and non-discriminatory market such as actual measurements (e.g., electrical load and power production, which are here denoted by the superscript h), day-ahead forecasts of renewable generation and electrical load (denoted here by the superscript f). Additional information, such as the schedules of conventional generation and merit

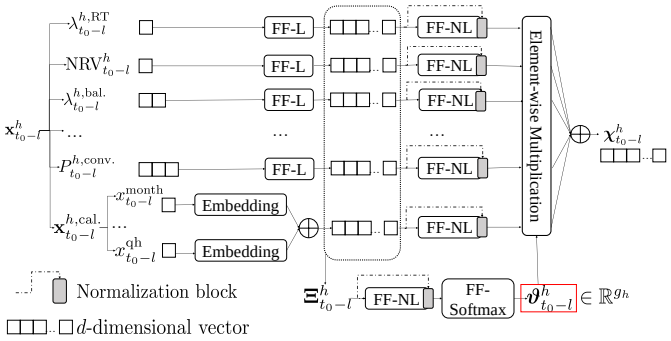


Fig. 3. Variable selection layer for the time step $t_0 - l$.

order estimates of operational balancing prices, may also be provided (also denoted by the superscript f).

In the Belgian power system, besides calendar information, we have at our disposal $m_h = 14$ historical covariates $\mathbf{x}^h_{:t_0+1}$ and $m_f = 15$ known future information $\mathbf{x}^f_{t_0+1:}$. These inputs are gathered in $g_h = 8$ and $g_f = 6$ groups as followed:

- the imbalance price ($\lambda^{h,\text{RT}} \in \mathbb{R}^1$).
- the net activated volume of balancing reserves ($\text{NRV}^h \in \mathbb{R}^1$).
- the upward and downward balancing prices ($\lambda^{h,\text{bal.}} \in \mathbb{R}^2$).
- the physical cross-border energy flows with France and Netherlands ($\phi^h \in \mathbb{R}^2$).
- the produced and forecasted wind and photovoltaic powers with their associated installed capacities ($P^{\{h,f\},\text{renew.}} \in \mathbb{R}^4$).
- the produced and scheduled powers of conventional generators ($P^{\{h,f\},\text{conv.}} \in \mathbb{R}^3$), composed of pump-hydro, gas and nuclear units.
- the measured and forecasted electrical load of the grid ($L^{\{h,f\}} \in \mathbb{R}^1$).
- the day-ahead electricity prices ($\lambda^{f,\text{DA}} \in \mathbb{R}^1$).
- the merit order estimates of operational balancing prices, i.e., the TSO expected prices corresponding to different volumes of activated reserves $\{-600, -300, -100, 100, 300, 600\}$ MW ($\lambda^{f,\text{bal.}} \in \mathbb{R}^6$).

Note that the database was already cleaned once by the TSO, which greatly facilitates the replacement of outliers. In this work, a simple linear interpolation scheme is sufficient for replacing them. In addition, to be enclosed in the non-linearity region of the activation functions, the data are then min-max normalized between $[-1, 1]$ before entering the neural model. The calendar information ($\mathbf{x}^{\{h,f\},\text{cal.}} \in \mathbb{R}^6$) are categorical variables characterizing working days, the day of the week, the hour, the quarter hour, the month and the absolute position of the time step. Overall, the set of historical covariates $\mathbf{x}^h_{:t_0+1}$ is composed of $\{\lambda^{h,\text{RT}}, \text{NRV}^h, \lambda^{h,\text{bal.}}, \phi^h, L^h, P^{h,\text{renew.}}, P^{h,\text{conv.}}, \mathbf{x}^{h,\text{cal.}}\}$, while the set of future known information $\mathbf{x}^f_{t_0+1:}$ contains $\{L^f, P^{f,\text{renew.}}, P^{f,\text{conv.}}, \lambda^{f,\text{DA}}, \lambda^{f,\text{bal.}}, \mathbf{x}^{f,\text{cal.}}\}$.

The level of relevance of the input variables for predicting a target can be hardly anticipated. Hence, we train dedicated subnetworks, i.e., the variable selection layers, jointly with the model to filter out any irrelevant input. This process is showcased in Fig. 3 for the past observed inputs $\mathbf{x}^h_{t_0-1}$ at time step $t_0 - l$. First, each group within the inputs $\mathbf{x}^h_{t_0-l}$ is mapped into a d-dimensional vector, either linearly for the continuous variables or through entity embeddings for the calendar

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

information [47]. The entity embeddings learn to map each calendar information to numerical features in a d-dimensional space. In contrast to the one-hot encoding methodology, this continuous representation identifies and leverages similarities between time steps. Then, all the embedding vectors are averaged in a unique d-dimensional vector that condenses all the calendar information. The use of a common representation space $\mathbb{R}^d$ throughout the model enables residual connections, which facilitates its training (see Subsection II-G).

The vector $\boldsymbol{\Xi}^h_{t_0-l} \in \mathbb{R}^{g_h \cdot d}$ in Fig. 3 represents the concatenation of all the transformed inputs. Once non-linearly transformed, this vector is used as a basis to compute the feature importance variables $\boldsymbol{\vartheta}^h_{t_0-l}$, framed in red in Fig. 3. They are obtained via a feed-forward network with a softmax function that outputs a vector of g_h-dimension. The softmax function ensures that the values of the output vector sum up to 1 and be positive. The final d-dimensional input $\boldsymbol{\chi}^{\text{h}}_{t_0-l}$ for the time step $t_0 - l$ is then obtained by combining each transformed group of inputs, weighted by their corresponding value in $\boldsymbol{\vartheta}^h_{t_0-l}$. Hence, the elements of the vector $\boldsymbol{\vartheta}^h_{t_0-l}$ yields a probability distribution of the relative importance of each group in $\boldsymbol{\chi}^{\text{h}}_{t_0-l}$, thereby providing interpretable outcomes (see Table IV of Subsection IV-C).

E. Local Temporal Processing Layer

The input sequences $\boldsymbol{\chi}^h, \boldsymbol{\chi}^f$ are then respectively processed by two distinct BLSTM networks, whose internal representations are exchanged at the forecast creation time t_0. The BLSTM is composed of two LSTM networks that process the input sequence in both positive and negative time directions, which allows to capture both forward and backward local time dependencies. Without loss of generality, the output of the BLSTM for the time step $t_0 - l$ is expressed as:

$$\mathbf{h}^{\text{forward}}_{t_0-l} = \mathcal{H}^{h,LSTM}(\boldsymbol{\chi}^h_{t_0-l}, \mathbf{h}^{\text{forward}}_{t_0-l-1}), \tag{4a}$$

$$\mathbf{h}^{\text{backward}}_{t_0-l} = \mathcal{H}^{h,LSTM}(\boldsymbol{\chi}^h_{t_0-l}, \mathbf{h}^{\text{backward}}_{t_0-l+1}) \tag{4b}$$

$$\boldsymbol{\upsilon}^h_{t_0-l} = \frac{\mathbf{h}^{\text{forward}}_{t_0-l} + \mathbf{h}^{\text{backward}}_{t_0-l}}{2} \tag{4c}$$

where $\mathcal{H}^{LSTM}$ is the composite LSTM function [48] and $\{\mathbf{h}^{\text{forward}}_t, \mathbf{h}^{\text{backward}}_t\}$ are the internal states of the LSTMs. The output $\boldsymbol{\upsilon}^h_{t_0-l}$ is an average of both forward and backward internal states for keeping the same d-dimensional representation throughout the model.

Overall, the roles of the BLSTMs are to provide i) an appropriate inductive bias for the time ordering of the input sequence, and ii) awareness of the surrounding elements in the input sequence. Leveraging both time position and local context have proved to be key elements for computing the attention scores in the Transformer-based attention layer [49].

F. Transformer-based Attention Layer

Attention mechanisms are computing layers that provide an abstract representation of an input sequence by dynamically weighting its different time steps. The process is showcased in Fig. 4 for the time step $t_0 + \tau$, where the sequence $\boldsymbol{\phi}^{\{h,f\}} \in \mathbb{R}^{T \times d}$ (with $T = l_{\max} + \tau_{\max}$) is obtained from $\boldsymbol{\upsilon}^{\{h,f\}}$ using FF-NL layers. The sequence $\boldsymbol{\phi}^{\{h,f\}}$ is linearly transformed in three different vectors, i.e., a query $\mathbf{Q}_{t_0+\tau} \in \mathbb{R}^d$, keys $\mathbf{K} \in \mathbb{R}^{T \times d}$ and values $\mathbf{V} \in \mathbb{R}^{T \times d}$, via FF-L layers. The abstract representation $\mathbf{A}_{t_0+\tau} \in \mathbb{R}^d$ is then obtained by weighting the values $\mathbf{V}$ with attention scores $\boldsymbol{\alpha}_{t_0+\tau} \in \mathbb{R}^T$, obtained by quantifying the level of matching between the query $\mathbf{Q}_{t_0+\tau}$ and the keys $\mathbf{K}$:

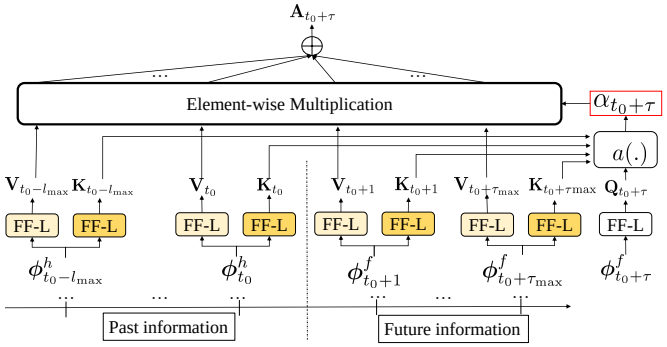


Fig. 4. The Transformer-based attention layer for the time step $t_0 + \tau$.

$$\mathbf{A}_{t_0+\tau} = a(\mathbf{Q}_{t_0+\tau}, \mathbf{K})\mathbf{V} \tag{5}$$

where $a(.)$ is the matching function.

