Interpretable Probabilistic Forecasting of Imbalances in Renewable-Dominated Electricity Systems

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

***Abstract*—High penetration of renewable energy such as wind power and photovoltaic (PV) requires large amounts of flexibility to balance their inherent variability. Making an accurate prediction of the future power system imbalance is an efficient approach to reduce these balancing costs. However, the imbalance is affected not only by renewables but also by complex market dynamics and technology constraints, for which the dependence structure is unknown. Therefore, this paper introduces a new architecture of sequence-to-sequence recurrent neural networks to efficiently process time-based information in an interpretable fashion. To that end, the selection of relevant variables is internalized into the model, which provides insights on the relative importance of individual inputs, while bypassing the cumbersome need for data preprocessing. Then, the model is further enriched with an attention mechanism that is tailored to focus on the relevant contextual information, which is useful to better understand the underlying dynamics such as seasonal patterns. Outcomes show that adding modules to generate explainable forecasts makes the model more efficient and robust, thus leading to enhanced performance.**

***Index Terms*—Attention mechanisms, balancing market, deep learning, interpretability, multi-horizon forecasting.**

I. Introduction

THE costs of renewable energies, in particular wind power and photovoltaic (PV), have significantly decreased in recent years. Even though these technologies start becoming economically viable without external economic incentive [1], they are associated with a major economic issue from a system perspective since their cost of generation is not reflective of the final cost incurred to the end-users. Indeed, their uncertain and intermittent nature needs to be continuously compensated by other flexible/balancing resources, which may incur significant balancing costs [2].

Manuscript received March 11, 2021; revised May 23, 2021; accepted June 21, 2021. Date of publication June 24, 2021; date of current version March 22, 2022. This work was supported via the energy transition funds project "EPOC 2030-2050" organized by the FPS economy, S.M.E.s, Self-employed and Energy. Paper no. TSTE-00256-2021. (*Corresponding author: Yi Wang.)*

Jean-François Toubeau is with the post-doctoral research fellow of the National Fund of Scientific Research (FNRS). He is working at the the Power Systems and Markets Research Group, University of Mons 7000 Mons, Belgium (e-mail: Jean-Francois.TOUBEAU@umons.ac.be).

Jérémie Bottieau and François Vallée are with the Power Systems and Markets Research Group, University of Mons, 7000 Mons, Belgium (e-mail: Jeremie.bottieau@umons.ac.be; Francois.VALLEE@umons.ac.be).

Yi Wang is with the Power Systems Laboratory, Department of Information Technology and Electrical Engineering, ETH Zürich ETH Zürich, 8092 Zürich, Switzerland (e-mail: yiwang@eeh.ee.ethz.ch).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSTE.2021.3092137.

Digital Object Identifier 10.1109/TSTE.2021.3092137

In the competitive European framework, the balancing task is managed through three complementary policies. First, each market actor is responsible for maintaining the energy balance within its portfolio on a quarter-hourly basis, which is achieved by trading (surpluses and shortages of) energy in different market floors, ranging from years ahead to day-ahead and intraday stages [3]. Second, in case of a real-time imbalance of the system arising from the aggregated deviations from individual agents, the Transmission System Operator (TSO) is responsible for restoring the frequency in its control area. This is achieved through a market-based mechanism [4] in which the TSO anticipatively buys the upward and downward power capacity in day-ahead to the different actors, such that these power reserves can be deployed in case of need. Finally, balancing costs from the activation of reserves are billed to market agents contributing to the frequency deviation via the imbalance settlement mechanism [5].

In this context, the development of a reliable forecasting tool for upcoming system imbalances is an inexpensive solution that can create substantial value for both the TSO and market actors. First, relying on such forecasts is an important milestone towards a more dynamic (and closer to real-time) sizing and procurement of operating reserves [6]. Indeed, better quantifying the imbalance risk may avoid oversizing the reserve capacity, thus reducing system costs [7]. Then, such multi-horizon forecasts may also support the evolution from a curative balancing stage (in which the activation of reserves passively responds to observed frequency signals) towards pro-active management of balancing resources with the goal of minimizing the expected activation costs [8]. Finally, reserve providers can also benefit from these frequency forecasts [9]. For instance, wind power producers can rely on this information for deloading strategies whereby they operate below their maximum power to create margins for upward regulation [10].

In general, the imbalance is commonly referred to as Area Control Error (ACE), which is the difference between scheduled and actual values of the power exchanged in the TSO control area, i.e., the power imbalance as if no balancing control was performed by the TSO [11]. In the last decades, the ACE was assumed to be fully driven by load and wind forecasting errors [12], [13]. However, in modern power systems, the ACE is guided by complex and volatile market-related conditions. In particular, frequency deviations are mainly driven by both the chaotic dynamics of atmospheric systems (e.g., for renewable-based generation and electric heating) and human behavior (e.g., regarding the risk-based dispatch of resources in electricity

markets), and therefore exhibit significant variability and uncertainty. For these reasons, the literature on ACE forecasting is still very sparse.

A pioneering work is presented in [14], where it is shown that feedforward neural networks (FFNNs) yield better results than econometric models for predicting the daily imbalance medians. Such observations are confirmed in [15] wherein FFNNs are also used for the deterministic forecast of the grid frequency. These contributions are completed in [16], where a comparative study of different forecasters is performed for lead times from 1 to 10 minutes. Results show that Seasonal Auto-Regressive Moving Average (SARMA) models and FFNNs can be outperformed by an improved version of the exponential smoothing method enriched with a Kalman filter. These methods only rely on past values of the ACE time series, thus neglecting the information from other relevant covariates. In [17], additional features, such as load and wind forecasts, temperature data, and market-related variables, have been used to improve the predictions of hourly imbalances. Based on this, [18] also capitalizes on exogenous variables and shows that recent advances in recurrent neural networks (RNNs), i.e., deep learning architectures tailored to capture time dependencies and non-linear behaviors, can improve the prediction accuracy. However, all these models were developed in a deterministic setting. An extension to a probabilistic framework is thus presented in [19] using quantile regression forests. Finally, the modeling capabilities of a multi-attention recurrent neural network are combined with mixture density networks to generate probabilistic forecasts of primary frequency data in [20].

Overall, all the above models lack an automated procedure for selecting relevant features among the different available variables, thus relying on a cumbersome manual preprocessing phase. Moreover, they mainly focus on prediction performance, which is done at the expense of interpretability, i.e., forecasters are developed without the ability to provide insights on how they exploit the different inputs, which is a major barrier for the industrial acceptability of such models. In this paper, a new generic methodology is developed for the multi-horizon probabilistic forecasting of the ACE with the objective of efficiently coupling high performance with interpretability. To boost accuracy, the model is built upon an encoder-decoder architecture, in which contextual data are treated by an encoder that summarizes this information, which is then fed into a decoder that yields the desired predictions [21]. Encoder and decoder blocks are composed of Long Short Term Memory (LSTM) RNNs due to their ability to represent non-linear time dynamics [22]. The model is here augmented with two modules dedicated to interpretability. First, an automated selection procedure of relevant inputs is internalized into the model, which provides direct insights on the relative importance of individual inputs. Second, two attention-based layers are designed for identifying the salient past and future time dependencies.

The contributions of this paper are threefold.

1) The traditional encoder-decoder model is enriched by designing two different encoder blocks, respectively processing past-observed and future-known inputs. The encoded vectors are then combined and treated by a decoder to generate the multi-horizon predictions. This solution allows the integration of time-varying data over a horizon that differs from the prediction window, thus fully leveraging all contextual dependencies.
2) The model is coupled with an attention mechanism, i.e., a computing layer dedicated to selectively concentrate on salient time information while ignoring irrelevant parts of the input sequence. Attention layers are implemented for both past and known future data. This serves two complementary objectives, i.e., capturing long-term dependencies across all (past and future) time steps while enabling interpretability by identifying the relevant time steps in the multi-horizon forecast [23].
3) A generic feature selection process is internalized into the proposed bi-attentional model through the incorporation of new modules able to differentiate the importance of each covariate (at each time step) on the prediction [24]. Interestingly, these variable selection modules directly provide insights into which variables are driving the forecast outcome. Moreover, they automatically select inputs without requiring expert knowledge or task-specific engineering [25].

