Probabilistic Forecasting of Battery Energy Storage State-of-Charge under Primary Frequency Control

Mashlakov Aleksei, Lensu Lasse, Kaarna Arto, Tikka Ville, Honkapuro Samuli

Probabilistic Forecasting of Battery Energy Storage State-of-Charge under Primary Frequency Control

Aleksei Mashlakov, Lasse Lensu, Arto Kaarna, Ville Tikka, Samuli Honkapuro

***Abstract*—Multi-service market optimization of battery energy storage system (BESS) requires assessing the forecasting uncertainty arising from coupled resources and processes. For the primary frequency control (PFC), which is one of the highest-value applications of BESS, this uncertainty is linked to the changes of BESS state-of-charge (SOC) under stochastic frequency variations. In order to quantify this uncertainty, this paper aims to exploit one of the recent achievements in the field of deep learning, i.e. multi-attention recurrent neural network (MARNN), for BESS SOC forecasting under PFC. Furthermore, we extend the MARNN model for probabilistic forecasting with a hybrid approach combining Mixture Density Networks and Monte Carlo dropout that incorporate the uncertainties of the data noise and the model parameters in the form of prediction interval (PI). The performance of the model is studied on BESS SOC datasets that are simulated based on real frequency measurements from three European synchronous areas in Great Britain, Continental Europe, and Northern Europe and validated by three PI evaluation indexes. Compared with the state-of-the-art quantile regression algorithms, the proposed hybrid model performed well with respect to the coverage probability of PIs for the different regulatory environments of the PFC.**

***Index Terms*—Attention-based neural network, battery energy storage system (BESS), frequency control, mixture density networks, Monte Carlo dropout, prediction intervals, probabilistic forecasting, state-of-charge (SOC).**

I. Introduction

BATTERY Energy Storage Systems (BESSs) are considered as one of the essential building blocks for a transition towards more sustainable and intelligent power systems. A wide spectrum of system- or grid-oriented BESS applications [1] includes an integration of renewable generation, energy arbitrage, local grid support, and system balancing, just to name a few. A comprehensive management of these applications enables more flexible, reliable, and resilient grid operation capable to handle growing system intermittency and complexity caused by the increasing penetration of renewables and distributed energy resources. It is expected that the costs of stationary and mobile BESSs will continue falling triggering more utility- and residential-scale BESS adoption at different grid levels and for a variety of services [2], [3]. Consequently, new business cases arise for sophisticated analysis and control of BESSs and provoke a substantial body of research that aims to optimize battery operation for boosting economic benefit from all available revenue streams.

A simultaneous provision of multiple services is one of the most common approaches in order to achieve maximum profitability and return on investment from a standalone battery storage. In most of the service combinations summarized in [4], the value stacking from multiple revenue streams is achieved by a service coupling with different requirements such as combining power and energy intensive services, active and reactive power services, or different service timescales. A decision-making process for finding the optimal combination and capacity allocation across the benefits of multiple services is naturally formulated as an optimization problem. This optimization for BESSs is difficult because of complex dependencies and uncertainties of different factors such as market profitability, BESS operational costs, coupled resource power output, etc. Therefore, the success of the BESS operational strategies is strongly dependent on the knowledge of these uncertainties at multiple timescales.

The probabilistic forecasts are considered to be a robust tool for the risk management and efficient decision making under the presence of uncertainties [5]. These forecasts are quantified in the form of prediction intervals (PIs), scenarios, density functions, or probability distributions that allow assessing the uncertainty in the forecasts. This approach facilitates the limitations of traditional point forecasts that define only the conditional mean of the signal and are restricted by very limited information about the forecast uncertainty, as well as sensitivity to forecast errors and unexpected events [6]. The increased popularity of the probabilistic forecasting in comparison to point forecast highlighted in [7] is also seen in an energy industry. Recent applications of the univariate probabilistic forecasting methods in smart grids are focused on the forecasting of electricity market prices [6], [8], renewable power generation [9]–[11], and electricity load [12], [13]. Reviews of the methods for these applications can be found in [7], [14]–[16].

The literature in relation to forecasting the behavior of

Manuscript received June 7, 2019; revised September 15, 2019; accepted October 1, 2019. Date of publication XXXXX XX, 2019; date of current version October 14, 2019. This work was supported by the DIGI-USER research platform of Lappeenranta-Lahti University of Technology LUT, Finland. The work of A. Mashlakov was supported in part by a grant of the Finnish Foundation for Technology Promotion. *(Corresponding author: Aleksei Mashlakov.)*

A. Mashlakov, V. Tikka, and S. Honkapuro are with the Laboratory of Electricity Market and Power Systems, School of Energy Systems, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, 53850 Finland (e-mail: aleksei.mashlakov@lut.fi; ville.tikka@lut.fi; samuli.honkapuro@lut.fi).

L. Lensu and A. Kaarna are with the Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT (e-mail: lasse.lensu@lut.fi; arto.kaarna@lut.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier XX.XXXX/JSAC.2020.XXXXXXX

BESS state-of-charge (SOC) under frequency control is scarce and restricted by forecasting of grid frequency [17], [18]. The complexity of the BESS SOC forecasting incorporates the stochastic nature of the power system frequency and the absence of clear spatial information on large macrogrids that could be used to support the forecasting. Moreover, the forecasting algorithm should be capable of achieving acceptable prediction performance in different regulatory environments. Consequently, the forecast errors belong to the challenges, and estimation of these errors is infeasible to achieve with point forecasts.