Following [41], we use the scaled dot-product attention as the matching function $a(.)$:

$$a(\mathbf{Q}_{t_0+\tau}, \mathbf{K}) = \text{softmax}\big(\frac{\mathbf{Q}_{t_0+\tau}\mathbf{K}}{\sqrt{d}}\big) \tag{6}$$

The dot-product yields the similarity of vector $\mathbf{Q}_{t_0+\tau}$ with regard to the keys $\mathbf{K}$. Higher values of the dot-product correspond to higher relevance between the given key and the proposed query. The scaling factor $\sqrt{d_k}$ is introduced to reduce the magnitude of the dot-product. Then, the softmax function renders the attention scores $\boldsymbol{\alpha}_{t_0+\tau}$ as a probability distribution over all keys $\mathbf{K}$ with regards to $\mathbf{Q}_{t_0+\tau}$. The magnitude of the attention scores $\boldsymbol{\alpha}_{t_0+\tau}$ provide direct insights on the contributions of each element of the input sequence $\boldsymbol{\phi}^{\{h,f\}}$ to predict the real-time price at $t_0 + \tau$.

The attention mechanism provides two keys benefits: i) the model is able to directly access to the most salient contextual information for each time step of the prediction horizon, and ii) it allows to learn regime-specific temporal dynamics by using distinct attention score patterns for each regime. These two benefits are respectively showcased in Fig. 10 and Fig. 11 of Subsection IV-C.

G. Normalization Block

For controlling the depth of the model and facilitating the backpropagation of gradients, we use a normalization block whenever a non-linear transformation is performed. This is illustrated in Fig. 5 for the time step $t_0 - l$ when using the BLSTM. The normalization block is composed of three components, i.e., a layer normalization [43], a gated linear unit [37] and a residual connection [50].

Layer normalization fixes the mean and variance of the distributions of the inputs at each neural layer, which allows reducing internal covariate shift during training [43]. Practically, for this example, the output vector of the layer normalization is computed as $\boldsymbol{\chi}^{\text{LNorm},h}_{t_0-l} = \gamma \frac{\boldsymbol{\chi}^h_{t_0-l} - \mu}{\sigma} + \beta$, in which μ, σ are the

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

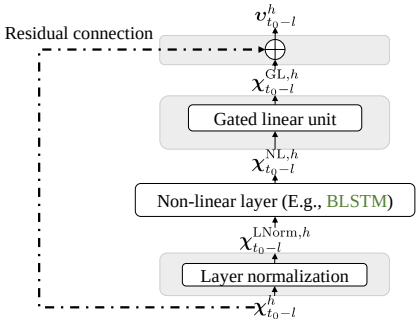


Fig. 5. Illustration of the normalization block applied to the BLSTM for the time step $t_0 - l$.

mean and standard deviation of the elements in $\boldsymbol{\chi}^h_{t_0-l}$, and γ, β are the gain and bias parameters to be trained, respectively.

The gated linear unit allows the model to control the magnitude of the non-linear transformation of the previous layer. The gated linear unit, which takes as input $\boldsymbol{\chi}^{\mathrm{NL},h}_{t_0-l}$, yields:

$$\boldsymbol{\chi}^{\mathrm{GL},h}_{t_0-l} = f^{\sigma}(W^4 \boldsymbol{\chi}^{\mathrm{NL},h}_{t_0-l} + b^4) \odot (W^5 \boldsymbol{\chi}^{\mathrm{NL},h}_{t_0-l} + b^5) \quad (7)$$

where f^{σ} is the sigmoid function, $W^{\{4,5\}} \in \mathbb{R}^{d\times d}$, $b^{\{4,5\}} \in \mathbb{R}^d$ are weights and biases, $\odot$ is the element-wise Hadamard product and d is the dimension of the model. If necessary, the gated linear unit could suppress the non-linear transformation by outputting values all close to 0.

Residual connections allow the model to learn residual functions, which have been proved to be easier to optimize in deeper architecture [50]. The residual connection simply performs an identity mapping, which is added to the output of the gated linear unit (neither extra parameters nor computational complexity is added).

The gain of performance of using the normalization blocks is analyzed in Fig. 8 of Subsection IV-B.

H. Output layer

The simultaneous prediction of the q-quantiles $\hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q}, \forall\tau \in \{1,...,\tau_{\max}\}, \forall q \in Q$, with Q the set of quantiles to predict, are achieved by a FF-L layer at each time step. To produce these quantiles, the model is trained using the smooth approximation of the pinball loss [27], where the Huber norm $H(.)$ is introduced for differentiability issues at the origin [51]. The loss function is computed as:

$$E^H_{t_0+\tau} = \sum_{q\in Q} \begin{cases} q \cdot H(\lambda^{\mathrm{RT}}_{t_0+\tau}, \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q}) & \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q} < \lambda^{\mathrm{RT}}_{t_0+\tau} \\ (1-q) \cdot H(\lambda^{\mathrm{RT}}_{t_0+\tau}, \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q}) & \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q} \geq \lambda^{\mathrm{RT}}_{t_0+\tau} \end{cases} \quad (8)$$

Quantile crossing issue may arise when fitting separately different quantiles. In this paper, we conduct naive rearrangement of the predicted q-quantiles in ex-post, i.e., we sort in ascending order the q-quantiles at each time step of the prediction horizon after they are predicted [52]. Note that this procedure is also performed for the benchmark.

In this work, a mini-batch mode is preferred for training the model, which consists in updating the weight and bias parameters based on the loss function of subsets of samples (i.e., 96 samples representing a daily sequence in our case study), thereby providing a compromise between the batch and online learning modes. The gradient descent procedure is carried out with the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ [41], which automatically and individually adapts the learning rate δ for each network parameter in order to escape local optima during the training phase. The upper limit of the learning rate δ varies over the number n of mini-batches, according to the formula:

$$f^{\mathrm{LR}}(.) = \frac{\delta}{\sqrt{d}} \min\left(\frac{1}{\sqrt{n}}, \frac{n}{n^{1.5}_{\mathrm{warmup}}}\right) \quad (9)$$

where d is the dimension of the model, while $\delta = 0.001$ and $n_{\mathrm{warmup}} = 4000$ are hyperparameters that determine the highest learning rate achieved and the number of steps to reach it, respectively.

III. Performance Measures and Competing Methods

This section introduces the skill scores for evaluating the probabilistic forecasts and competing methods.

A. Performance Measures

The quality of probabilistic forecasts is dominated by two concepts, i.e., reliability and sharpness. Reliable forecasts ensure that the forecast probabilities are consistent with the observed ones, while sharper forecasts are able to tightly encapsulate the uncertainty around the variable of interest. In this paper, we assess the overall quality of the probabilistic forecasts based on three skill scores.

First, we use the Continuous Ranked Probability Score (abbreviated CRPS hereafter), defined as:

$$E^{\mathrm{CRPS}}_{t_0+\tau} = \int_x \left(F(x) - \theta(x - \lambda^{\mathrm{RT}}_{t_0+\tau})\right)^2 \mathrm{d}x \quad (10)$$

where $F(.)$ is the Cumulative Distribution Function (abbreviated CDF hereafter) defined by the predicted q-quantiles $\hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q}$, and $\theta(.)$ is the Heaviside step function, taking the value 1 for $x \geq \lambda^{\mathrm{RT}}_{t_0+\tau}$ and 0 otherwise.

Eq. (10) is a quadratic measure of the difference between the predicted CDF and the observation, which is null in case of a perfect probabilistic forecast [53]. It measures both reliability and sharpness, and has the same unit than the variable of interest. As we evaluate non-parametric predictive densities, the CRPS score can be obtained using numerical integration [54].

Then, we also use the pinball loss $E^Q_{t_0+\tau}$ weighted across all q-quantiles of interest:

$$E^Q_{t_0+\tau} = \sum_{q\in Q} q \max\left(0, \lambda^{\mathrm{RT}}_{t_0+\tau} - \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q}\right) + (1-q)\max\left(0, \hat{\lambda}^{\mathrm{RT}}_{t_0+\tau,q} - \lambda^{\mathrm{RT}}_{t_0+\tau}\right) \quad (11)$$

where a lower $E^Q_{t_0+\tau}$ score indicates a better probabilistic forecast.