Overall, the proposed bi-attentional model is generic and offers a natural and efficient way to deal with the heterogeneity of data sources (such as past ACE observations, past values of exogenous variables, or the known information about the future). The model moreover provides quantile forecasts which are free of any distributional assumption, and can also be easily extended to perform forecasts with different lead times at different granularity levels.

The value of the methodology is tested on real-world data from the Belgian power system [26] by investigating the ACE dynamics learned by the proposed model. Then, the performance of the architectural variations is also investigated by performing an ablation analysis (in which we quantify the loss of accuracy when each individual component is removed from the model). Besides, comprehensive comparisons with other successful approaches in probabilistic forecasting such as econometric models, gradient boosting methods, and random forests are conducted.

The rest of the paper is organized as follows. Section II presents the different components of the proposed prediction model. Section III introduces the benchmarks. Section IV conducts case studies on a real-world dataset from Belgium and compares different models. Finally, Section V summarizes important outcomes and perspectives.

II. Methodology

The objective of the paper is to predict, at the forecast creation time t_0, the next ACE values with a granularity of 15 minutes, together with an accurate quantification of the forecast uncertainty. This is achieved by training a parametric model f_θ (whose parameters θ have to be learned) for generating the conditional distribution of the future ACE signal over the horizon $\{t_0 + 1, \ldots, t_0 + \tau_{\max}\}$.

In addition to past ACE values $y_t \in \mathbb{R}$, a wide variety of time-dependent data can be accessed for the multi-horizon

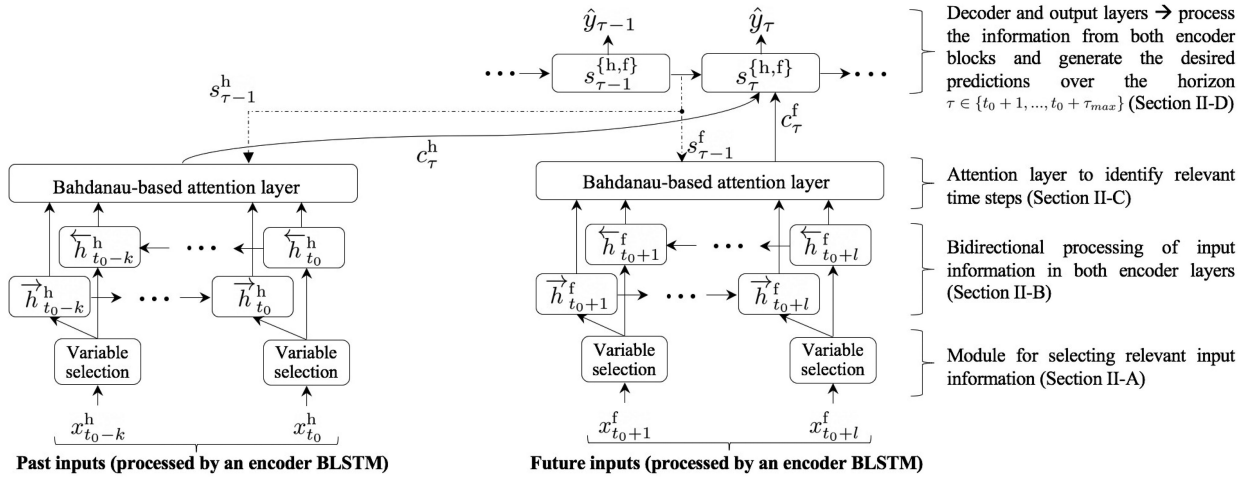

Fig. 1. Global model characterized by a bi-attentional architecture to generate interpretable multi-horizon predictions with heterogeneous input data.

ACE forecasting. These exogenous variables can be divided into two classes, i.e., the m_h historical covariates $x_t^{\mathrm{h}} \in \mathbb{R}^{m_h}$ which usually come from measurements (and whose values are thus sequentially gathered as time t elapses), and the m_f inputs $x_t^{\mathrm{f}} \in \mathbb{R}^{m_f}$ whose future value are known (e.g., calendar information or market quantities cleared in forward stages). Incorporating these exogenous variables may benefit the forecasting performance. At this stage, it is interesting to notice that we use t-indices when referring to the conditioning range (of inputs) $\{t_0 - k, \ldots, t_0 + l\}$, while τ is used for the forecast horizon (of outputs) $\{t_0 + 1, \ldots, t_0 + \tau_{\max}\}$.

To optimally leverage such heterogeneous information, a novel sequence-to-sequence architecture is designed to process the m_h past and m_f future inputs in two different blocks. The global architecture of the forecaster, which is developed to achieve both high forecasting performance and interpretability of the outcomes, is depicted in Fig. 1. First, input features are fed into variable selection modules that filter out the contributions of irrelevant data (Section II-A). The resulting information is processed by an encoder-decoder model, which is dedicated to efficiently capture (local) time correlations (Section II-B). The architecture is augmented with attention mechanisms to learn (long-term) dependencies. These additional layers are constructed for both past and future data, thus forming a global model with two temporal attention layers (Section II-C). The resulting bi-attentional model is trained to generate probabilistic forecasts of the future ACE values y_τ, which is achieved by minimizing the quantile loss. This yields quantiles $\hat{y}_\tau = \{\hat{y}_\tau^{(q)}\}_{q \in \mathcal{Q}}$ of the target distribution for different relevant probability levels $q \in \mathcal{Q}$ (Section II-D). Finally, we discuss how the model can yield interpretable outcomes (Section II-E).

At this stage, it should be noted that all modules of the global model are jointly trained in the learning phase, thus guaranteeing the consistency of the framework.

A. Variable Selection

With the increased willingness to promote a transparent and competitive market, many variables become available, but their actual relevance for predicting the ACE (or any other market indice) is typically unknown. Historical covariates x_t^{h} and the known future information x_t^{f} are summarized in Table I.

TABLE I
EXOGENOUS INPUT FEATURES OF THE FORECASTING TOOL

Past data x_t^h	Known future data x_t^f
Calendar information	
Load and wind generation, cross-border exchanges, actual dispatch of gas and pumped-hydro energy storage	Forecast of load and wind generation, day-ahead commitment of gas and pumped-hydro energy storage

The forecasts of load and wind generation come from the Belgian TSO, which provides this information with the goal of promoting transparent and competitive markets [27]. For confidentiality reasons, the underlying modeling framework cannot be disclosed. Also, it should be noted that only past realizations of net imports/exports are incorporated into the model since they correspond to the actual physical flows measured at the interface with neighboring countries. The net balance of (future) scheduled cross-border flows, which are estimated based on commercial energy transactions, are not considered. This choice is supported by preliminary analyses revealing that these future values bring no additional explanatory power to the ACE forecast.

In this work, we aim to avoid a manual data selection, which requires a complex multi-variate analysis deciphering intricate (high-dimensional) dependencies between inputs. Indeed, some input features may be found irrelevant in univariate analysis while offering valuable information when coupled with other variables.