The goal of this study is to implement and analyze the probabilistic forecasting to assess the uncertainty of BESS SOC under the primary frequency control (PFC). The results should complement the published work on model predictive optimization of BESS economic dispatch for a provision of multiple services. The basis for the forecasting is a multi-attention recurrent neural network (MARNN), a deep learning framework designed to capture the most relevant contextual information for sequence forecasting. Moreover, this framework is extended to realize a variational MARNN for providing a robust probabilistic forecast. This extension is implemented with Mixture Density Networks (MDNs) and Monte Carlo (MC) dropout allowing to quantify the forecasts in the form of the PIs based on the point forecasting and the error obtained by the uncertainties in the inherent noise in the data and the model parameters. In order to evaluate the performance of this hybrid probabilistic forecasting model for different regulatory environments of PFC, it is tested on BESS SOC datasets simulated based on real power grid frequency measurements from three European synchronous areas in Great Britain (GB) [19], Continental Europe (CE) [20], and Northern Europe (NE) [21], respectively.

The contributions of this study are described as follows:
(1) An extension of MARNN to implement the variational MARNN with a hybrid approach combining the MDN and MC dropout, incorporating both aleatoric and epistemic uncertainties. These approaches are appealing due to their scalability and applicability to any existing neural networks;
(2) Novel application of the variational MARNN for probabilistic forecasting of BESS SOC under the provision of PFC service. This forecasting approach takes an important step forward to the optimal decision-making in smart grids under uncertainties of related processes and can be potentially used at large scale for a variety of other applications;
(3) Performance validation of the variational MARNN for probabilistic BESS SOC forecasting in three European synchronous areas with different droop curve characteristics of primary frequency control. The validation is done with multiple PI evaluation indexes and compared with several quantile regression algorithms that served as the benchmark.

The rest of this paper is organized as follows: Section II provides background information about the PFC and related work devoted to the uncertainties of BESS under simultaneous provision of the PFC along with other services. A theoretical ground for the deep learning framework with the extension to probabilistic forecasting, i.e. the variational MARNN with MC dropout and MDNs, is described in Section III. In Section IV, case studies, feature selection process, MARNN model design and implementation details, benchmark models and criteria to validate the performance efficiency of probabilistic forecasting are presented. The results are demonstrated and discussed in Section V. Finally, the conclusions are drawn and future work is planned in Section VI.

II. Background and related work

This section provides a necessary background information about the present PFC practices and principles of the BESS operation under the PFC. Moreover, a brief review of the research assessing the uncertainty of sub-processes related to the optimization of BESS performance for PFC along with other services is introduced.

A. Primary Frequency Control

In the context of power systems, a pursuit of low-carbon principles is inherent in a displacement of maneuverable fossil-fuel based power plants by intermittent renewable generation and, eventually, imposes challenges of more variable supply and a reduction in system inertia [22]. As a consequence, the stability of a power system is being endangered by more frequent and large frequency deviations, and special control strategies are necessary to compensate for these variations. Flexible loads and the BESSs are expected to become one of the main tools to support system stability in the case of high renewable penetration. Many studies have concluded the quality of power system frequency is improved with the integration of these resources for frequency regulation [23], [24]. Moreover, the same effect has been also demonstrated by simulating aggregated small-scale BESSs [25].

Frequency regulation is generally implemented in three levels with primary, secondary, and tertiary frequency control also referred to as frequency containment reserves (FCR), frequency restoration reserves (FRR), and replacement reserves (RR), respectively. A detailed overview of these services can be found in [26]. This paper is focused on the PFC that is one of the most common BESS applications due to the appropriate technical capabilities of the BESS [27] and possibly higher market dividends compared to other services [28].

The PFC is the first resort that is activated to guarantee the frequency stability of the power system compensating for the offset between the production and the demand. Each of the synchronous areas can have different requirements for the PFC exposed by corresponding droop curve parameters. The exact values of these parameters have an impact on the system dynamics that can be seen from different frequency distributions [29]. The PFC deploys fast-acting automatic resources that aim to hold the frequency within the dead-band (DB) limits Δf_{db} by responding to the frequency deviations $\Delta f(t)$ from the nominal system frequency f_{N}:

$$\Delta f(t) = f(t) - f_{\text{N}}, \tag{1}$$

where $f(t)$ is the locally measured frequency at time t. The response is expressed by the reference BESS power output

$P_{\mathrm{FCR}}(t)$ at every moment according to a governing droop curve as follows:

$$P_{\mathrm{FCR}}(t) = \begin{cases} 0, & |\Delta f(t)| \leq |\Delta f_{\mathrm{db}}| \\ P_{\mathrm{FCR}}^{\max}\left(\frac{\Delta f(t)}{|\Delta f_{\max}|}\right), & |\Delta f_{\mathrm{db}}| < |\Delta f(t)| < |\Delta f_{\max}| \\ P_{\mathrm{FCR}}^{\max}\left(\frac{\Delta f(t)}{|\Delta f(t)|}\right), & |\Delta f(t)| \geq |\Delta f_{\max}| \end{cases}, \tag{2}$$

where $\Delta f_{\max}$ is the full activation frequency deviation (FAFD). A negative frequency deviation below the DB leads to BESS discharging, while BESS charging is provoked by a positive deviation above the DB. If the frequency deviation is within the DB, the power output is equal to zero, otherwise it is proportionally increased with coefficient $\Delta f(t)/|\Delta f_{\max}|$ until the full activation frequency power limit is reached. Exceeding this threshold requires continuous provision of the maximum reference power output $P_{\mathrm{FCR}}^{\max}$ from the BESS during a specified time duration. Moreover, regulatory rules set the time requirements for the full activation that can be extremely small for the BESS.