The $E^{\mathrm{CRPS}}_{t_0+\tau}$ and $E^Q_{t_0+\tau}$ scores are complemented with the Winkler score, which quantifies the forecast quality for different prediction intervals. For a prediction interval of $(1-\beta)100\%$, the Winkler score $E^W_{t_0+\tau}$ is defined as:

$$E^W_{t_0+\tau} = \begin{cases} \epsilon_{t_0+\tau}, & L_{t_0+\tau} \leq \lambda^{\mathrm{RT}}_{t_0+\tau} \leq U_{t_0+\tau}, \\ \epsilon_{t_0+\tau} + 2(L_{t_0+\tau} - \lambda^{\mathrm{RT}}_{t_0+\tau})/\beta, & \lambda^{\mathrm{RT}}_{t_0+\tau} < L_{t_0+\tau}, \\ \epsilon_{t_0+\tau} + 2(\lambda^{\mathrm{RT}}_{t_0+\tau} - U_{t_0+\tau})/\beta, & \lambda^{\mathrm{RT}}_{t_0+\tau} > U_{t_0+\tau}, \end{cases} \quad (12)$$

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

where $L_{t_0+\tau} = \hat{\lambda}^{\text{RT}}_{t_0+\tau,\beta/2}$ and $U_{t_0+\tau} = \hat{\lambda}^{\text{RT}}_{t_0+\tau,1-\beta/2}$ are the lower and upper bounds of the prediction interval defined by the confidence level β, and $\epsilon_{t_0+\tau} = U_{t_0+\tau} - L_{t_0+\tau}$ is the interval width.

If a real-time price realization $\lambda^{\text{RT}}_{t_0+\tau}$ is within the predicted interval $[L_{t_0+\tau}, U_{t_0+\tau}]$, the Winkler score $E^{W}_{t_0+\tau}$ is a direct measure of sharpness. Otherwise, a penalty term, whose value depends on the severity of the forecast error, is added for reflecting the deficiency in reliability.

The Winkler score (encompassing both reliability and sharpness aspects) is additionally supported by i) the Prediction Interval Coverage Probability (abbreviated PICP hereafter), which empirically indicates the coverage probability of the prediction intervals [55], and ii) the Prediction Interval Normalized Average Width (abbreviated PINAW hereafter), which measures the width of the prediction intervals [56].

The PICP is calculated through counting the covered real-time price realization $\lambda^{\text{RT}}_{t_0+\tau}$ between the lower $L_{t_0+\tau} = \hat{\lambda}^{\text{RT}}_{t_0+\tau,\beta/2}$ and upper $U_{t_0+\tau} = \hat{\lambda}^{\text{RT}}_{t_0+\tau,1-\beta/2}$ bounds of the prediction interval defined by the confidence level β. The counting of the PICP is computed as follows:

$$E^{\text{PICP}}_{t_0+\tau} = \begin{cases} 1, & \text{if } \lambda^{\text{RT}}_{t_0+\tau} \in [L_{t_0+\tau}, U_{t_0+\tau}] \\ 0, & \text{if } \lambda^{\text{RT}}_{t_0+\tau} \notin [L_{t_0+\tau}, U_{t_0+\tau}] \end{cases} \tag{13}$$

When averaged over the test set, the PICP should be as close as possible to the nominal value $(1-\beta)100\%$ of the associated prediction interval. A lower PICP than the nominal value indicates a deficiency in reliability for the prediction interval, which may lead to ex-post disappointments for the subsequent decision maker. Note that a higher PICP than the nominal value simply notifies that the prediction interval is more reliable than anticipated.

However, an excess of reliability may produce very large prediction intervals, which may be of no use for subsequent decision makers as they convey too many uncertainties. Hence, the sharpness of the prediction interval is also an important aspect for determining its level of informativeness. This aspect can be measured via the PINAW. The PINAW is given by:

$$E^{\text{PINAW}}_{t_0+\tau} = \frac{(U_{t_0+\tau} - L_{t_0+\tau})}{\overline{\lambda}^{\text{RT}} - \underline{\lambda}^{\text{RT}}} \tag{14}$$

where $\overline{\lambda}^{\text{RT}}$, $\underline{\lambda}^{\text{RT}}$ are the maximum and minimum values of real-time prices over the test set, normalizing $E^{\text{PINAW}}_{t_0+\tau}$ in percentage. The closer to 0, the sharper and thus more informative is the prediction interval.

In this paper, the Winkler score, PICP and PINAW are calculated for $\beta = \{0.1, 0.5, 0.9\}$.

B. Competing Methods

The proposed model is compared with a wide range of forecasting techniques. First, two naive methodologies are implemented:

- A step-wise averaging model (Step-Avg), where the real-time price distribution of each forecasting time step is computed based on the average of all past observations corresponding to this specific period of the day.
- A probabilistic generalization of persistence (Pers) based on a random walk model. The forecast assumes a Gaussian distribution where the mean is given by the last available real-time price realization, and the variance is determined by exponential smoothing of previous squared errors [57].

Six state-of-the-art models in time series forecasting are also implemented:

- An Auto-Regressive Moving Average (ARMA) model, for which we compute prediction intervals assuming that the residuals are uncorrelated and normally distributed [58].
- A quantile regression forest (QRF), i.e., a bagging-based ensemble method, in which the outcomes of independent regression trees are used for estimating the conditional distribution [59].
- A gradient boosting regression tree (QGBRT) trained with the quantile loss. New regression trees are sequentially created to predict the residuals of the previously generated ones [60].
- A deep feed-forward neural network (S-FFNN), where several hidden layers are stacked on top of each other. The model is trained using the smooth approximation of the pinball loss.
- The traditional sequence-to-sequence model (Seq2Seq) based on LSTM networks [27]. The model is trained using the smooth approximation of the pinball loss.
- The Bahdanau-based sequence-to-sequence model (B-Seq2Seq) proposed in [40], which is also trained using the smooth approximation of the pinball loss.

It should be noted that the ARMA model is only fed with past imbalance price observations, while the machine learning models, i.e., QRF, QGBRT, S-FFNN, Seq2Seq, and B-Seq2Seq, have access to the same input data as the proposed Transformer-based model. In addition, we conduct a hyperparameter optimization to identify the optimal model complexity of each forecaster. This is achieved through a random search, where the same number of iterations is used across all benchmarks [61]. While this benchmark provides a representative snapshot of currently existing time series forecasting methods, this is by no means all-encompassing. Indeed, for instance, gaussian processes are also used for time series prediction, e.g., in [62], but their scalability for larger datasets is still under research [63]. In addition, hybrid boosting-bagging algorithms for tree-based ensemble methods is also present in the literature [64]–[66], but their development is still limited for probabilistic time series forecasting problems.

We also perform an ablation study, in which we investigate the loss in performance of the proposed model (denoted by Ref) when removing important parts of its architecture.

- Ref-Att is the Ref model without the Transformer-based attention layer.
- Ref-VarSel is the Ref model where the variable selection networks are removed.
- Ref-Bidir is the Red model where the sequential information fed to the attention mechanism is injected via sinusoidal functions of different frequencies [41] instead of the BLSTM networks.

- Ref-NB is the Ref model without the normalization blocks.

IV. Case Studies

We conduct the case study on publicly available data obtained from the website of Elia [28], i.e., the Belgian Transmission System Operator, on an Intel® Core™ i7-3770 CPU @ 3.4 GHz with 16 Gb of RAM. The variable of interest is the Belgian imbalance price λ^{RT}. The thirteen forecasting models are implemented using the scikit-learn package, statsmodels package, and TensorFlow package in Python 3.6. The data spans from 2016-1-1 to 2019-12-31, for a total of four years of data. Specifically, the first three years of data (from 2016-1-1 to 2018-12-31) are used to train and validate the models with a ratio of 85%-15%. Hence, the parameters of the models are updated using 85% of these three years, while the remaining 15% is used for tuning the hyper-parameters (i.e., the model parameters are not updated based on the signal errors of the validation set). Once trained and validated, the forecasting models are then benchmarked using the entire year 2019 (test set) for assessing their probabilistic performance, which consists of approximatively 35,000 novel (unseen) input conditions. Each quarter-hourly step of the database is used as a forecast creation time t_0. A prediction horizon of 4 hours is selected, which corresponds to $\tau_{\text{max}} = 16$ time steps, and we compute the 5th, 15th, 25th, 35th, 45th, 50th, 55th, 65th,75th, 85th, 95th percentiles of the target distribution (i.e., $|Q| = 11$) for each of these time periods.