The variable selection is thus internalized into the global model through the use of variable selection networks [24]. Practically, inputs at each time step $\{t_0 - k, \ldots, t_0 + l\}$ are fed into a dedicated processing layer that learns which are the most salient variables. The methodology is here presented for past inputs (but the same principle applies for future inputs). The inputs x_t^{h} at time $t \in \{t_0 - k, \ldots, t_0\}$ are individually standardized in the range $[-1, 1]$ before being fed into the variable selection block,

wherein the normalized inputs $\bar{x}_t^{\mathrm{h}}$ are transformed into a new vector through the linear mapping:

$$x_t^{\mathrm{h,trans}} = W_t^{\mathrm{h}} \bar{x}_t^{\mathrm{h}} + b_t^{\mathrm{h}} \tag{1}$$

where W_t^{h} and b_t^{h} are parameters that need to be optimized during the training phase. The linear transformed input $x_t^{\mathrm{h,trans}} \in \mathbb{R}^{m_h}$ are then processed by a softmax layer that normalizes its inputs, i.e., all m_h variables are converted in the range [0, 1], and their sum is equal to 1, so that they can be seen as probabilities:

$$v_t^{\mathrm{h}} = \mathrm{softmax}(x_t^{\mathrm{h,trans}}) \tag{2}$$

where $v_t^{\mathrm{h}} \in \mathbb{R}^{m_h}$ is the vector defining the importance of variables at time t. The raw features $\bar{x}_t^{\mathrm{h}}$ are then weighted by:

$$\tilde{x}_t^{\mathrm{h}} = v_t^{\mathrm{h}} \circ \bar{x}_t^{\mathrm{h}} \tag{3}$$

where $\circ$ is the element-wise multiplication, and $\tilde{x}_t^{\mathrm{h}}$ naturally filters out irrelevant features (before being incorporated into the forecaster) that could negatively affect accuracy.

In parallel, each calendar information (month of the year, day of the week, quarter-of-an-hour of the day, day type) is processed through an embedding layer, which is transforming the calendar variable into a fixed-dimensional vector. This reduces the dimensionality of the input space (e.g., by avoiding the use of 96 binary variables to encode all quarter-of-an-hour of the day), while providing a new learned meaningful representation able to capture their relative significance. In this way, time steps with similar properties are placed close to each other in the embedding vector, which cannot be achieved with traditional techniques such as one-hot encoding.

At each time step, both calendar and other inputs are then combined and integrated into the forecaster.

B. Encoder-Decoder Architecture

Efficiently processing the past observed values and the known information about the future is a non-trivial task due to the necessity to feed models with a fixed-dimensional input vector (since the number m_h of past covariates is likely to differ from the number m_f of known inputs about the future). A successful solution is provided by sequence-to-sequence models, which are composed of two different blocks and are thus sometimes referred to as encoder-decoder models. These architectures have shown promising results for challenging tasks such as translation applications [28].

Traditionally, the encoder processes the past data (known at the forecast creation time t_0) over a look-back window of k past steps $\{t_0 - k, \ldots, t_0\}$, with the goal of mapping this input sequence to a latent state vector c^{enc}. This conditioning step serves to capture the past dynamics, and the decoder subsequently exploits the representation c^{enc}, along with the known information about the future, to generate the multi-horizon predictions.

In such a framework, the future context that can be incorporated into the model is limited to the length of the multi-horizon forecast, which may prevent to fully leverage contextual information. To alleviate this issue, as shown in Fig 1, we implement a novel architecture wherein two distinct encoder neural networks are defined. These are respectively processing past data x_t^{h} for look-back steps $t \in \{t_0 - k, \ldots, t_0\}$, and future data x_t^{f} for look-ahead steps $t \in \{t_0 + 1, \ldots, t_0 + l\}$. The resulting (transformed) information c^{h} and c^{f} are then jointly treated into a decoder layer that provides the prediction over the horizon $\{t_0 + 1, \ldots, t_0 + \tau_{\max}\}$.

In general, encoder and decoder blocks can be represented by any modeling framework. Here, with the goal of capturing the complex ACE dynamics, we focus on the Long Short Term Memory (LSTM) RNNs, which are characterized by a time-dependent hidden state providing an internal representation of past events for propagating information through time [29]. For clarity, hidden states associated with encoder blocks are denoted by h_t, while s_t is used for hidden states of the decoder. For both encoders, the LSTM model is further enriched with bidirectional processing of the input features, resulting in a Bidirectional LSTM (BLSTM) model that can process the context in both positive and negative time direction (thus capturing both forward and backward time dependencies) [30]. Practically, at each time step t of the horizon, the BLSTM model has access to two hidden states, i.e., $\overrightarrow{h}_t$ that provides a representation of previous events, and $\overleftarrow{h}_t$ that summarizes the information of the following time steps. We define h_t as the concatenation of both forward and backward states, i.e., $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$. It thus contains a summary of both preceding and following time horizons, but with a focus on the information at time t.

As previously mentioned, we use two separate BLSTM models for respectively encoding past $\{t_0 - k, \ldots, t_0\}$ and future $\{t_0 + 1, \ldots, t_0 + l\}$ data, such that two context vectors c^{h} and c^{f} are generated. Without loss of generality, we present the encoder block processing the past time steps $t \in \{t_0 - k, \ldots, t_0\}$, i.e.,

$$\overrightarrow{h}_t^{\mathrm{h}} = \mathcal{H}_{LSTM}\left(x_t^{\mathrm{h}}, \overrightarrow{h}_{t-1}^{\mathrm{h}}\right), \overleftarrow{h}_t^{\mathrm{h}} = \mathcal{H}_{LSTM}\left(x_t^{\mathrm{h}}, \overleftarrow{h}_{t+1}^{\mathrm{h}}\right) \tag{4}$$

$$\begin{aligned} c^{\mathrm{h}} &= f\left(\left\{\overrightarrow{h}_{t_0-k}^{\mathrm{h}}, \overleftarrow{h}_{t_0-k}^{\mathrm{h}}, \ldots, \overrightarrow{h}_{t_0}^{\mathrm{h}}, \overleftarrow{h}_{t_0}^{\mathrm{h}}\right\}\right) \\ &= f\left(\{h_{t_0-k}^{\mathrm{h}}, \ldots, h_{t_0}^{\mathrm{h}}\}\right) \end{aligned} \tag{5}$$

where $\mathcal{H}_{LSTM}$ is the LSTM composite function [29], and f is a user-defined nonlinear function, i.e., a common solution is $f(\{h_{t_0-k}^{\mathrm{h}}, \ldots, h_{t_0}^{\mathrm{h}}\}) = h_{t_0}^{\mathrm{h}}$. The same procedure is applied for processing future time steps $t \in \{t_0 + 1, \ldots, t_0 + l\}$ to yield the context vector c^{f}.

Then, the decoder reads this transformed information c^{h} and c^{f} to yield the predictive distribution over the horizon $\tau \in \{t_0 + 1, \ldots, t_0 + \tau_{\max}\}$:

$$p\left(y_\tau \mid \mathbf{y}_{t_0-k:t_0}, \mathbf{x}_{t_0-k:t_0}^{\mathrm{h}}, \mathbf{x}_{t_0+1:t_0+l}^{\mathrm{f}}\right) = g\left(s_\tau^{\{h,f\}}, c^{\mathrm{h}}, c^{\mathrm{f}}\right) \tag{6}$$

where g is a nonlinear function that provides the probability of y_τ, and $s_\tau^{\{h,f\}}$ is the hidden state of the decoder layer.

C. Attention Mechanism

In traditional sequence-to-sequence models, the use of a fixed-length vector c^{enc} to exchange information between encoder and decoder modules is a bottleneck for the performance. Indeed, this makes it difficult for the model to deal with long horizons and to capture multi-scale characteristics. A solution to this problem is to rely on an attention-based mechanism. The principle is

to add a computing layer on top of the encoder wherein the input information is mapped into a sequence of vectors $\mathbf{c}^{\text{enc}} = \{c^{\text{enc}}_{t_0+1}, \ldots, c^{\text{enc}}_{t_0+\tau_{\max}}\}$, instead of a single vector c^{enc}. The goal is to selectively adapt the relevant encoder-side information for each time step of the prediction horizon $\{t_0+1, \ldots, t_0+\tau_{\max}\}$.

Different architectures have been developed to derive the context vector $\mathbf{c}^{\text{enc}}$. In this work, we use an attention mechanism inspired by the Bahdanau layer [23]. In that framework, all the hidden states of the encoder are jointly used for constructing each element of the $\tau_{\max}$-dimensional context vector c^{enc}_{τ}, thus facilitating the representation of long-term dependencies. However, in its traditional form, the Bahdanau-based encoder-decoder is limited to exploit the attention from the past information x^{h}_t. The principle is here extended to allow the attention for both past and known future data, giving rise to a global model with two temporal attention layers, one for each of both BLSTM encoder networks (Fig. 1). The resulting architecture is able to learn complex alignments between the different time steps of the studied horizon. As depicted in Fig. 1, both (past and future) encoder blocks are connected to a decoder layer via these attention-based mechanisms.