Taking into account the BESS efficiency η, the change of the BESS SOC within the time period $\Delta t = t_i - t_{i-1}$ is defined in percentage as follows:

$$S(t) = 100\% \int_{t_{i-1}}^{t_i} \frac{\eta P_{\mathrm{FCR}}(t)}{E_{\mathrm{rated}}} dt, \tag{3}$$

where E_{rated} is the nominal energy capacity of the BESS. Since the BESS operation at the PFC market directly affects the BESS life time, the weighting factor between the possible dividends and operation costs should be evaluated to optimize the economic dispatch of the BESS for the PFC.

B. Uncertainties of Primary Frequency Control

A generic algorithm presented in [4] takes into account the uncertainty in the forecasted power and energy requirements for services of dispatching the operation of an active distribution feeder and PFC to allocate the portion of the battery power and energy capability. However, it utilizes a simplified approach for the uncertainty estimation of the required PFC power by setting it to the maximum value. The forecasting uncertainty of photo-voltaic (PV) generation is utilized in [30] for model-predictive optimization for simultaneous provision of local and PFC services by aggregating energy storage units. A simultaneous offering of the BESS in day-ahead energy, spinning reserve, and regulation markets considering the uncertainties in the predicted market prices as well as in the energy deployment in spinning reserve and regulation markets is proposed in [31]. A study in [32] considers using a battery to simultaneously provide frequency regulation service and peak shaving with stochastic joint optimization that captures both the uncertainty of future demand and the uncertainty of future frequency regulation signals.

Thus, when the PFC is provided by the BESS, the linked uncertainties include but are not limited to the power generation output of coupled resource, customer power consumption, market prices, and frequency regulation response. Therefore, the forecasting tools that can support the optimal decision-making under risks of these uncertainties are crucial.

III. Probabilistic Forecasting Model

In a model-dependent probabilistic forecasting, the model uncertainty, or similarly the model estimation error while predicting the outcome of a stochastic process can be explained by the noise in the training data sample and the uncertainties in the models themselves [33]. According to the Bayesian viewpoint, these uncertainties are also referred to as aleatoric and epistemic, and can be captured with Bayesian inference. This procedure assumes a formalization of the uncertainties as posterior probability distributions over either the model outputs, or model parameters, respectively.

To provide a probabilistic forecast of BESS SOC, we quantify the aleatoric and epistemic uncertainties via a MDN and MC dropout, respectively. In what follows, we first introduce the MARNN model as the basis for sequence forecasting, and then proceed to formulate the probabilistic extension with the MDN and MC dropout.

A. Forecasting Framework

Recurrent neural networks (RNNs) are advanced deep learning-based structures that have shown high potential in processing sequentially dependent data. These networks were initially developed for language modeling [34], but nowadays they are also applied for solving sequential forecasting problems in the energy sector [35]. The notable learning ability of the RNNs for sequential forecasting is explained by their structure that is designed to hold relevant information from the past inputs. A vanilla architecture of RNNs is composed of an encoder and decoder that are generally implemented with gated recurrent unit (GRU) or long short term memory (LSTM) unit RNNs. The encoder aims to convert an input sequence into a latent state representation vector that is further transformed by a decoder into an output sequence. However, when a complete sequence of information is encoded in the single vector, it becomes challenging to decode the first inputs and long-range dependencies due to the vanishing gradient problem [36].

Attention mechanism introduced in [37] is one of the latest advances in neural machine translation that led to significant performance improvements of deep RNN models in memorizing long source sentences. In this study, we use MARNN model $f^{\omega}(\cdot)$ that is deploying lag values from multiple previous input sequences $\mathrm{X} = [\mathrm{x}_1, ..., \mathrm{x}_N]$ at decoding time via estimation of the corresponding target variable of output sequence $\mathrm{Y} = [y_1, ..., y_N]$. In this scenario, a combined sequence of encoder hidden states for the new input sequence $\mathrm{x}_n = [x_1, ..., x_T]$ can be defined as $\mathrm{h}_n = [h_1, ...h_T]$. Consequently, these hidden states are retrieved by K attention heads in order to compute the attention vector a_n of the input sequence as follows:

$$a_n = \sum_{k=1}^{K} \sum_{t=1}^{T} \alpha_k h_t, \tag{4}$$

where α_k is an attention weight assigned to the encoder state $h_t = e(x_t, h_{t-1})$. This attention vector is fed into a fully

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2019.2952195, IEEE Journal on Selected Areas in Communications

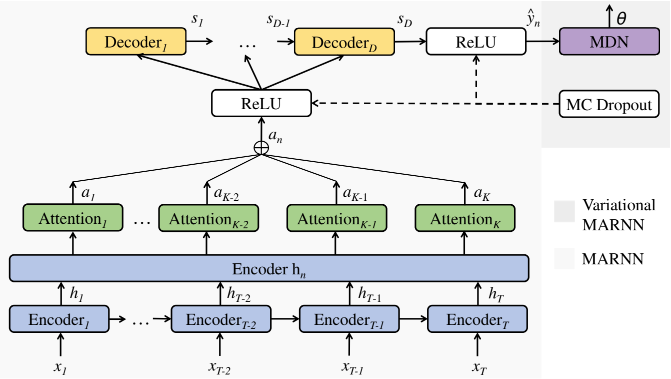


Fig. 1: Extension of multi-attention recurrent neural network (MARNN) model with Mixture Density Network (MDN) layer and Monte Carlo (MC) dropout.

connected layer with ReLU activation function prior to input to the decoder layer

$$\mathrm{ReLU}(a_n) = \max(0, a_n). \tag{5}$$

The last hidden state output of the decoder layer s_D is defined as

$$s_D = g\Big(s_{D-1}, \mathrm{ReLU}(a_n)\Big) \tag{6}$$

where s_{D-1} is the previous decoder state. Finally, the conditional probability over a distinct attention vector a_n for the new target variable $\hat{y}_n$ is then given by

$$\hat{y}_n = p(f^{\omega}(\mathrm{x}_n)|\mathrm{x}_n) = \max(0, s_D). \tag{7}$$

The structure of such MARNN model is illustrated in Fig.1. This architecture of the MARNN enables retrieving the meaningful information by the decoder for each of the output that significantly improves the model performance for forecasting.