The final configurations of the probabilistic forecasting methods (along with their search spaces) are:

- ARMA model, which considers 12 lagged values and 2 previous values of past errors, i.e., the autoregressive part $p = 12$ and the moving average part $q = 2$. The search ranges of $\{p, q\}$ are respectively $\{1, 2, 3, 4, 8, 12, 16, 20\}$ and $\{0, 1, 2, 3, 4\}$.
- QRF model, with a population of $N^{RF} = 500$ trees fully extended, and a ratio of maximum features per split of 0.05. The look-back window is set to $l_{\text{max}} = 32$. The search range of the ratio of maximum features considered at each split is $\{0.05, 0.1, 0.2, 0.5, 0.9, 1\}$, while the one of l_{max} is $\{4, 8, 12, 16, 24, 32\}$.
- QGBRT model, with a learning rate of 0.1, a maximum depth of 8 per tree, and a ratio of maximum features per split of 0.2. The number of boosting stages is determined by using early stopping, whose upper limit is fixed at 100. The look-back window is also set to $l_{\text{max}} = 32$. The search ranges of additional hyperparameters, i.e., maximum depth and learning rate, are respectively $\{4, 8, 12\}$ and $\{0.1, 0.01, 0.001\}$.

Concerning neural models, the search ranges of hyper-parameters are: i) the number of processing units by layer, which is included in $\{6, 12, 24, 48\}$, ii) the range of the look-back window, which is contained in $\{4, 8, 12, 16, 24, 32\}$, and iii) the number of hidden layers for the S-FFNN model, which varies between [1,3]. The initial learning rate of the Adam optimizer is set to the default value 10^{-3}. For maximizing the generalization capability of neural models, the early stopping criterion is also adopted. This criterion stops the optimization process when no performance improvement is apparent on the validation set, and selects the final model parameters based on the minimum error of the validation set. The final configurations of the neural models are:

- S-FFNN model, with a look-back window of $l_{\text{max}} = 4$ and 2 hidden layers of 48 processing units.
- Seq2Seq model, with 64 Long Short Term Memory processing units, and a look-back window of $l_{\text{max}} = 24$.
- B-Seq2Seq model, with 32 Long Short Term Memory processing units, and a look-back window of $l_{\text{max}} = 24$.
- Ref model, where the dimension d is set to 12. The look-back window is $l_{\text{max}} = 32$.

TABLE I
Number of parameters, training and inference times of probabilistic forecasting methods.

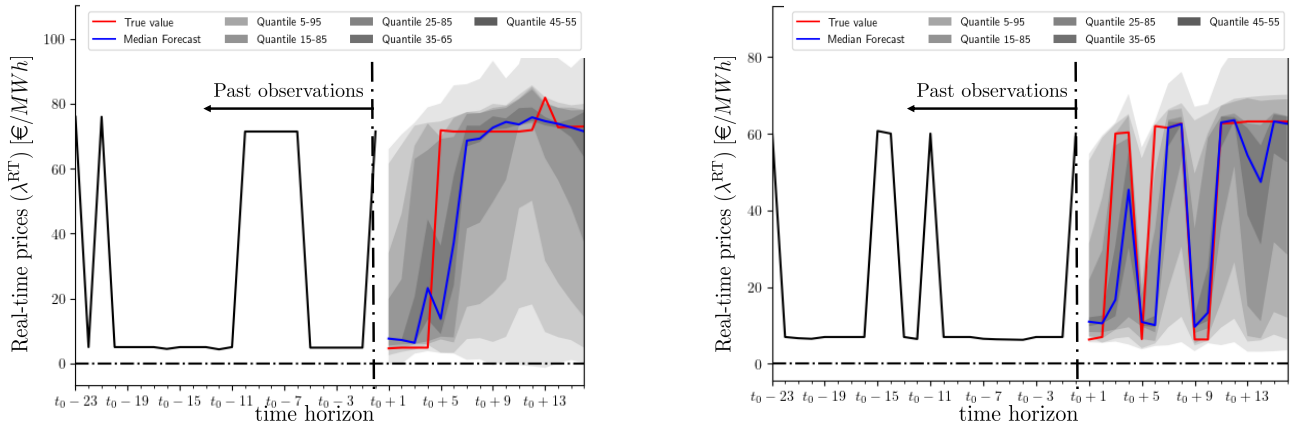
Models	Parameters	Training Time [s]	Inference Time [s]
ARMA	15	120	0.01
QRF	$\tau_{\text{max}} \cdot N^{\text{RF}} \cdot 37k$	$\tau_{\text{max}} \cdot 4200$	0.2
QGBRT	$\tau_{\text{max}} \cdot \lvert Q\rvert \cdot 22k$	$\tau_{\text{max}} \cdot \lvert Q\rvert \cdot 180$	0.4
S-FFNN	$58.5k$	36.4	0.04
Seq2Seq	$95.5k$	725	0.1
B-Seq2Seq	$41k$	1150	0.12
Ref	$63k$	1750	0.15

Concerning the number of parameters, the training and inference times, Table I provides a brief overview for each probabilistic forecasting model. Note that the number of parameters for tree-based ensemble models (i.e., QRF and QGBRT) is given by their number of nodes. Concerning neural models, the S-FFNN model with 2 hidden layers requires less training time than the other neural variants. The Seq2Seq model has the greater number of parameters, which can be explained by its higher number of processing units (i.e., 64) compared to the Ref model (i.e., 12) and the B-Seq2Seq model (i.e., 32). Interestingly, the S-FFNN and Ref models have more parameters than the B-Seq2Seq model. This may come from the nature of the attention mechanisms, where the attention scores from the Ref model are solely computed based on FFNNs, while the ones from B-Seq2Seq model are essentially extracted from hidden states of recurrent neural networks (which is thus less demanding in terms of parameters for long input sequences). For tree-based ensemble models, QRF necessitates a different model per prediction step τ, while QGBRT trains a different model per prediction step τ and forecasted quantile q. As a consequence, this augments the training time of these models for a long forecasting horizon. Overall, the training time of all time series forecasting models still remains manageable with a standard configuration of a computer (with a peak of 18 hours for the τ-QRF models in our experiments). Finally, we can observe that the inference time for generating new predictions is lower than one second for all prediction models, which renders them operational for close-to-real-time purposes.

A. Forecast Evaluation

Fig. 6 illustrates the probabilistic forecasts obtained using the proposed model for the 14th April 2019 at 06H00 and the 14th September at 16H00. It can be observed that the real-time price signal is properly embedded by the predic-

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970



(a) 14th April 2019 at 06:00 (b) 14th September 2019 at 16:00

Fig. 6. Multi-horizon probabilistic forecasts of λ^{RT} on the 14th April 2019 at 06:00 (Fig. 6a) and on the 14th September 2019 at 16:00 (Fig. 6b).

TABLE II
EVOLUTION OF THE CRPS SCORES [€/MWH] OVER THE ENTIRE FORECASTING HORIZON FOR ALL THE MODELS, WHERE TOT. IS THE AGGREGATION OF THE CRPS SCORES.

Models	Tot.	t_0+1	t_0+2	t_0+3	t_0+4	t_0+5	t_0+6	t_0+7	t_0+8	t_0+9	t_0+10	t_0+11	t_0+12	t_0+13	t_0+16
Step-Avg	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2	18.2
Pers	23.2	18.47	20.93	21.61	21.4	22.59	23.43	23.74	23.74	24.24	24.43	24.41	24.34	24.34	24.72
ARMA	20.92	17.17	19.22	19.86	20.24	20.84	21.18	21.35	21.44	21.59	21.62	21.63	21.66	21.67	21.79
QRF	18.67	15.69	17.27	17.75	18.02	18.67	18.92	19.01	19.08	19.16	19.17	19.23	19.27	19.35	19.43
QGBRT	16.82	13.07	15.18	15.86	16.37	16.89	17.24	17.32	17.45	17.5	17.4	17.36	17.47	17.57	17.47
S-FFNN	16.77	14.41	15.64	15.87	16.15	16.63	16.83	17.46	16.93	17.07	16.96	17.21	17.12	17.33	17.74
Seq2Seq	16.81	15.39	16.36	16.71	16.9	16.98	17.05	17.06	17	16.96	16.94	16.91	16.94	16.93	16.98
B-Seq2Seq	16.34	14.7	15.88	16.14	16.25	16.42	16.51	16.53	16.52	16.49	16.48	16.46	16.52	16.54	16.74
Ref	**15.6**	**12.88**	**14.5**	**15.2**	**15.54**	**15.79**	**15.98**	**16.05**	**16.04**	**16.02**	**15.95**	**15.97**	**15.95**	**15.9**	**15.98**

TABLE III
THE PINBALL LOSS [€/MWH] AVERAGED OVER THE ENTIRE PREDICTION HORIZON (TOT.), THE FIRST SIX TIME STEPS $\{t_0+1,...,t_0+6\}$ AND THE LAST TEN TIME STEPS $\{t_0+7,...,t_0+16\}$ OF THE PREDICTION HORIZON.