Practically, the outputs $\{h^{\text{h}}_{t_0-k}, \ldots, h^{\text{h}}_{t_0}\}$ and $\{h^{\text{f}}_{t_0+1}, \ldots, h^{\text{f}}_{t_0+l}\}$ of both BLSTM encoders are fed into distinct attention layers to extract the context vectors $\{c^{\text{h}}_{t_0+1}, \ldots, c^{\text{h}}_{t_0+\tau_{\max}}\}$ and $\{c^{\text{f}}_{t_0+1}, \ldots, c^{\text{f}}_{t_0+\tau_{\max}}\}$, respectively. Finding these context vectors is achieved through an alignment model, which is here described for the encoder block that processes past time steps:

$$c^{\text{h}}_{\tau} = \sum_{t=t_0-k}^{t_0} \alpha_{\tau t} h^{\text{h}}_t \tag{7}$$

where h^{h}_t is the hidden state of the corresponding encoder BLSTM, and the alignment vector $\alpha_{\tau t}$ is computed by

$$\alpha_{\tau t} = \frac{e^{r_{\tau t}}}{\sum_{t=t_0-k}^{t_0} e^{r_{\tau t}}} \tag{8}$$

with $r_{\tau t}$ quantifying the degree of correspondence between the encoder inputs at time t and the prediction outcome at time τ. This alignment is computed with a FFNN, using the encoder state h^{h}_t at time t and the hidden state $s^{\text{h}}_{\tau-1}$ of the corresponding LSTM network (of the decoder) at time $\tau-1$.

$$r_{\tau t} = \text{align}\left(h^{\text{h}}_t, s^{\text{h}}_{\tau-1}\right) \tag{9}$$

A similar alignment architecture is used to obtain the context information c^{f}_{τ} for the encoder dedicated to process future information x^{f}_t.

D. Decoder and Output Layers

The outputs c^{h}_{τ} and c^{f}_{τ} of both attentional layers are then processed by a decoder composed of two distinct LSTM layers, which respectively yield hidden states s^{h}_{τ} and s^{f}_{τ}:

$$s^{\text{h}}_{\tau} = \mathcal{H}_{LSTM}\left(c^{\text{h}}_{\tau}, s^{\text{h}}_{\tau-1}\right),\ s^{\text{f}}_{\tau} = \mathcal{H}_{LSTM}\left(c^{\text{f}}_{\tau}, s^{\text{f}}_{\tau-1}\right) \tag{10}$$

Finally, the outputs of both LSTM layers are combined $[s^{\text{h}}_{\tau}; s^{\text{f}}_{\tau}]$ and processed by a FFNN to generate the predictions $\{\hat{y}_{t_0+1}, \ldots, \hat{y}_{t_0+\tau_{\max}}\}$.

With the goal that the forecasts can be used for tuning risk management in dedicated decision-making processes, we predict (with the global model f_θ) the conditional distribution of the future ACE values. In particular, this target distribution is approximated using Quantile Regression (QR) [31], by outputting the conditional quantiles $\hat{y}_{\tau} = \{\hat{y}^{(q)}_{\tau}\}_{q\in\mathcal{Q}}$ for different relevant probability levels $q \in \mathcal{Q}$. Practically, the model parameters θ are learned by minimizing the total quantile loss (also referred to as pinball loss) over each sequence of the training database [31]. However, when multiple quantiles are jointly predicted, one may risk running into the well-known issue of quantile crossing. This issue is checked in post-processing, but no occurrence was observed over the whole test set. In the literature, some augmented loss functions have been introduced to tackle this problem [32], but they are out of the scope of this work.

At each time step τ of the prediction horizon, the QR forecasting model thus provides a $|\mathcal{Q}|$-dimensional output (i.e., one for each quantile of interest):

$$\hat{\mathbf{y}}_{t_0+1:t_0+\tau_{\max}} = f_\theta\left(\mathbf{y}_{t_0-k:t_0}, \mathbf{x}^{\text{h}}_{t_0-k:t_0}, \mathbf{x}^{\text{f}}_{t_0+1:t_0+l}\right) \tag{11}$$

Note that the quantiles $\hat{y}^{(q)}_{\tau}$ are determined without any prior assumption on the form of the target distribution, thus providing high flexibility in the uncertainty representation.

E. Model Interpretability

The bi-attentional model shown in Fig. 1 combines two complementary modules that yield interpretable outcomes, i.e., the attention mechanisms identify significant temporal patterns, and the variable selection blocks discriminate the importance of different input features at each time step.

First, in addition to aligning the relevant information at different time steps with a specific target (thus simplifying the task of the decoder layer in making accurate predictions), another key advantage of the attention mechanism (explained in Section II-C) is to enhance explainability. In particular, the magnitude of attention weights $\alpha_{\tau t}$ in both encoding layers provides insights on how well all inputs around time t are related to the output $\hat{y}_{\tau}$ at position τ.

From (8), these $\alpha_{\tau t}$-values can be intuitively interpreted as a probability that the output $\hat{y}_{\tau}$ is aligned with information $\{x^{\text{h}}_t, x^{\text{f}}_t\}$ at time t, thereby reflecting the importance of these inputs. In our global architecture wherein past and known future information are processed within separate blocks, it is thus possible to identify the learned importance of persistent patterns in the data as well as backward dependencies.

Second, the direct incorporation of feature selection blocks (described in Section II-A) into the global model makes it inherently interpretable. The vector of variable selection weights v^{h}_t in (2) has the dimensionality of the inputs, and each of its components explicitly quantifies the importance of the corresponding variable (at time step t).

From a practical perspective, these insights are essential for the acceptance and practical deployment of the model at the industry level, where experts want to understand the underlying mechanisms by which a model works. Moreover, these insights on how the model generates predictions can also be leveraged

by model builders to further improve performance. For instance, it can be used to increase the size of both look-back and look-ahead windows if attention peaks are observed at the endpoints or to look for new features that could further complement those selected by the model.

III. Benchmarks

We compare our proposed bi-attentional model with a wide range of techniques for multi-horizon forecasting.

First, we implement three different naive methodologies that provide interesting baselines.

- A probabilistic generalization of persistence (Persistence) based on a random walk model. The forecast assumes a Gaussian distribution where the mean is given by the last available ACE realization, and the variance is determined by exponential smoothing of previous squared errors [33].
- An averaging model (Average) in which the predicted ACE distribution is composed of all ACE values of the historical database. The same distribution is thus applied for each time step of the horizon.
- A step-wise averaging model (Step-average) in which a different ACE distribution is computed for each individual time step, based on all past observations corresponding to this specific period of the day.

Second, we construct three state-of-the-art models in time series forecasting.

- A probabilistic Auto-Regressive Integrated Moving Average (ARIMA), which assumes a constant variance and a linear time correlation.
- A quantile regression forest (QRF), i.e., a method that generalizes random forests (in which the outcomes of independent decision trees are averaged) for estimating quantiles. The number of trees is set to 500.
- A gradient boosting regression tree (GradBoost) trained with quantile regression to generate probabilistic predictions. This approach sequentially creates new models (in an additive fashion) to forecast the residuals of the global model obtained at the previous stage. The number of boosting stages is fixed to 100.

It should be noted that the ARIMA model is only fed with past ACE observations, while tree-based models (QRF and GradBoost) have access to the same input data as the proposed bi-attentional model.

Third, in order to understand the contribution of the different architectural variations of the proposed reference model (Ref), we perform an ablation study. This consists of investigating the decrease in the model performance when certain components are removed.

- Bidirectional processing (Ref-Bidir): the bidirectional processing of data in both encoder blocks is replaced by a traditional unidirectional LSTM layer with 24 neurons.
- Attention mechanisms (Ref-Att): the attention-based layers are removed from the model such that a single fixed-length vector $c^{\{h,f\}}$ is fed by each encoder into its corresponding decoder layer.