B. Mixture Density Networks

The MDNs were proposed in [38] with the motivation to expand the restricted univariate point predictions of conventional neural networks with the multivariate probability distribution of the continuous target variables. These networks exploit the capabilities of Gaussian mixture models (GMMs) [39] to model arbitrary probability density functions of the target variable conditioned on the corresponding input vector using a sufficient number of mixture components.

Based on the assumption of the Gaussian conditional distribution of the target data, and, the fact that the least-squares formalism used in conventional neural networks can be obtained using the maximum likelihood [38], the probability density of the target variable $\hat{y}_n$ is then represented as a linear combination of kernel functions in the form

$$p_{\theta}(\hat{y}_n|\mathrm{x}_n, \theta) = \sum_{m=1}^{M} \gamma_m(\hat{y}_n)\phi_m\Big(\hat{y}_n|\sigma_m(\mathrm{x}_n), \mu_m(\mathrm{x}_n)\Big), \tag{8}$$

where $\theta = \{\gamma_m, \mu_m, \sigma_m\}_{m=1}^{M}$ is a set of M GMM components corresponding to the mixture weights, mean, and variance that can be added on the top of the neural network with MDN layer without any other modifications, as illustrated in Fig.1. These mixture parameters are derived from θ as follows:

$$\gamma_m(\mathrm{x}_n) = \frac{\exp(\theta_m^{\gamma})}{\sum_{m=1}^{M}\exp(\theta_m^{\gamma})} \tag{9}$$

$$\sigma_m(\mathrm{x}_n) = \exp(\theta_m^{\sigma}) \tag{10}$$

$$\mu_m(\mathrm{x}_n) = \exp(\theta_m^{\mu}). \tag{11}$$

The conditional density function ϕ_m is represented in a Gaussian form as follows:

$$\phi_m(\hat{y}_n|\mathrm{x}_n) = \frac{1}{(2\pi)^{c/2}\sigma_m(\mathrm{x}_n)^c}\exp\frac{\|\hat{y}_n - \mu_m(\mathrm{x}_n)\|^2}{2\sigma_m(\mathrm{x}_n)^2}, \tag{12}$$

where c is the number of outputs of the MARNN model that is defined by the width of the decoder dense layer.

Training of the MDNs on top of the MARNN is implemented with standard back-propagation through time algorithm, and it aims to maximize the log-likelihood of the linear combination of the kernel functions, which is equal to minimizing the negative logarithm of the likelihood:

$$\log\mathcal{L}(\theta) = -\log\big(p_{\theta}(\hat{y}_n|\mathrm{x}_n)\big) = -\log\Big(\sum_{m=1}^{M}\gamma_m(\mathrm{x}_n)\phi_m(\hat{y}_n|\mathrm{x}_n)\Big). \tag{13}$$

Thus, the output of the MDN prediction consists of $(c+2)M$ outputs and is further approximated as a Gaussian normal distribution $\mathcal{N}(\mu_{\theta}, \sigma_{\theta}^2)$ whose mean and variance are defined as follows:

$$\mu_{\theta}(\mathrm{x}_n) = \frac{1}{M}\sum_{m=1}^{M}\gamma_m(\mathrm{x}_n)\mu_m(\mathrm{x}_n), \tag{14}$$

$$\sigma_{\theta}^2(\mathrm{x}_n) = \frac{1}{M}\sum_{m=1}^{M}\gamma_m(\mathrm{x}_n)\Big(\sigma_m^2(\mathrm{x}_n) + \big\|\mu_m(\mathrm{x}_n) - \mu_{\theta}(\mathrm{x}_n)\big\|^2\Big). \tag{15}$$

C. Monte Carlo Dropout

The main idea behind the dropout in a neural network is known as a stochastic regularization technique that is used to prevent the network overfitting to the training data by randomly switching off a subset of the hidden neurons during the training with a given probability [40]. However, recent findings in [41] suggest that the dropout could also be leveraged as an approximation of a probabilistic Gaussian process to evaluate the model uncertainty with respect to an observed sample.

The theoretical grounding of the new dropout variant is based on finding the posterior distribution over the model parameter space ω following normal prior distributions of a function $f^{\omega}(\cdot)$ that defines the neural network model architecture. Intractable in general, this target is assessed by approximating the variational distribution of the parameter space $q(\omega)$ with a mixture of Gaussians with small variances and the mean of one Gaussian fixed at zero, and then by averaging this approximation with MC integration. For the sample $\hat{\omega}_z \sim q(\omega)$, the prediction (probabilistic model likelihood) of a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2019.2952195, IEEE Journal on Selected Areas in Communications

new model output $\hat{y}_n$ given new input x_n at test time is defined in [42] by

$$p(\hat{y}_n|\mathrm{x}_n,\mathrm{X},\mathrm{Y}) \approx \int p(\hat{y}_n|\mathrm{x}_n,\omega)q(\omega)d\omega \tag{16}$$

$$\approx \frac{1}{Z}\sum_{z=1}^{Z} p(\hat{y}_n|\mathrm{x}_n,\hat{\omega}_z), \tag{17}$$

where Z is the number of variation parameters in ω and X,Y is a set of prior observations. The expectation of $\hat{y}_n$ defines the predictive mean of the model, while its variance represents the predictive uncertainty.