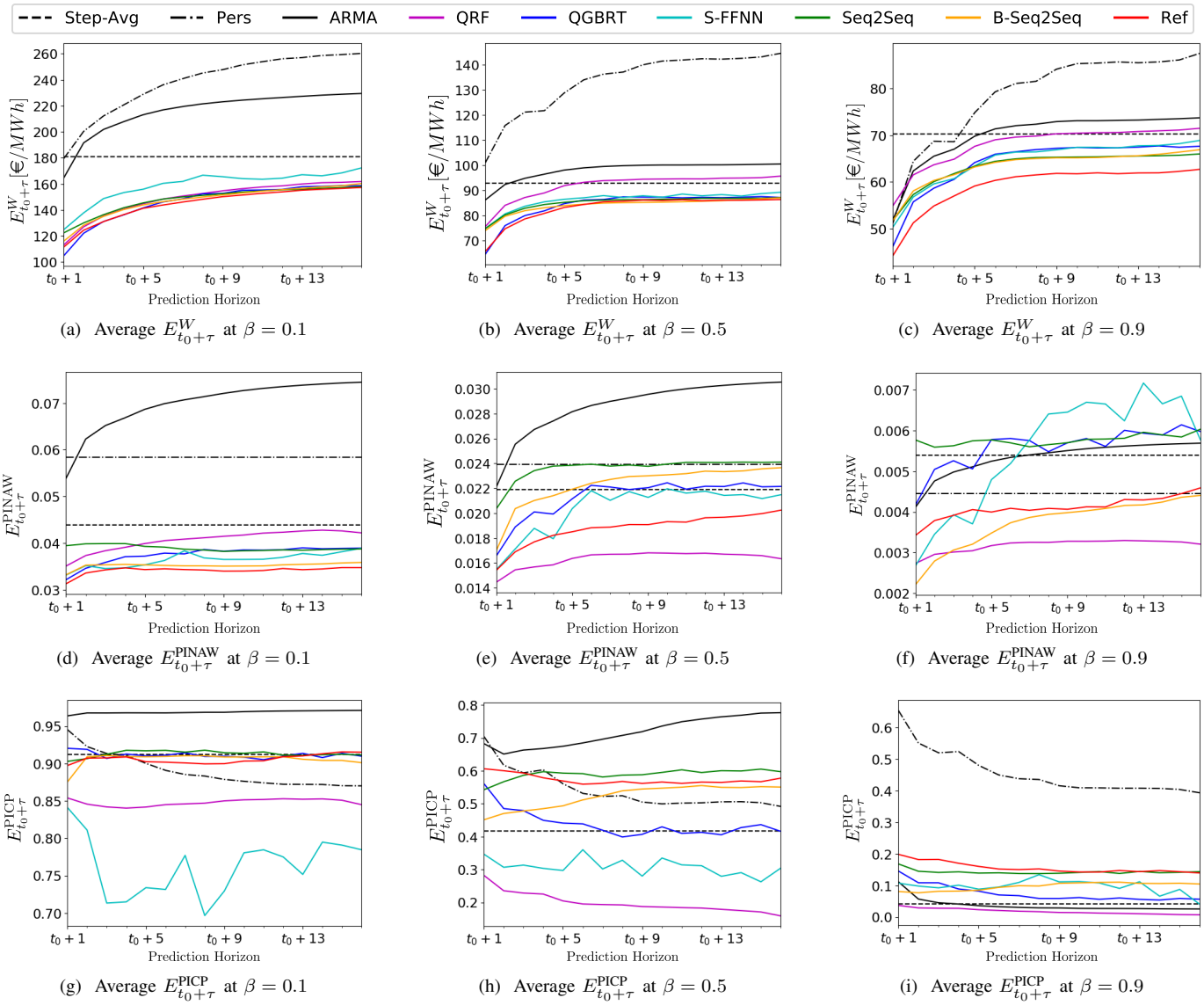
Models	Tot.	$\{t_0+1,...,t_0+6\}$	$\{t_0+7,...,t_0+16\}$
Step-Avg	125.18	125.18	125.18
Pers	158.93	140.74	169.85
ARMA	130.87	123	135.6
QRF	120.27	112.9	124.7
QGBRT	112.62	103.89	117.85
S-FFNN	114.9	107.96	119.07
Seq2Seq	112.41	106.98	115.67
B-Seq2Seq	112	106.4	115.36
Ref	**107.32**	**99.67**	**111.9**

tion intervals. Interestingly, larger prediction intervals encompass both price regimes, while narrower intervals, e.g., $\{\hat{\lambda}^{RT}_{t_0+\tau,0.35},\hat{\lambda}^{RT}_{t_0+\tau,0.65}\}$, attempt to predict the future price regime.

Table II provides the CRPS scores of the different models at each prediction step, which are averaged over the entire test set. The best individual scores are denoted in bold figures. We observe that the proposed model (Ref) provides the lowest CRPS scores over the entire prediction horizon, while the other Machine Learning (ML) methods, i.e., B-Seq2Seq, S-FFNN, Seq2Seq, QGBRT, and QRF, are the second-best models. One reason explaining the gap between ML methods and the other methods is that only ML methods fully leverage all the available input information. This tends to indicate that including exogenous variables in the forecasting models has a positive impact on accuracy. Interestingly, the naive Step-Avg model achieves an overall better performance than the autoregressive models, i.e., the Pers and ARMA models. It is aligned with previous observations [22], [40] where naive forecasts can be hard to beat for real-time market variables. The Pers model has the worst performance within the benchmark. By simply propagating the most recent past realization, the model does not have the ability to infer the most likely future price-regime of the real-time prices. Even if it includes a larger look-back window of past realizations, the ARMA model is unable to perform better than the naive Step-Avg model. Overall, it can also be observed that the CRPS scores for all models (except for the Pers) saturate when the prediction horizon is longer than one hour and a half, i.e., for t_0+6.

Concerning the ML models, we can see that the neural models, i.e., the Ref, B-Seq2Seq, Seq2Seq, and S-FFNN, tend to outperform the tree-based ensemble methods (QRF and QGBRT) over the last time steps $\{t_0+7,...,t_0+16\}$. This can be explained by the fact that different models are defined independently at each prediction step $t_0+\tau$ for tree-based ensemble methods, whereas the parameters of the neural models are shared over the prediction horizon [24]. The S-FFNN model directly generates all the probabilistic predictions, while the final layer of other neural variants is iterated over the entire forecasting horizon. By sharing the parameters in their output layer, the neural models are able to better capture temporal dependencies between outputs. However, it can be observed that the tree-based ensemble methods remain very competitive over the first six time steps and that the QGBRT model performs even better than the B-Seq2Seq over the first three time steps. We also observe that QRF performs worse than QGBRT, which can be explained by the fact that QRF

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970



(a) Average $E^{W}_{t_0+\tau}$ at $\beta = 0.1$ (b) Average $E^{W}_{t_0+\tau}$ at $\beta = 0.5$ (c) Average $E^{W}_{t_0+\tau}$ at $\beta = 0.9$

(d) Average $E^{\text{PINAW}}_{t_0+\tau}$ at $\beta = 0.1$ (e) Average $E^{\text{PINAW}}_{t_0+\tau}$ at $\beta = 0.5$ (f) Average $E^{\text{PINAW}}_{t_0+\tau}$ at $\beta = 0.9$

(g) Average $E^{\text{PICP}}_{t_0+\tau}$ at $\beta = 0.1$ (h) Average $E^{\text{PICP}}_{t_0+\tau}$ at $\beta = 0.5$ (i) Average $E^{\text{PICP}}_{t_0+\tau}$ at $\beta = 0.9$

Fig. 7. Average Winkler score, PINAW and PICP of all models over the test set for $\beta = \{0.1, 0.5, 0.9\}$ at each prediction step.

gives an estimate of the conditional distribution from which quantiles are extracted, whereas the q-quantiles of QGBRT are directly computed through the minimization of the quantile loss. Besides, the S-FFNN model performs worse than the B-SeqSeq model and our proposed model, which shows the importance of aligning the architecture of the neural network with the temporal characteristic inherent to time series forecasting problems. Finally, both intrinsic interpretable neural models, i.e., Ref and B-Seq2Seq, provide the best averaged scores, which suggests that adding interpretable components within their architecture do not hinder their prediction performance.

In addition, Table III provides the pinball loss, which is averaged over i) the entire prediction horizon (Tot.), ii) the first six time steps $\{t_0 + 1, ..., t_0 + 6\}$, and iii) the last ten time steps $\{t_0 + 7, ..., t_0 + 16\}$ of the prediction horizon. Concerning the pinball loss, the observations that are drawn in Table II are even more pronounced. Overall, the Ref model achieves a higher accuracy in each column compared to the other forecasting models. The Seq2Seq and B-Seq2Seq models outperform the QGBRT model over the whole horizon, but QGBRT is the second best model for the first six time steps. In this Table, it can be also observed that the S-FFNN model performs less than the QGBRT, B-Seq2Seq and B-Seq2Seq models.