TABLE II
Training and Inference Times of Different Models

Models	Training time [s]	Inference time [s]
ARIMA	184	0.01
QRF	$\tau_{\max}$· 4,400	0.22
GBDT	$\tau_{\max}$· 175	0.63
Ref	951	0.12

- Variable selection (Ref-VarSel): the variable selection networks are removed from the global model, i.e., the inputs x_t^{h} and x_t^{f} are directly fed into their corresponding BLSTM encoder.

For each forecaster (except the parameter-free naive approaches), an hyperparameter optimization is conducted to identify the optimal model complexity. This is achieved through an extensive random search. The same number of iterations is used across all benchmarks.

IV. Case Study

The procedure is performed on actual data obtained from Elia, the Belgian TSO [27]. The database is composed of 6 years of data (from 2015 to 2020), which are divided into sequences of length $\{t_0 - k, \ldots, t_0 + l\}$ for the inputs with their corresponding targets covering time steps $\{t_0 + 1, \ldots, t_0 + \tau_{\max}\}$. Note that 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 $\tau_{\max} = 16$ time steps, and we compute the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles of the target distribution (i.e., $|\mathcal{Q}| = 9$) for each of these time periods.

The sequences are then partitioned into three parts, i.e., a training set (for learning the relationship between the ACE and the inputs), a validation set (for tuning the model hyperparameters), and a test set (for evaluating the model performance). Here, the full year 2020 is used as a test set, while the sequences related to years 2015-2019 are split into a training set (containing 85% of these data) and a validation set (with 15 % of the data). The training time of the different (non-naïve) models is provided in Table II. Simulations have been performed on an Intel Core i7-3770 CPU @ 3.4 GHz with 16 Gb of RAM. All variants from the reference bi-attentional model have similar training times, and they are thus not differentiated in the results.

Note that, to avoid tree-based models to have an output of dimension $\tau_{\max} \cdot |Q|$, a different model is trained for each time step τ. This reduces the output size of each model (thus facilitating training), but at the expense of an increased training time. In particular, each of the $\tau_{\max} = 16$ GBDT models takes around 3 minutes to train (for a total time of more than 45 minutes), while each of the QRF models takes 73 minutes.

The reference bi-attentional model takes around 16 minutes for training one configuration. However, once the model is trained, the (online) inference time for generating new predictions is insignificant (lower than one second). This property applies for all (naïve, tree-based, and deep learning) models, which makes them well suited for close to real-time purposes. It should be noted that if one is interested in integrating the

new information that is continuously revealed over time, the reference bi-attentional model can be dynamically adapted using exclusively the new data (by applying gradient descent only on these samples). This solution provides quick and efficient updates, and circumvents the need to retrain the global architecture with the whole set of historical data.

To find the optimal model complexity (e.g., number of neurons per layers) during training, different configurations need to be tested, which requires each time to retrain a new model from scratch. Interestingly, this phase of hyper-parameter optimization is significantly facilitated for the reference model. Indeed, thanks to the different modules to improve the interpretability, we observe that the resulting performance is more robust to hyper-parameters, i.e., small variations in accuracy among different architectures are recorded. This property is a valuable advantage of the proposed model with respect to tree-based or traditional deep learning approaches (that can be very sensitive to changes in the model complexity). Indeed, no strong expert knowledge is necessary to properly hand-tune the model, which may strongly boost its acceptability in the industry.

Practically, the search ranges for hyper-parameters of the bi-attentional model are listed hereafter:

- number of neurons by layer = [6, 12, 24, 48]
- number of past time steps k = [4, 8, 12, 16, 20, 32, 96]
- number of future time steps l = [4, 8, 12, 16, 20, 32, 96]
- learning rate = $[10^{-2}, 10^{-3}, 10^{-4}]$
- dropout rate = [0, 0.1]

Outcomes reveal that the optimal numbers of past and future steps to process are $k = 20$ and $l = 16$, respectively. A single hidden layer is imposed for all encoder and decoder blocks. Both BLSTM encoders are composed of 24 neurons within forward and backward layers, while the LSTM decoders have 12 neurons. The training is performed with a batch size of 96 daily sequences using the Adam algorithm [34], in which the initial learning rate is fixed at 0.001. Early stopping is implemented to prevent overfitting, which consists of terminating the training phase before the model starts to memorize the data instead of learning the underlying dependencies. This regularization technique seems sufficient since the best model was obtained without dropout in the learning phase.

A. Evaluation Metrics

Two important aspects need to be jointly considered when evaluating the performance of a probabilistic forecaster, i.e., reliability and sharpness. Reliability measures the statistical correctness between the predicted quantiles and the actual observations. It is computed using the percentage of the ex-post samples (in the test set) that are actually lower than the corresponding quantile forecast. Sharpness is a direct measure of the width of the prediction intervals. To evaluate this trade-off between both concepts, three probabilistic scores are used, i.e., the quantile loss function, the Winkler score, and the continuous ranked probability score (CRPS).

First, we use the quantile loss E_τ^Q, i.e., the same function as the one used to train the probabilistic forecaster, to quantify the accuracy of the predictions at each time step of the test set (12).

$$E_\tau^Q = \sum_{q\in\mathcal{Q}} q \max\left(y_\tau - \hat{y}_\tau^{(q)}, 0\right) + (1-q)\max\left(\hat{y}_\tau^{(q)} - y_\tau, 0\right) \tag{12}$$

where the values $\hat{y}_\tau^{(q)}$ are the predicted quantiles (i.e., outputs of the forecaster), and y_τ the actual ACE observations.

The quantile loss E_τ^Q is complemented with the Winkler score, which quantifies the forecast quality at different probability levels [35]. For a prediction interval of $(1-\beta)\cdot 100\%$, the Winkler score E_τ^W is defined as:

$$E_\tau^W = \begin{cases} \delta_\tau & L_\tau \le y_\tau \le U_\tau \\ \delta_\tau + 2(L_\tau - y_\tau)/\beta & y_\tau < L_\tau \\ \delta_\tau + 2(y_\tau - U_\tau)/\beta & y_\tau > U_\tau \end{cases} \tag{13}$$

where $L_\tau = \hat{y}_\tau^{\beta/2}$ and $U_\tau = \hat{y}_\tau^{1-\beta/2}$ are the lower and upper bounds of the prediction interval defined by the confidence level β, and $\delta_\tau = U_\tau - L_\tau$ is the interval width (i.e., sharpness). If an ACE realization y_τ is within the predicted interval $[L_\tau, U_\tau]$, the Winkler score E_τ^W 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. In this paper, E_τ^W is calculated for confidence levels $\beta = \{0.02, 0.1, 0.2, 0.5\}$.

Third, the CRPS computes the quadratic difference between the forecast cumulative distribution function (CDF) F_τ and the empirical CDF of the actual observation y_τ [36]:

$$\text{CRPS}(F_\tau, y_\tau) = \int_{-\infty}^{\infty} (F_\tau(z) - H(z - y_\tau))^2 \, dz \tag{14}$$

where H is the shifted unit step function (i.e., its value is equal to 0 for arguments z lower than y_τ, and is equal to 1 when arguments z are higher than y_τ). Interestingly, the CRPS measures both reliability and sharpness, and has the same unit as the forecasted variable, which makes it easily interpretable [37]. A low CRPS is associated with an accurate probabilistic forecast. In this work, the CRPS is approximated using simulated samples [38] with:

$$\text{CRPS}(\hat{F}_\tau, y_\tau) = \frac{1}{M}\sum_{i=1}^{M} |X_{\tau,i} - y_\tau| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} |X_{\tau,i} - X_{\tau,j}| \tag{15}$$

where the predicted quantiles $\hat{y}_\tau^{(q)}$ are interpolated using a cubic spline to obtain a continuous forecast CDF $\hat{F}_\tau$. Then, inverse random sampling of $\hat{F}_\tau$ is used to obtain M samples $X_{\tau,1}, \cdots, X_{\tau,M}$.