Here in order to obtain this uncertainty, we apply the MC dropout for G times with certain probability p for each fully-connected layer at the test time and collect the outputs of the MDN predictions. At each test step, the mean of the prediction $\mu_{\omega_g}(\mathrm{x}_n)$ is defined according to (14). Then, the variance of model uncertainty can be estimated from the variance of mean of the MDN predictions of the trained network:

$$\sigma^2_{\omega}(\mathrm{x}_n) = \frac{1}{G-1}\sum_{g=1}^{G}\Big(\mu_{\omega_g}(\mathrm{x}_n) - \mu_{\omega}(\mathrm{x}_n)\Big)^2, \tag{18}$$

where $\mu_{\omega_g}(\mathrm{x}_n) \sim f^{\omega}(\mathrm{x}_n)$ and $\mu_{\omega}(\mathrm{x}_n)$ is the mean of all G outputs that is defined as follows:

$$\mu_{\omega}(\mathrm{x}_n) = \frac{1}{G}\sum_{g=1}^{G}\mu_{\omega_g}(\mathrm{x}_n). \tag{19}$$

The MC dropout in the attention-based RNN model corresponds to randomly dropping the attention head in the sequence, and can be interpreted as forcing the model not to rely on some attention heads for its task.

D. Variational Multi-Attention Recurrent Neural Network

Incorporating the MDNs and MC dropout in the structure of MARNN allows extending the capabilities of the latter to implement a variational MARNN and assess the probabilistic forecasting [42]. The regression mean of this network $f^{\omega}(\mathrm{x}_n)$ for input sequence x_n can be defined as the mean of MDN and MC dropout predictions as follows:

$$\hat{y}_n = f^{\omega}(\mathrm{x}_n) = \mu_{\mathrm{total}}(\mathrm{x}_n) = \frac{1}{2}\Big(\mu_{\omega}(\mathrm{x}_n) + \mu_{\theta}(\mathrm{x}_n)\Big). \tag{20}$$

Under the assumption of statistical independence of the estimation error and noise, the variance of the total prediction errors can be obtained through the summation of the variance of model uncertainty $\sigma^2_{\omega}(\mathrm{x}_n)$ and the variance of noise $\sigma^2_{\theta}(\mathrm{x}_n)$:

$$\sigma^2_{\mathrm{total}}(\mathrm{x}_n) = \sigma^2_{\omega}(\mathrm{x}_n) + \sigma^2_{\theta}(\mathrm{x}_n). \tag{21}$$

Algorithm 1 summarizes the process of finding these parameters via probabilistic forecasting with the variational MARNN. Then, these parameters are used to define the upper bounds $U^{\delta}_n(\mathrm{x}_n)$ and the lower bounds $L^{\delta}_n(\mathrm{x}_n)$ of the PI by the following group of equations:

$$\begin{cases} L^{\delta}_n(\mathrm{x}_n) = \mu_{\mathrm{total}}(\mathrm{x}_n) - z_{1-\delta/2}\sqrt{\sigma^2_{\mathrm{total}}(\mathrm{x}_n)} \\ U^{\delta}_n(\mathrm{x}_n) = \mu_{\mathrm{total}}(\mathrm{x}_n) + z_{1-\delta/2}\sqrt{\sigma^2_{\mathrm{total}}(\mathrm{x}_n)} \end{cases}, \tag{22}$$

where $z_{1-\delta/2}$ is the standard normal distribution critical value that depends on the selected tail confidence level δ. This level is defined by the prediction interval nominal confidence (PINC) that corresponds to the expectation of $\hat{y}_n$ to be within the PIs limits $[L^{\delta}_n(\mathrm{x}_n), U^{\delta}_n(\mathrm{x}_n)]$ with the nominal probability $100(1-\delta)$ %:

$$\mathbb{E}\Big(\hat{y}_n \in [L^{\delta}_n(\mathrm{x}_n), U^{\delta}_n(\mathrm{x}_n)]\Big) = 100(1-\delta)\ \%. \tag{23}$$

Algorithm 1 Prediction process with variational MARNN

```
Input: input sequence $x_n$, trained variational MARNN model
    $f^{\hat{\omega}}(\cdot)$, MC dropout probability $p$, number of iterations $G$
Output: prediction mean $\hat{y}$, variance $\sigma^2_{total}(x_n)$
  // Compute MDN prediction:
  $\theta \leftarrow \{\gamma_m, \mu_m, \sigma_m\}_{m=1}^{M} \leftarrow f^{\hat{\omega}}(x_n)$
  // Split up the mixture parameters:
  $\gamma_m(x_n) \leftarrow \frac{\exp(\theta^{\gamma}_m)}{\sum_{m=1}^{M}\exp(\theta^{\gamma}_m)}$
  $\sigma_m(x_n) \leftarrow \exp(\theta^{\sigma}_m)$
  $\mu_m(x_n) \leftarrow \exp(\theta^{\mu}_m)$
  // Compute mean and variance of MDN prediction:
  $\mu_\theta(x_n) \leftarrow \frac{1}{M}\sum_{m=1}^{M}\gamma_m(x_n)\mu_m(x_n)$
  $\sigma^2_\theta(x_n) \leftarrow \frac{1}{M}\sum_{m=1}^{M}\gamma_m(x_n)\Big(\sigma^2_m(x_n)+\big\|\mu_m(x_n)-\mu_\theta(x_n)\big\|^2\Big)$
  // Compute mean and variance with MC dropout:
  for $g = 1$ to $G$ do
      $\mu_{\omega_g}(x_n) \leftarrow steps(2:5) \leftarrow f^{\hat{\omega}}\Big(x_n|MCdropout(p)\Big)$
  end for
  $\mu_\omega(x_n) \leftarrow \frac{1}{G}\sum_{g=1}^{G}\mu_{\omega_g}(x_n)$
  $\sigma^2_\omega(x_n) \leftarrow \frac{1}{G-1}\sum_{g=1}^{G}\Big(\mu_{\omega_g}(x_n) - \mu_\omega(x_n)\Big)^2$
  // Compute total mean and variance:
  $\hat{y}_n \leftarrow f^{\omega}(x_n) \leftarrow \mu_{total}(x_n) \leftarrow \frac{1}{2}\Big(\mu_\omega(x_n) + \mu_\theta(x_n)\Big)$
  $\sigma^2_{total}(x_n) \leftarrow \Big(\sigma^2_\omega(x_n) + \sigma^2_\theta(x_n)\Big)$
  return $\hat{y}$, $\sigma^2_{total}(x_n)$
```