To complement these results, Fig. 7 showcases the Winkler score, PINAW and PICP of all models for the confidence levels $\beta = \{0.1, 0.5, 0.9\}$ over the entire prediction horizon. The confidence levels $\beta = \{0.1, 0.5, 0.9\}$ correspond to the prediction intervals $\{\hat{\lambda}^{\text{RT}}_{t_0+\tau,0.05}, \hat{\lambda}^{\text{RT}}_{t_0+\tau,0.95}\}$, $\{\hat{\lambda}^{\text{RT}}_{t_0+\tau,0.25}, \hat{\lambda}^{\text{RT}}_{t_0+\tau,0.75}\}$, $\{\hat{\lambda}^{\text{RT}}_{t_0+\tau,0.45}, \hat{\lambda}^{\text{RT}}_{t_0+\tau,0.55}\}$, respectively. The first row concerns the Winkler scores (i.e., encompassing both the sharpness and reliability aspects), the second row indicates the PINAW (i.e., the sharpness aspect) and the last row covers the PICP (i.e., the reliability aspect). For a larger interval at $\beta = 0.1$, the metrics of the ML models are very close to each other and are significantly below the other models' metrics. For such a large interval, the models provide predictions encompassing both the low- and high-price regimes (in a narrower fashion

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970

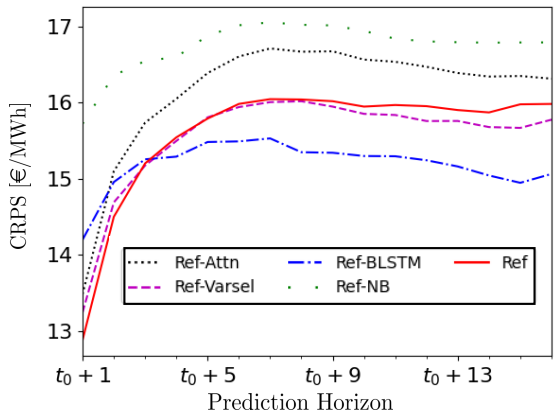


Fig. 8. CRPS scores for the ablation analysis over the entire prediction horizon.

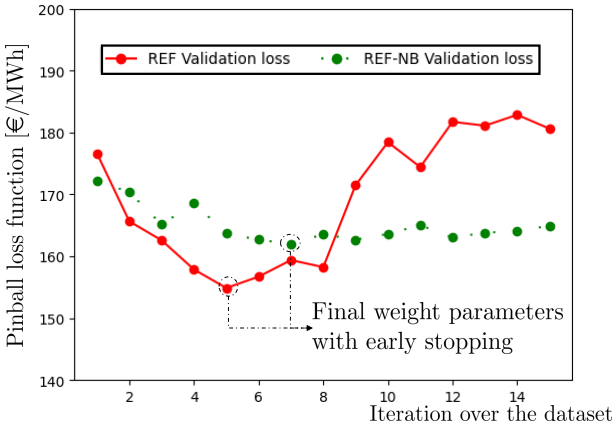


Fig. 9. Validation losses of the Ref model with and without the normalization blocks during the training procedure.

for the ML models), but none of the models are able to differentiate the price regime. Fig. 7d, indicates that the Ref model provides the sharpest interval $\{\hat{\lambda}^{\text{RT}}_{t_0+\tau,0.05}, \hat{\lambda}^{\text{RT}}_{t_0+\tau,0.95}\}$, while Fig. 7g shows that the level of reliability of 90% is met. We can also observe in Fig. 7g that the QRF and S-FFNN models have a deficit of reliability for this prediction interval. Fig. 7b shows the Winkler scores at $\beta = 0.5$. In this case, the metrics of Ref and QGBRT are practically equal, while the B-Seq2Seq and Seq2Seq perform worse over the first time steps. We also observe that the performance of the QRF model starts to deteriorate for narrower prediction intervals. Indeed, the QRF model provides a very sharp prediction interval but with low reliability (purple curves in Fig. 7e and 7h). In contrast, the red curves in Fig. 7e and 7h show that the Ref model manages to produce one of the sharpest interval, while attaining a level of reliability above the nominal value of 50%. Concerning the narrowest interval at $\beta = 0.9$ (Fig. 7c), the Winkler scores are more stratified. The Ref model clearly outperforms all other models, which is highly valuable since the 45th and 55th quantiles provide direct information on the price-regime of the real-time prices. This is highlighted by Fig. 7f and 7i, where the Ref model provides one of the sharpest and most reliable prediction interval. More particularly, Fig. 7f indicates that the Ref and B-Seq2Seq are the two best models in terms of sharpness for this prediction interval. Note that the QRF model provides a very sharp interval but with too low reliability. Fig. 7i indicates that the Ref model provide the most reliable prediction interval, with a level of reliability above the nominal value of 10%. This tends to demonstrate that the proposed model is able to better detect the likely future regime of real-time prices than the other forecasting models.

B. Ablation Analysis

Fig. 8 shows the average CRPS scores over the entire forecasting horizon resulting from the ablation analysis. First, the Ref-NB model achieves the worst performance. This highlights the importance of the normalization blocks, which adapt the depth of our proposed model to the dataset and facilitate the backpropagation of gradients. This is illustrated in Fig. 9, which shows the validation losses of the Ref and Ref-NB models during the training procedure. It can be seen that the Ref model achieves a lower validation loss than the Ref-NB at the time the training is stopped with early stopping. Returning to Fig. 8, the non-attentional model, denoted Ref-Attn, is the second-worst model in terms of accuracy, which highlights the benefits of this alignment procedure that provides direct connections with relevant time steps of the surrounding horizon. This observation tends to reflect the importance of the attention mechanism to capture different temporal patterns for differentiating the different regimes of price. Interestingly, we see that injecting the sequential information with sinusoidal functions of different frequencies (instead of the BLSTM) also worsens the results during the first quarter hours, but provides better forecasts for the remaining prediction horizon. Finally, the metrics of the models Ref-Varsel and Ref are comparable over the entire forecasting horizon. This suggests that the main interest of adding the variable selection layers consists in providing more interpretable outcomes.

C. Interpretability

In this Section, we analyse the temporal attributions of the input features given by our model both globally and locally. Firstly, we quantify the global feature importances by analysing the variable selection weights described in Section II-D. Then, we visualize the persistent temporal patterns based on the attention weights explained in Section II-F. Finally, we illustrate the behavior of our model for a case-specific outcome, where we show the difference in the temporal patterns of the most dominant drivers during a regime-switching event.

1) Global Analysis

The interpretable outcomes of the variable selection layers are shown in Table IV. Practically, $\{\overline{\vartheta}^h, \overline{\vartheta}^f\}$ are selection variables that are aggregated for each feature across the entire test set. Results show that the proposed model extracts only a subset of key inputs (highlighted in bold) that intuitively play a significant role in the predictions. Regarding the past available information, the lagged values of the real-time prices are critical as expected. In addition, the net activated regulation volume and the calendar information also emerge as important drivers for the model. Remarkably, the lagged values of renewable generation bring an additional explanation power

TABLE IV
AVERAGED REPRESENTATION OF VARIABLE SELECTION WEIGHTS OVER BOTH PAST (UPPER TABLE) AND FUTURE (LOWER TABLE) DATA.

	$\lambda^{h,\text{RT}}$	NRV^h	$\lambda^{h,\text{bal.}}$	ϕ^h	L^h	$P^{h,\text{renew.}}$	$P^{h,\text{conv.}}$	$x^{h,\text{cal.}}$
$\overline{\vartheta}^h$	**0.27**	**0.26**	0.05	0.03	0.01	**0.11**	0.05	**0.22**

	L^f	$P^{f,\text{renew.}}$	$P^{f,\text{conv.}}$	$x^{f,\text{cal.}}$	$\lambda^{f,\text{DA}}$	$\lambda^{\text{f,bal.}}$
$\overline{\vartheta}^f$	0.03	0.04	< 0.01	0.02	**0.86**	0.05

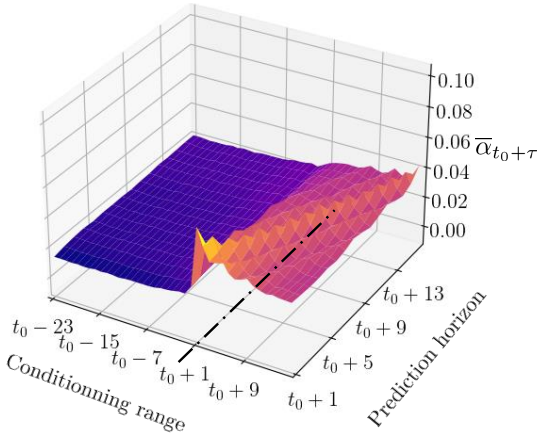


Fig. 10. Averaged temporal attention of the model over the entire prediction horizon for both past and future conditionning ranges

to the model. For the known inputs in $\mathbf{x}^f_{t_0+1:}$, the most dominant driver is the day-ahead electricity price. However, in our experimental set-up, the merit order proxies of operational balancing reserves provided by the TSO, i.e., $\lambda^{\text{f,bal.}}$, play only a minor role in the proposed model.

Next, we analyze persistent temporal patterns, which are often key to understanding the time-dependent relationships between inputs-outputs. To do so, we average the attention weights over the entire test set, which produces the averaged attention patterns $\overline{\alpha}_{t_0+\tau}$ depicted in Fig 10. In this plot, each contour line perpendicular to the 'conditioning range' axis represents the intensity of the model's temporal attention for each time step of the prediction horizon. Over the whole prediction horizon, it can be seen that the model is mostly focused on the time steps between $t_0 - 7$ and $t_0 + 16$. Such outcomes can be expected since the real-time price is a signal which includes quick and abrupt changes.