B. Comparison of Models

Fig. 2 presents the forecast accuracy in terms of quantile loss E^Q of the different models as a function of the prediction horizon to analyze up to which time span it is relevant to anticipate grid imbalances.

The results are averaged over all sequences of the test set and may provide valuable insights for policy-makers regarding the

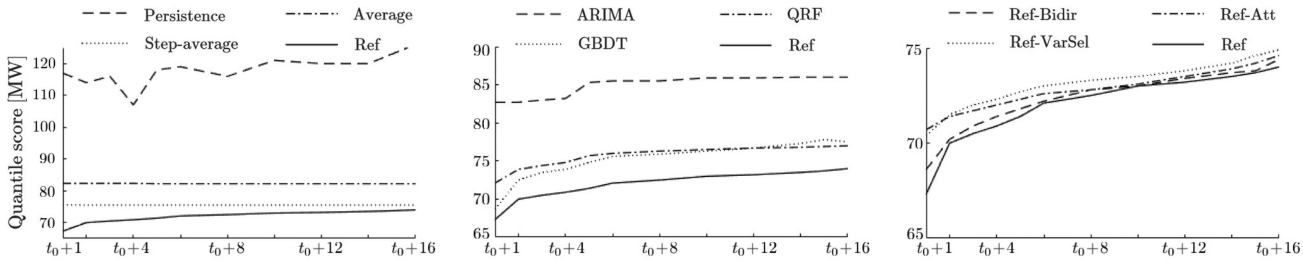

Fig. 2. Quantile score of different methods (averaged over the whole test set) for the prediction horizon $\{t_0+1,\ldots,t_0+\tau_{\max}\}$. Outcomes are split into three sub-figures, i.e., the first one comparing the reference model (Ref) with naïve benchmarks, the second one comparing state-of-the-art models, while the third one shows the outcome of the ablation study (which allows to clearly visualize the impact on performance of the different components of the Ref model).

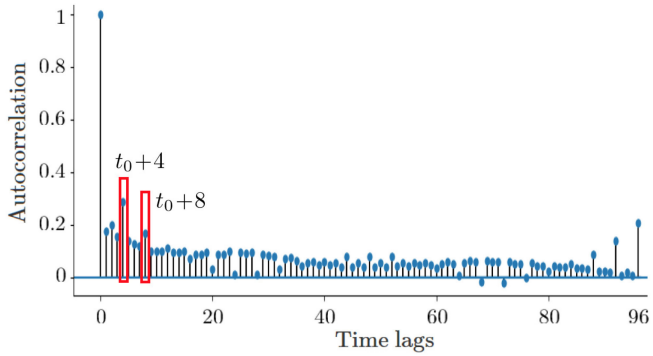

Fig. 3. Autocorrelation function of the ACE.

definition of suitable balancing rules, e.g., by identifying the optimal gate closure times for the reserve provision.

Due to the high variability of the ACE, the Persistence model performs very badly. Indeed, it simply propagates past observations without the ability to infer the most likely future realizations. This behavior can be explained by the autocorrelation function of the ACE (Fig. 3), which shows moderate time dependencies among consecutive ACE realizations. However, we observe peaks at t_0+4 and t_0+8 (corresponding to hourly dependencies), which explains the improved accuracy of Persistence at these time steps (with respect to surrounding steps). As further explained in Section IV-C, this corresponds to the hourly market transitions (when ramping effects are taking place).

Then, it can be observed in Fig. 2 that the bi-attentional model outperforms all (naive and state-of-the-art) benchmarks over the 4-hour prediction horizon. In particular, it yields an average improvement of 3.6% (over the first four quarter-hourly periods $\{t_0+1,\ldots,t_0+4\}$ of the prediction horizon) compared to GradBoost, which is the next best model. QRF also provides reliable outcomes in the first time steps, but the accuracy quickly decreases to drop below naive models after only one hour. Although tree-based models have the advantage of processing any type of (continuous and integer) inputs without the need to normalize these data, they still lack the processing abilities of more advanced architectures.

Overall, the differences in performance (between all forecasters) quickly decrease over time since all methods tend towards a naive representation of the historical distributions (differentiated between the quarter-of-an-hour), thus illustrating the difficulty to capture the extreme volatility and unpredictability of the ACE over long horizons. In that respect, the naive benchmark (Step-average) estimating the future ACE distributions independently for each time step (based on the historical ACE values measured each day at the corresponding period) provides a competitive baseline, which is not easily overcome.

We also see that the ARIMA model achieves poor prediction performance, which may arise from different reasons. First, it only leverages past ACE values (thus neglecting the potential of exogenous information) in a linear framework. Second, the prediction intervals of the ARIMA are computed analytically, assuming that residuals are uncorrelated and normally distributed, which may lead to poor accuracy when these assumptions are violated [39]. Third, only the uncertainty associated with the random error term is considered to find the prediction interval, thereby disregarding the uncertainty in the choice of the model and in the parameter estimates [40]. Hence, although econometric models make fewer structural assumptions and can thus be seen as more flexible than other approaches, they are of limited value for modeling the uncertainty of complex variables such as the ACE.

From the ablation analysis, it can be observed that the attention mechanisms, as well as the variable selection blocks, have both a significant impact on performance, with an averaged increase in the quantile loss of around 2.5% in both cases (on the first hour of the horizon). The reduced accuracy of the non-attentional model is a direct result of the lack of direct connections with relevant time steps of the surrounding horizon, thus highlighting the benefits of this alignment procedure. This observation tends to reflect the existence of important seasonality patterns in the data, which is revealed in the following of the paper with the analysis of the ACE signal (in Fig. 4). Likewise, making the model interpretable (by incorporating the feature selection into the architecture) enables to focus on the relevant information at each time step, thereby improving the model performance.

To complement these results, we conduct a probabilistic quantification of the future imbalance conditions. To that end, the CRPS and the Winkler scores for different reliability levels are computed for all models for the first step t_0+1 of the horizon, and the results are provided in Table III. We observe similar trends in both metrics. In particular, the proposed reference (Ref) model achieves a higher accuracy for all intervals. Such outcomes are interesting since gains in accuracy for extreme quantiles can lead to significant cost savings from a system

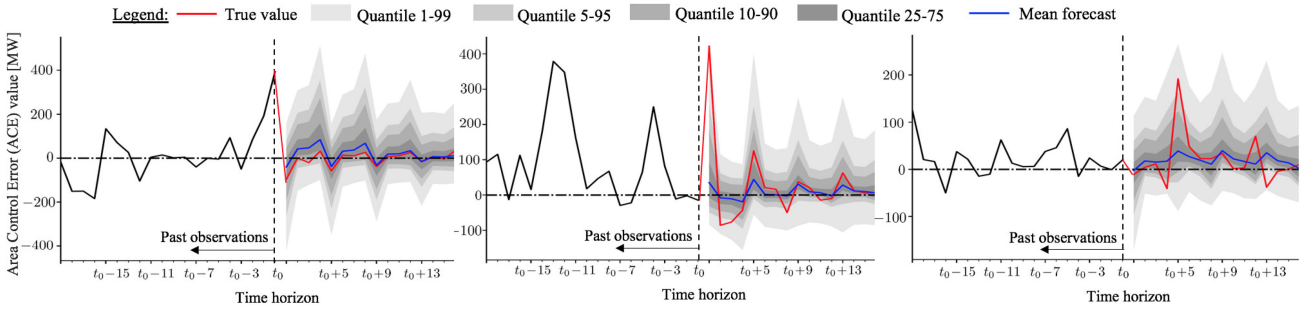

Fig. 4. Multi-horizon probabilistic forecasts of ACE for the 13th April 2020 at 3 different periods, i.e., 1:00 am, 5:00 pm, and 11:00 pm, respectively.