IV. Case Study

This section presents the input data, benchmark models, optimization of the model hyper-parameters, implementation details, and evaluation indexes that were used to comprehensively study the performance of the variational MARNN for the PI forecasting of BESS SOC.

A. Battery Energy Storage System State-of-Charge Modeling

The evaluation of the model has been carried out in three different regulatory environments corresponding to the PFC by BESS in CE, NE, and GB European synchronous areas. For each of the cases, a simple BESS model was applied to simulate the BESS SOC according to (3) and the PFC droop curve parameters set by the regulatory rules in the area. The real frequency measurements from the three areas for the period of four years (2015 - 2018) served as an input for the BESS model. The frequency data are publicly available and can be accessed for an evaluation [19]–[21]. The original time resolution for the datasets are 0.01, 1, and 10 seconds. Prior to the simulation, the frequency measurements were resampled via linear interpolation to a resolution of 1 second. The simulation was conducted assuming that the BESS under control does not cause the frequency deviation. The BESS power-to-energy ratio was set to 1 as it is one of the most common ratios for the PFC according to [43]. The discharge and charge BESS efficiency was equal to 98.5 %, and no degradation was considered in order to simulate the average battery response for every new measurement. Also, an assumption was made that the full activation time is set to less than one second.

For the BESS SOC modeling in this study, the PFC, wide enhanced frequency response (EFR-Wide), and FCR for Normal operation (FCR-N) characteristics of the frequency response droop curve in Germany, Great Britain, and Finland, respectively, were utilized. These characteristics are summarized in Table I and reflect three different patterns of the droop-curve characteristics. In the CE – PFC, the DB is the lowest, while the allowable frequency deviation is between the highest in GB – EFR-Wide and the lowest in NE – FCR-N. Moreover, the characteristics of GB – EFR-Wide correspond to the highest values for the frequency deviation and deadband, while NE – FCR-N has the highest DB and the lowest deviation.

TABLE I: Characteristics of the selected frequency response reserves in the studying areas [29]

Parameter	CE – PFC	GB – EFR-Wide	NE – FCR-N
FAFD	±200 mHz	±500 mHz	±100 mHz
DB	±10 mHz	±50 mHz	±50 mHz

The outcome of the simulation was three time series consisting of BESS SOC data with one-second resolution for the period of four years that were further re-sampled to an hourly sum of BESS SOC. The characteristics of these datasets can be seen in Fig. 2. The autocorrelation and partial autocorrelation plots of the datasets demonstrate that the 24-hour lag values are statistically significant. However, the scales and duration of the positively correlated spikes vary from the GB with the lowest values and shortest period to the CE with the highest correlation and longest period. Moreover, the histograms of the datasets illustrate the difference of underlying frequency distribution of continuous BESS SOC data. According to the densities of the areas, the amount of under-frequency hours that correspond to the negative sign of BESS SOC and power injection into the grid is prevailing over the over-frequency hours. Moreover, for the case of NE, many of the hours are fluctuating at the values close to zero. The deviation of most SOC values for GB and CE is within 10 % per hour, while for the NE, these extreme values are closer to 50 %. For hourly distribution of the values, it can be noticed that it is relatively stable during the day in the GB area with minor negative deviation of the median at the morning hours and variations during the daily hours. In contrast, the other areas have more unique hours with different median, interquantile range and variations between the maximum and minimum values.

Thus, these data representations demonstrate that the presented areas have different BESS SOC distributions that reflect the difference of regulatory environments of frequency control and overall properties of power system dynamics in the areas. A study of the variational MARNN model performance on all of the datasets is crucial in order to understand its uniformity to the applications of BESS SOC forecasting. In specific, it provokes the questions of the attention performance with different correlation levels at the multi-attention lag hours as well as the ability of MDNs to capture diverse distributions.

B. Feature Selection

Feature selection for the forecasting of BESS SOC is challenged by a large amount of factors that have an influence on the frequency in a specific synchronous area. The major circumstances include but are not limited to traditionally wide spatial characteristics of interconnected macrogrids, possibly different regulatory requirements for frequency regulation in the macrogrids, diverse generation and consumption mix, multiple direct current links between the areas. In this scenario, mining supporting features such as weather or market data from sub-areas is not always feasible, especially for a large synchronous areas such as CE. Consequently, a direct choice of features is restricted to datetime features, and derivatives of frequency and BESS SOC data. The latter two demonstrate the highest feature importance for the forecasting of BESS SOC even in comparison with the market data according to the study in [44] where the above-mentioned features were evaluated for point forecasting of BESS SOC. Here, the inputs used for forecasting the BESS SOC delta between the consecutive hours for the t-th hour of the next day, ΔS_t, included time, BESS SOC data, and frequency features. Their descriptions are listed in Table II, and their correlations with the target ΔS_t are visualized in Fig. 3. The datasets are publicly available for examination in [45].