2) Local Analysis

Finally, we conduct a case-specific interpretable analysis in Fig. 11 for the prediction time steps $\{t_0 + 1, t_0 + 5, t_0 + 13\}$ of the probabilistic forecasts on 14th April 2019 at 06H00. Recalling Fig. 6a, the forecaster predicts a low-price regime at $t_0 + 1$, then, it introduces a regime switch at $t_0 + 5$ in order, finally, to output a high-price regime distribution at $t_0 + 13$. Fig. 11 clearly demonstrates the dependency between the predicted regimes and the different temporal importance patterns of the most dominant drivers. Concretely, the temporal importance of each dominant driver is computed as $\varrho^i_{t_0+j} = \vartheta^i_{t_0+j} \cdot \alpha_{t_0+j}$, $\forall i \in \{\lambda^{h,\text{RT}}, \text{NRV}^h, P^{h,\text{renew}}, x^{h,\text{cal.}}, \lambda^{f,\text{DA}}\}$, $\forall j \in \{-l_{\max}, ..., \tau_{\max}\}$. It can be observed that the proposed model tends to rely on the day-ahead prices for predicting a low imbalance price, while it focuses on past information for predicting a high-price regime. These observations are further corroborated by the importance spikes occurring on the conditioning steps $\{t_0, t_{0-7}, t_{0-21}\}$ in Fig. 11b and 11c, which correspond to past time steps characterized by a high-price regime (see Fig. 6a).

V. CONCLUSION

The paper proposes a novel Transformer-based model for interpretable, high-performance multi-horizon probabilistic forecasting of real-time electricity prices. Such prices are important market signals for market players aiming at reducing their imbalance costs or maximizing balancing actions. However, their predictions are highly complex since the prices are characterized by regime switching behavior and spikes.

In this context, we leverage recent advances in deep neural networks to provide a new, well-suited approach for predicting real-time electricity prices. In a detailed case study, we illustrate that the proposed model is able to outperform state-of-the-art forecasting methods, with a respective decrease of 4.5% in the CRPS metrics compared with the first benchmark method, i.e., Bahd-Seq2Seq. In addition, a global analysis of the interpretable outcomes allow highlighting both the most important features, i.e., the set $\{\lambda^{h,\text{RT}}, \text{NRV}^h, P^{h,\text{renew}}, x^{h,\text{cal.}}, \lambda^{f,\text{DA}}\}$, and the features' temporal window of the proposed model (from $t_0 - 7$ to $t_0 + 16$). Finally, a case-specific interpretable analysis demonstrates the ability of the proposed model to capture different temporal attention patterns of each input according to the price-regime predicted.

Two complementary lines of research can be envisaged for future works. First, interpretability in probabilistic time series forecasting deserves more in-depth studies. For instance, an extensive benchmark study investigating both post-hoc and inherent interpretable probabilistic time series forecasting models could provide interesting insights on their usefulness for providing human interpretable outcomes. Second, the application of such interpretable time series forecasting models in other market frameworks could be also envisaged. Indeed, this will ensure of their efficiency in the presence of other market dynamics, while providing valuable outcomes about their respective drivers.

REFERENCES

[1] B. F. Hobbs and S. S. Oren, "Three waves of u.s. reforms: Following the path of wholesale electricity market restructuring," *IEEE Power Energy Mag*, vol. 17, no. 1, pp. 73–81, 2019.

[2] D. Lee, H. Shin, and R. Baldick, "Bivariate probabilistic wind power and real-time price forecasting and their applications to wind power bidding strategy development," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6087–6097, 2018.

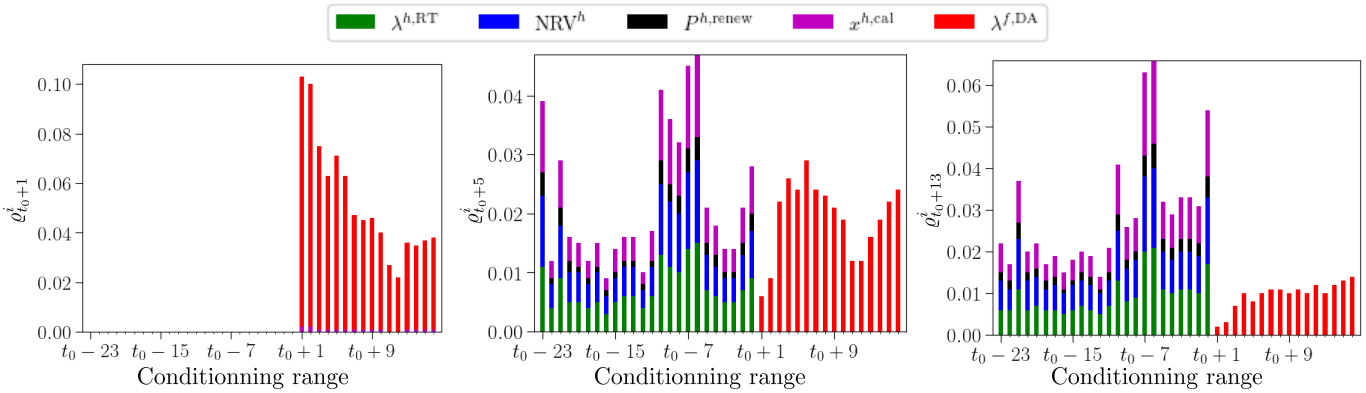
[3] B. Mohammadi-Ivatloo, H. Zareipour, M. Ehsan, and N. Amjady, "Economic impact of price forecasting inaccuracies on self-scheduling of generation companies," *Electr. Power Syst. Res.*, vol. 81, no. 2, pp. 617–624, 2011.

[4] H. Chitsaz, P. Zamani-Dehkordi, H. Zareipour, and P. P. Parikh, "Electricity price forecasting for operational scheduling of behind-the-meter storage systems," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6612–6622, 2018.

[5] J.-F. Toubeau, J. Bottieau, Z. De Grève, F. Vallée, and K. Bruninx, "Data-driven scheduling of energy storage in day-ahead energy and reserve markets with probabilistic guarantees on real-time delivery," *IEEE Trans. Power Syst.*, vol. 36, no. 4, pp. 2815–2828, 2021.

[6] R. Weron, "Electricity price forecasting: A review of the state-of-the-art

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2022.3195970



(a) Prediction time step $t_0 + 1$ (b) Prediction time step $t_0 + 5$ (c) Prediction time step $t_0 + 13$

Fig. 11. The attention-weighted importances of the most dominant drivers over both past and future conditionning ranges, i.e., $\varrho^{i}_{t_0+j} = \vartheta^{i}_{t_0+j} \cdot \alpha_{t_0+j}, \forall i \in \{\lambda^{\mathrm{RT}}, \mathrm{NRV}, P^{h,\mathrm{renew}}, x^{\mathrm{cal.}}, \lambda^{\mathrm{DA}}\}, \forall j \in \{-l_{\max}, ..., \tau_{\max}\}$, when performing the probabilistic forecasts on the 14th April 2019 at 06H00 for the prediction time steps $\{t_0 + 1, t_0 + 5, t_0 + 13\}$ (which respectively corresponds to Fig. 11a, 11b, and 11c).

with a look into the future," *Int J Forecast*, vol. 30, pp. 1030–1081, 2014.

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

[8] L. Hirth and I. Ziegenhagen, "Balancing power and variable renewables: Three links," *Renew. Sust. Energ. Rev.*, vol. 50, pp. 1035–1051, 2015.

[9] B. Hua, D. A. Schiro, T. Zheng, R. Baldick, and E. Litvinov, "Pricing in multi-interval real-time markets," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2696–2705, 2019.

[10] A. Papavasiliou and G. Bertrand, "Market design options for scarcity pricing in european balancing markets," *IEEE Trans. Power Syst.*, pp. 1–1, 2021.

[11] E. Litvinov, F. Zhao, and T. Zheng, "Electricity markets in the united states: Power industry restructuring processes for the present and future," *IEEE Power Energy Mag.*, vol. 17, no. 1, pp. 32–42, 2019.

[12] T. Gomez, I. Herrero, P. Rodilla, R. Escobar, S. Lanza, I. de la Fuente, M. L. Llorens, and P. Junco, "European union electricity markets: Current practice and future view," *IEEE Power Energy Mag.*, vol. 17, no. 1, pp. 20–31, 2019.

[13] S. Vejdan and S. Grijalva, "The value of real-time energy arbitrage with energy storage systems," in *2018 IEEE PESGM*, 2018, pp. 1–5.

[14] H. Ding, P. Pinson, Z. Hu, and Y. Song, "Optimal offering and operating strategies for wind-storage systems with linear decision rules," *IEEE Transact. Power Syst.*, vol. 31, no. 6, pp. 4755–4764, 2016.

[15] D. W. Bunn, J. N. Inekwe, and D. MacGeehan, "Analysis of the fundamental predictability of prices in the british balancing market," *IEEE Trans. Power Syst.*, vol. 36, no. 2, pp. 1309–1316, 2021.

[16] P. Wang, H. Zareipour, and W. D. Rosehart, "Descriptive models for reserve and regulation prices in competitive electricity markets," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 471–479, 2014.