TABLE III
WINKLER SCORE OF DIFFERENT METHODS FOR THE FIRST STEP OF THE PREDICTION HORIZON $t_0 + 1$

Models	CRPS [MW]	Winkler score [MW]			
		$\beta = 0.02$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 0.5$
Persistence	33.6	488.3	266.3	192.7	111.3
Average	22.0	502.9	273.2	194.9	112.7
Step-average	20.7	446.1	241.2	175.3	105.0
ARIMA	23.4	490.5	259.8	195.7	118.3
QRF	19.9	389.2	224.0	166.7	102.0
GBDT	20.3	358.4	208.6	157.3	98.7
Ref-Bidir	19.5	356.9	207.2	156.4	98.2
Ref-Att	20.1	372.4	218.3	162.9	100.5
Ref-VarSel	19.9	378.6	217.6	162.1	99.7
Ref	**19.2**	**354.1**	**203.7**	**153.8**	**96.2**

perspective (e.g., through reduced needs in terms of reserve sizing or through improved risk-based activation of reserves).

For illustrating the quality of the outcomes obtained using the bi-attentional model, probabilistic forecasts (over the 4 hours horizon) for April 13, 2020, at three different periods, i.e., 1:00 am, 5:00 pm, and 11:00 pm, are depicted in Fig. 4. The predicted intervals tend to properly embed the actual ACE realizations, suggesting that the volatility of the signal is well captured. The ACE uncertainty follows a seasonal pattern with clear spikes at hourly intervals, i.e., higher imbalances occur in the first quarter-of-an-hour of each new hour. This observation arises from the fact that European day-ahead markets are characterized by an hourly time resolution, i.e., the electricity is traded as energy in hourly blocks [kWh/h]. This results in a dispatch of power plants with ramping trajectories between consecutive hours that are causing imbalances in the TSO control area. It is worth noting that several theoretical solutions are proposed to tackle this problem, such as trading power trajectories instead of energy blocks [41], or by introducing asynchronous energy blocks [42].

C. Interpretability

In this part, we analyze how the bi-attentional model allows to interpret the predictions. In particular, we discuss the outcomes of the model for April 13, 2020, at 5.00 pm to provide reflections on the relationships that have been learned during training.

First, we focus on the variable selection blocks since the magnitude of weights v_t^{h} in (2) is an image of the importance of each individual input at time t. These weights are thus represented

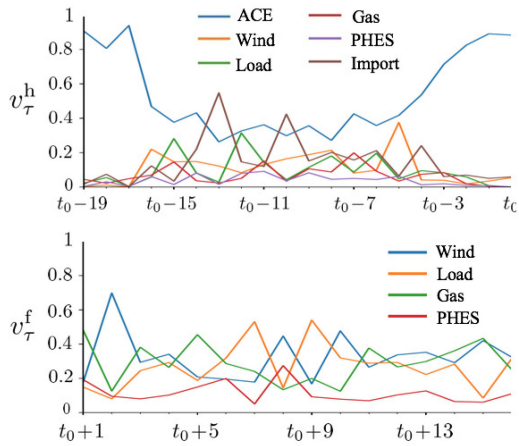

Fig. 5. Representation of variable selection weights over both past (upper graph) and future (lower graph) horizons.

over both past $\{t_0 - k, \ldots, t_0\}$ and future $\{t_0 + 1, \ldots, t_0 + l\}$ conditioning horizons, and are depicted in Fig. 5.

Outcomes from the past horizon show that the model extracts information from only a subset of inputs to generate the predictions. As expected, the past ACE observations are the most relevant inputs since these values contain the recent dynamics of the target signal. For more distant time steps, the ACE becomes less important, and the model diversifies the relevant variables. In particular, net imports/exports occurring between 3 to 4 hours before the forecast creation time (at 5:00 pm) seem to bring valuable information. By analyzing raw data, this can be explained by the fact that net imports are lower during this period due to a peak in PV generation, such that the resulting energy mix is subject to more uncertainty that is ultimately affecting the ACE. This uncertainty is translated into larger prediction intervals in the later afternoon (when there is still PV generation), which are then progressively decreasing. Also, variables such as wind power and total load also seem to help the model. Indeed, they allow to inform on the expected variability of the system imbalance (e.g., the magnitude of prediction errors in wind generation is higher for windy days). In contrast, information from pumped hydro energy storage (PHES) is of little value.

However, these observations should be put into perspective with the information provided by attention weights (in both encoding layers) that shed light on the most important time steps to make the predictions. These learned weights are represented

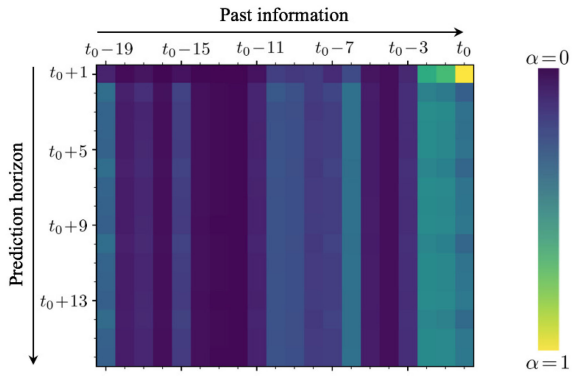

Fig. 6. Importance of the past conditioning range for the prediction over the horizon of interest $\{t_0+1,\ldots,t_0+\tau_{\max}\}$.

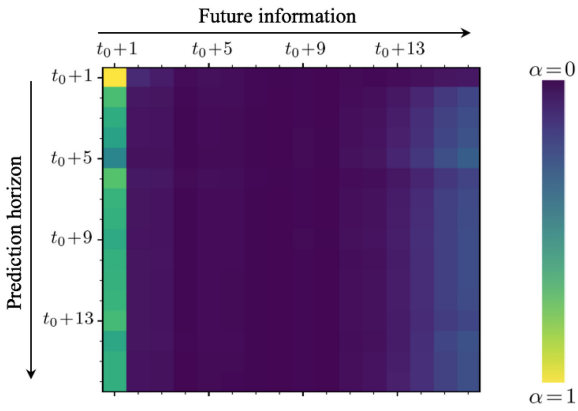

Fig. 7. Importance of the future conditioning range for the prediction over the horizon of interest $\{t_0+1,\ldots,t_0+\tau_{\max}\}$.

using a matrix plot in Fig. 6 for the $k=20$ past time steps. The rows of the matrix indicate the steps of the prediction horizon, while the columns represent the conditioning range. From this, we clearly see that the last time steps are significantly dominating the prediction importance. Hence, the ranking of variables for time steps of more than 1 h back is less significant than the recent observations that show the strong importance of the ACE.

Then, the same analysis is carried out with the $l=16$ steps of the future context, and the results are presented in Fig. 7. All prediction steps seem to focus almost exclusively on the information related to t_0+1, thus disregarding the information from $\{t_0+2,\ldots,t_0+l\}$. Interestingly, such conclusions are supported by our additional simulations revealing that replacing future wind and load values with their perfect forecasts brings limited value to the model, which clearly shows that the ACE forecaster mainly relies on past data to make the probabilistic predictions.

In contrast to econometric methods, which are based on arbitrary assumptions for capturing seasonality patterns, the attention layers can learn such (forward and backward) dependencies directly from historical data. This enables to easily conclude on which positions of the input sequence the model focuses its attention for each step of the prediction horizon. Such a capability is very useful to build trust in the model since human experts can judge the relevance of the outcomes through their knowledge of the environment. In this way, both modules for interpretability act as a guarantee that meaningful variables impact the output, i.e., it is possible to check that the model has captured a truthful causality between explanatory variables and the predicted outputs.

V. Conclusion

The paper is devoted to the multi-horizon probabilistic forecasting of the ACE, which measures the system imbalance when no balancing actions are performed. This information is indeed essential for reducing the balancing costs while ensuring the reliability of the power system, thus paving the way towards the large-scale integration of renewable sources in electricity markets. The ACE is a complex signal affected by time-dependent market and technical constraints. The main objective of the work is to avoid relying on black-box models that do not explain how the available inputs are used to generate the predictions.