TABLE II: Inputs for a day-ahead BESS SOC forecast at the t-th hour

Input	Input description
$S_{t-48} \ldots S_{t-168}$[a]	Hourly BESS SOC that is 48, 72, 96, 120, 144, 168 hours prior to the t-th hour
ΔS_{t-48}	Hourly difference of BESS SOC that is 48 hours prior to the t-th hour
$\overline{F}_{t-48}$	Hourly mean frequency that is 48 hours prior to the t-th hour
$\sum N_{t-48}^{1\text{up}}$, $\sum N_{t-48}^{2\text{up}}$	Sum of the number of seconds when the frequency was above daily single and double standard deviation value at the hour that is 48 hours prior to the t-th hour
$\sum N_{t-48}^{1\text{down}}$, $\sum N_{t-48}^{2\text{down}}$	Sum of the number of seconds when the frequency was below daily single and double standard deviation value at the hour that is 48 hours prior to the t-th hour
$\overline{\sum N}_{t-48}^{\text{up}}$, $\overline{\sum N}_{t-48}^{\text{down}}$	Mean of positive $\sum N_{t-48}^{1\text{up}}$, $\sum N_{t-48}^{2\text{up}}$ and negative $\sum N_{t-48}^{1\text{down}}$, $\sum N_{t-48}^{2\text{down}}$ sums, respectively
$H_t^{\sin}$, $H_t^{\cos}$	Sine and cosine function of hourly values

[a] For the MARNN model, only S_{t-48} was included as an input feature of shifted BESS SOC, and the rest of lag features were expected to be retrieved by the model algorithm.

The forecasting period was selected based on the current structure of liberalized electricity markets and battery economic dispatch in a multi-objective environment. Here it is assumed that for the day-ahead multi-service optimization,

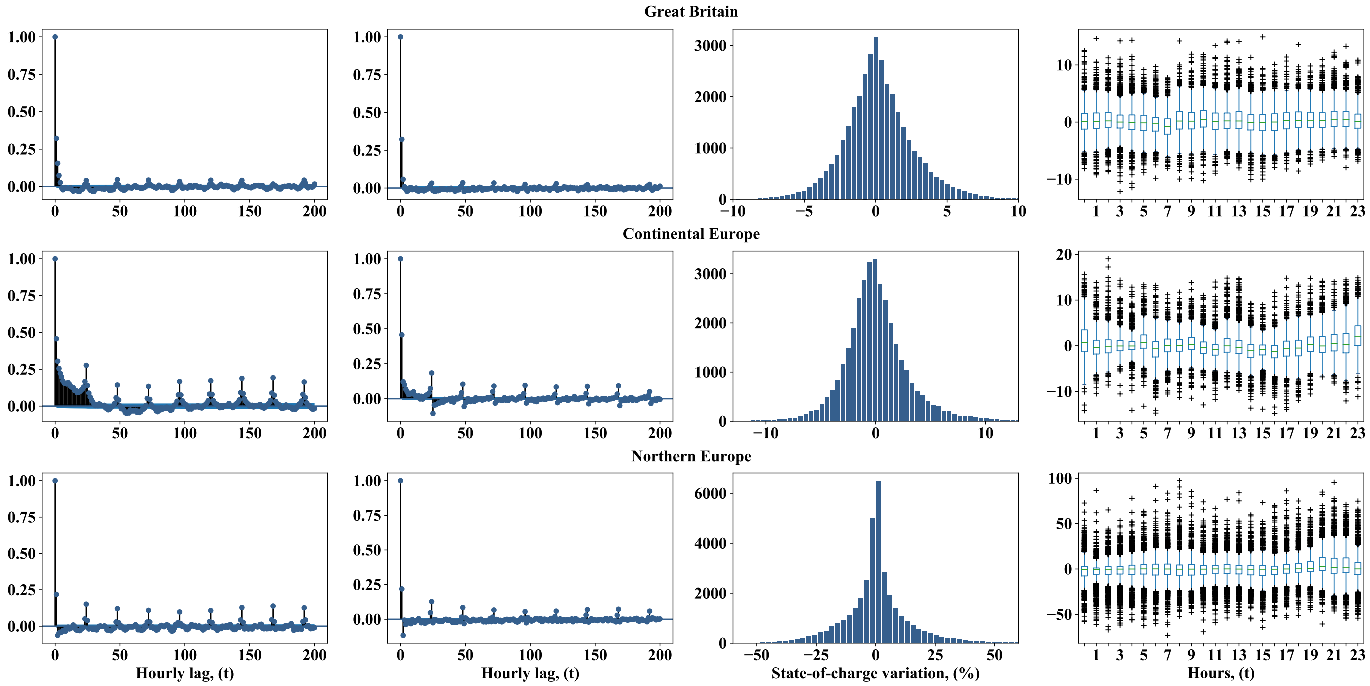


Fig. 2: Characteristics of hourly BESS SOC data in Great Britain, Continental Europe, and Northern Europe synchronous areas. From left to right: autocorrelation plot, partial correlation plot, histogram, and boxplot.

the optimal bidding strategy for the next day should be prepared before the closure in the day-ahead wholesale market (typically at 12h 00). Thus, the objective was forecasting the BESS SOC with one-hour resolution for a sequence of 36 hours ahead. Consequently, all of the features except the time are shifted by at least 48 hours due to the highest correlation at 24-th hourly lags and the need to exclude future values from the inputs for the forecasts from 24 to 36 hour steps. An example timeline for the forecasting of day-ahead BESS SOC delta ΔS_t is illustrated in Fig. 4. For the MARNN model, the input for the prediction of $t + P$ step ahead, where t is the current hour, consists of a sequence of $T = 48$ data points with feature dimension and includes the values for the period from $t - 96 + P$ to $t - 48 + P$. For the benchmark models, the input at every prediction step is the first value of the sequence at $t - 48 + P$ with dimension of the features.