[17] M. Olsson and L. Soder, "Modeling real-time balancing power market prices using combined sarima and markov processes," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 443–450, 2008.

[18] I. Dimoulkas, M. Amelin, and M. R. Hesamzadeh, "Forecasting balancing market prices using hidden markov models," in *2016 13th International Conference on the EEM*, 2016, pp. 1–5.

[19] M. B. Olsson and L. Söder, "Modeling swedish real-time balancing power prices using nonlinear time series models," in *2010 IEEE 11th International Conference on PMAPS*, 2010, pp. 358–363.

[20] S. Jaehnert, H. Farahmand, and G. L. Doorman, "Modelling of prices using the volume in the norwegian regulating power market," in *2009 IEEE Bucharest PowerTech*, 2009, pp. 1–7.

[21] T. Jónsson, P. Pinson, H. A. Nielsen, and H. Madsen, "Exponential smoothing approaches for prediction in real-time electricity markets," *Energies*, vol. 7, no. 6, pp. 3710–3732, 2014.

[22] G. Klaeboe, A. L. Eriksrud, and S.-E. Fleten, "Benchmarking time series based forecasting models for electricity balancing market prices," *Energy Systems*, vol. 6, no. 1, pp. 43–61, 2015.

[23] A. Lucas, K. Pegios, E. Kotsakis, and D. Clarke, "Price Forecasting for the Balancing Energy Market Using Machine-Learning Regression," *Energies*, vol. 13, no. 20, pp. 1–16, 2020.

[24] S. B. Taieb and A. F. Atiya, "A bias and variance analysis for multistep-ahead time series forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 62–76, 2016.

[25] X. Zhan, S. Zhang, W. Y. Szeto, and X. M. Chen, "Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree," *J. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 125–141, 2020.

[26] J. Dumas, I. Boukas, M. M. de Villena, S. Mathieu, and B. Cornélusse, "Probabilistic forecasting of imbalance prices in the belgian context," in *2019 16th International Conference on the EEM*, 2019, pp. 1–7.

[27] J. Bottieau, L. Hubert, Z. De Grève, F. Vallée, and J. Toubeau, "Very-short-term probabilistic forecasting for a risk-aware participation in the single price imbalance settlement," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1218–1230, 2020.

[28] "Elia Group elia grid data," http://www.elia.be/en/grid-data/data-download, accessed: 2021-07-10.

[29] D. He and W.-P. Chen, "A real-time electricity price forecasting based on the spike clustering analysis," in *2016 IEEE/PES T & D*, 2016, pp. 1–5.

[30] H. Yang and K. R. Schell, "Hfnet: Forecasting real-time electricity price via novel gru architectures," in *2020 International Conference on PMAPS*, 2020, pp. 1–6.

[31] E. Kraft, D. Keles, and W. Fichtner, "Modeling of frequency containment reserve prices with econometrics and artificial intelligence," *J. Forecast.*, vol. 39, no. 8, pp. 1179–1197, 2020.

[32] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, p. 68–77, 2019.

[33] R. Cynthia, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019.

[34] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 5, no. 5, pp. 726–742, 2021.

[35] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[36] Z. Yang, A. Zhang, and A. Sudjianto, "Enhancing explainability of neural networks through architecture constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2610–2621, 2021.

[37] B. Lim, S. . Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021.

[38] S. Sarp, M. Kuzlu, U. Cali, O. Elma, and O. Guler, "An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool," in *2021 IEEE Power & Energy Society ISGT*, 2021, pp. 1–5.

[39] A. A. Ismail, M. Gunady, H. C. Bravo, and S. Feizi, "Benchmarking deep learning interpretability in time series predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[40] J.-F. Toubeau, J. Bottieau, Y. Wang, and F. Vallee, "Interpretable probabilistic forecasting of imbalances in renewable-dominated electricity systems," *IEEE Trans. Sustain. Energy*, pp. 1–1, 2021.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, p. 6000–6010.

[42] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on FedCSIS*, 2020, pp. 179–183.

[43] T. Q. Nguyen and J. Salazar, "Transformers without tears: Improving the normalization of self-attention," *IWSLT*, 2019.

[44] J.-F. Toubeau, J. Bottieau, F. Vallée, and Z. De Grève, "Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1203–1215, 2019.

[45] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Phil. Trans. R. Soc. A.*, vol. 379, no. 2194, p. 20200209, 2021.

[46] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," 2016, arXiv preprint, 1511.07289.

[47] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," 2016, arXiv preprint, 1604.06737.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," 2020, arXiv preprint, 1907.00235.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, arXiv preprint, 1512.03385.

[51] Y. Wang, D. Gan, M. Sun, N. Zhang, C. Kang, and l. Zongxiang, "Probabilistic individual load forecasting using pinball loss guided lstm," *Applied Energy*, vol. 235, pp. 10–20, 02 2019.

[52] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen, and C. Kang, "Combining probabilistic load forecasts," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3664–3674, 2019.

[53] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[54] J. Morales, A. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Renewable Energy Sources—Modeling and Forecasting*, 2014, vol. 205, pp. 15–56.

[55] A. Khosravi, S. Nahavandi, and D. Creighton, "Construction of optimal prediction intervals for load forecasting problems," *IEEE Transact. Power Syst.*, vol. 25, no. 3, pp. 1496–1503, 2010.

[56] H. Quan, D. Srinivasan, and A. Khosravi, "Short-term load and wind power forecasting using neural network-based prediction intervals," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 303–315, 2014.

[57] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit—normal distributions," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 61, pp. 555–576, 08 2012.

[58] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[59] N. Meinshausen, "Quantile regression forests," *J. Mach. Learn. Res.*, vol. 7, p. 983–999, 2006.

[60] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, p. 1189–1232, 2021.

[61] J.-F. Toubeau, T. Morstyn, J. Bottieau, K. Zheng, D. Apostolopoulou, Z. De Grève, Y. Wang, and F. Vallée, "Capturing spatio-temporal dependencies in the probabilistic forecasting of distribution locational marginal prices," *IEEE Transact. Smart Grid*, vol. 12, no. 3, pp. 2663–2674, 2021.

[62] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted gaussian process regression," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 300–308, 2018.

[63] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When gaussian process meets big data: a review of scalable gps," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4405–4423, 2020.

[64] D. Y. Pavlov, A. Gorodilov, and C. Brunk, "Bagboo: a scalable hybrid bagging-the-boosting model," in *2010 ACM Int. Conf. Inf. Knowl.*, 2010, pp. 1897–1900.

[65] K. V. Rashmi and R. Gilad-Bachrach, "Dart:dropouts meet multiple additive regression trees," in *2015 AISTATS*, 2015, pp. 89–497.

[66] A. Rogozhnikov and T. Likhomanenko, "Infiniteboost: building infinite ensembles with gradient descent," 2018, arXiv preprint, 1706.01109.

Jeremie Bottieau (Student Member, IEEE) received the Master degree and the Ph.D. degree in electrical engineering from the University of Mons (Belgium) in 2017 and 2022, respectively.

He is currently a postdoctoral researcher in the Power Systems and Markets Research Group, University of Mons. His research interests include short-term forecasting and decision-making in electricity markets.

Yi Wang (Member, IEEE) received the B.Sc. degree from Huazhong University of Science and Technology in June 2014, and the Ph.D. degree from Tsinghua University in January 2019. He was a visiting student with the University of Washington from March 2017 to April 2018. He served as a Postdoctoral Researcher in the Power Systems Laboratory, ETH Zurich from February 2019 to August 2021.

He is currently an Assistant Professor with the Department of Electrical and Electronic Engineering, University of Hong Kong. His research interests include data analytics in smart grids, energy forecasting, multi-energy systems, Internet-of-things, cyber-physical-social energy systems.

Zacharie De Grève (Member, IEEE) received the Electrical and Electronics Engineering degree from the Faculty of Engineering, University of Mons, Mons, Belgium, in 2007. He was a Research Fellow of the Belgian Fund for Research (F.R.S/FNRS) until 2012, when he got the Ph.D. degree in electrical engineering from the University of Mons, where he is currently an Associate Professor with the Electrical Power Engineering Unit.

His main research interests deal with the application of Machine Learning and Operations Research to electric power systems, and energy systems more generally. He also develops expertise in computational electromagnetics.

François Vallée (Member, IEEE) received the degree in civil electrical engineering and the Ph.D. degree in electrical engineering from the Faculty of Engineering, University of Mons, Belgium, in 2003 and 2009, respectively. He is currently a Professor and leader of the "Power Systems and Markets Research Group" at the University of Mons. His Ph.D. work has been awarded by the SRBE/KBVE Robert Sinave Award in 2010. His research interests include PV and wind generation modeling for electrical system reliability studies in presence of dispersed generation.

Jean-François Toubeau (Member, IEEE) received the Master degree and the Ph.D. degree in electrical engineering, from the University of Mons (Belgium) in 2013 and 2018, respectively.

He is currently a postdoctoral researcher with the Belgian Fund for Research (F.R.S/FNRS) within the "Power Systems and Markets Research Group" of the University of Mons. His research mainly focuses on bridging the gap between machine learning and decision-making in modern power systems.