In contrast to customary beliefs considering that high-performance deep learning-based models are inevitably opaque, the application of our model to the Belgian power system clearly illustrates that the objectives of accuracy and interpretability are closely dependent. In particular, enriching the model with attributes dedicated to yield more explainable outcomes presents useful advantages. First, it allows the final user of the model to better understand the logic of the model. This work is thus an important step towards the acceptability of deep learning models in the industry by providing information on the underlying relationships learned by the forecaster. Then, the interpretability blocks also reduce the burden of human experts by avoiding time-consuming preprocessing stages to identify suitable inputs.

References

[1] E. Pursiheimo, H. Holttinen, and T. Koljonen, "Inter-sectoral effects of high renewable energy share in global energy system," *Renewable Energy*, vol. 136(C), pp. 1119–1129, 2019.

[2] L. Exizidis, J. Kazempour, P. Pinson, Z. De Grève, and F. Vallée, "Impact of public aggregate wind forecasts on electricity market outcomes," *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1394–1405, Oct. 2017.

[3] J.-F. Toubeau, Z. De Grève, and F. Vallée, "Medium-term multimarket optimization for virtual power plants: A stochastic-based decision environment," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1399–1410, Mar. 2018.

[4] G. Notton *et al.*, "Intermittent and stochastic character of renewable energy sources: Consequences, cost of intermittence and benefit of forecasting," *Renewable Sustain. Energy Rev.*, vol. 87(C), pp. 96–105, 2018.

[5] J.-F. Toubeau, J. Bottieau, Z. De Greve, F. Vallee, 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, Jul. 2021.

[6] H. Holttinen *et al.*, "Methodologies to determine operating reserves due to increased wind power," *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 713–723, Oct. 2012.

[7] N. Menemenlis, M. Huneault, and A. Robitaille, "Computation of dynamic operating balancing reserve for wind power integration for the time-horizon 1-48 hours," *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 692–702, Oct. 2012.

[8] M. Høaberg and G. Doorman, "A stochastic mixed integer linear programming formulation for the balancing energy activation problem under uncertainty," in *Proc. IEEE Manchester PowerTech.*, 2017, pp. 1–6.

[9] J. Bottieau, L. Hubert, Z. De Grève, F. Vallée, and J.-F. 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, Mar. 2020.

[10] S. A. Hosseini, J.-F. Toubeau, Z. De Grève, and F. Vallée, "An advanced day-ahead bidding strategy for wind power producers considering confidence level on the real-time reserve provision," *Appl. Energy*, vol. 280, 2020, Art. no. 115973.
[11] M. Høaberg and G. Doorman, "Classification of balancing markets based on different activation philosophies: Proactive and reactive designs," in *Proc. 13th Int. Conf. Eur. Energy Market*, 2016, pp. 1–5.
[12] R. Doherty and M. O'malley, "A new approach to quantify reserve demand in systems with significant installed wind capacity," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 587–595, May 2005.
[13] K. Bruninx and E. Delarue, "A statistical description of the error on wind power forecasts for probabilistic reserve sizing," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 995–1002, Jul. 2014.
[14] M. P. Garcia and D. S. Kirschen, "Forecasting system imbalance volumes in competitive electricity markets," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 240–248, Feb. 2006.
[15] N. Tomin, "Intelligent monitoring and forecasting of the expected operating conditions of electric power system," in *Proc. IEEE 3rd Int. Youth Conf. Energetics*, 2011, pp. 1–8.
[16] J. W. Taylor and M. B. Roberts, "Forecasting frequency-corrected electricity demand to support frequency control," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1925–1932, May 2016.
[17] C. Contreras, "System imbalance forecasting and short-term bidding strategy to minimize imbalance costs of transacting in the spanish electricity market," Master's thesis, Comillas Pontifical Univ., Madrid, Spain, 2016.
[18] O. Yurdakul, F. Eser, F. Sivrikaya, and S. Albayrak, "Very short-term power system frequency forecasting," *IEEE Access*, vol. 8, pp. 141234–141245, 2020.
[19] T. S. Salem, K. Kathuria, H. Ramampiaro, and H. Langseth, "Forecasting intra-hour imbalances in electric power systems," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 9595–9600.
[20] A. Mashlakov, L. Lensu, A. Kaarna, V. Tikka, and S. Honkapuro, "Probabilistic forecasting of battery energy storage state-of-charge under primary frequency control," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 1, pp. 96–109, Jan. 2020.
[21] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," 2017, *arXiv:1711.11053*.
[22] J.-F. Toubeau *et al.*, "Capturing spatio-temporal dependencies in the probabilistic forecasting of distribution locational marginal prices," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2663–2674, May 2021.
[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[24] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," 2019, *arXiv:1912.09363*.
[25] J. Wang, H. Zhong, X. Lai, Q. Xia, Y. Wang, and C. Kang, "Exploring key weather factors from analytical modeling toward improved solar power forecasting," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1417–1427, Mar. 2019.
[26] J.-F. Toubeau, P.-D. Dapoz, J. Bottieau, A. Wautier, Z. De Grève, and F. Vallée, "Recalibration of recurrent neural networks for short-term wind power forecasting," *Electric Power Syst. Res.*, vol. 190, 2021, Art. no. 106639.
[27] Elia NV, "Grid Data," Available at http://www.elia.be/en/grid-data, 2021.
[28] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[29] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[30] 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, Mar. 2019.
[31] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: J. Econometric Soc.*, vol. 46, no. 1, pp. 33–50, 1978.
[32] J. Gasthaus *et al.*, "Probabilistic forecasting with spline quantile function RNNs," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, vol. 89, 2019, pp. 1901–1910.
[33] P. Pinson, "Very short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *J. R. Statist. Soc.*, vol. 61, no. 4, pp. 555–576, 2012.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[35] R. L. Winkler, "A decision-theoretic approach to interval estimation," *J. Amer. Statist. Assoc.*, vol. 67, no. 337, pp. 187–191, 1972.
[36] J. E. Matheson and R. L. Winkler, "Scoring rules for continuous probability distributions," *Manage. Sci.*, vol. 22, no. 10, pp. 1087–1096, 1976.
[37] D. van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renewable Sustain. Energy Rev.*, vol. 81(P1), pp. 1484–1512, 2018.
[38] E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson, "The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification," *Quart. J. Roy. Meteorological Soc.*, vol. 132, no. 621 C, pp. 2925–2942, 2006.
[39] R. L. Schmoyer, "Asymptotically valid prediction intervals for linear models," *Technometrics*, vol. 34, no. 4, pp. 399–408, 1992.
[40] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, "A state space framework for automatic forecasting using exponential smoothing methods," *Int. J. Forecasting*, vol. 18, no. 3, pp. 439–454, 2002.
[41] R. Philipsen, G. Morales-España, M. de Weerdt, and L. De Vries, "Trading power instead of energy in day-ahead electricity markets," *Appl. Energy*, vol. 233-234, pp. 802–815, 2019.
[42] J. Verberk, R. Hermans, P. Van denA. BoschJokić, and J. Frunt, "Systematic design of market-based balancing arrangements for deregulated power systems: An asynchronous solution," in *Proc. IEEE Trondheim PowerTech.*, 2011, pp. 1–7.

Jean-François Toubeau (Member, IEEE) received a degree in civil electrical engineering and the Ph.D. degree in electrical engineering from the University of Mons, Mons, Belgium, in 2013 and 2018, respectively. He is currently a Postdoctoral Researcher of the Belgian Fund for Research (F.R.S/FNRS) within the Power Systems and Markets Research Group of the same University. His research mainly focuses on bridging the gap between machine learning and decision-making in modern power systems.

Jérémie Bottieau (Student Member, IEEE) received the Diploma in electrical engineering from the University of Mons, Mons, Belgium, where he has been working toward the Ph.D. degree with Power Systems and Markets Research Group since 2017. His research interests include short-term forecasting and optimization in electricity markets.

Yi Wang (Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2014, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2019. From 2017 to 2018, he was a Visiting Student with the University of Washington, Seattle, WA, USA.

He is currently a Postdoctoral Researcher with ETH Zürich, Zürich, Switzerland. His research interests include multienergy systems, data analytics, and wireless communication in the smart grid.

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