According to Fig. 3, in most of the cases, the correlation of the target variables with the features is relatively low and not exceeding 10 %. The highest correlation with the target is provided by shifted BESS SOC S_{t-48}, and this correlation is expected to be lower for the more distant shifts, not presented in Fig. 3. Apart from Continental Europe, most of the features are not well correlated with the time features. Concurrently, these BESS SOC and frequency features are well correlated among each other.

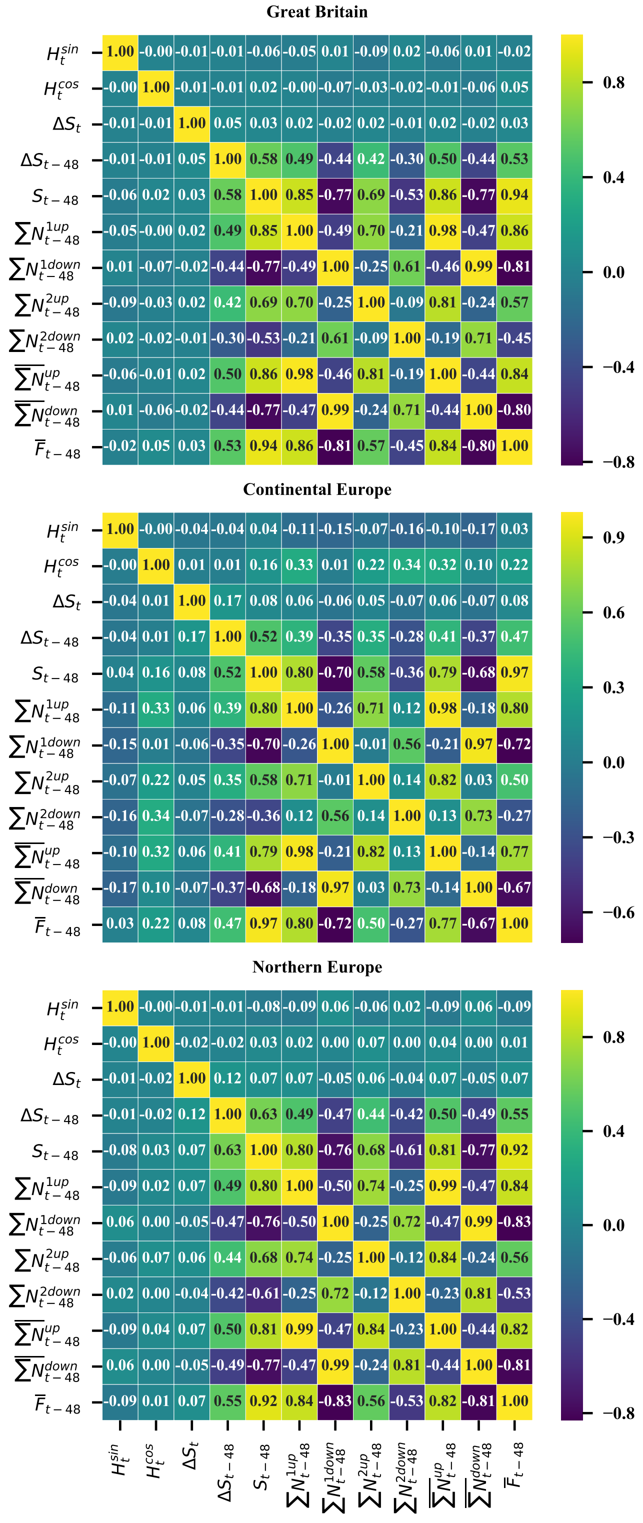
C. Benchmark

In order to evaluate the performance of the variational MARNN model with representative benchmark, it was compared with the following models that adopt quantile regression for construction of prediction intervals:

– Linear Quantile Regression (LQR) is a variation of linear least-squares regression that models not the conditional mean of the response variable but the conditional τ-th quantile of the response variable [46].
– Quantile Regression Forests (QRF) is a generalization of random forests that enables estimation of conditional quantiles for high-dimensional response variables. In contrast to the random forest that contains only the conditional mean of the observations that fall into the tree node, QRF expands this node to keep the value of all observations and returns the full conditional distribution of response value in its prediction [47].
– Quantile Gradient Boosting (QGB) is a modification of gradient boosting algorithm where a quantile loss function is used as a loss function for a gradient calculation to adjust the target of consecutive weak learner [48], [49].
– Quantile Regression Neural Network (QRNN) is an extension of the neural networks with quantile regression loss function for the estimation of the predictive distribution via conditional quantiles [50]. In this study, QRNN was implemented by a shallow neural network with two wide fully-connected layers and ReLU activation function.

The listed quantile regression models were extensively used in many Global Energy Forecasting Competitions [51], [52] and can be considered as a state-of-the-art in the topic of probabilistic energy forecasting.

In quantile regression the τ-th quantile level is defined as the value below which the proportion of the conditional response population is τ for $\tau \in (0,1)$. The limits of PI for a nominal coverage rate $1 - \delta$ are then constructed by the quantile levels $\delta/2$ and $1 - \delta/2$. The quantile loss function averaged over the whole dataset for the corresponding conditional quantile $f^{\omega}|\tau$



Great Britain

	H_t^{sin}	H_t^{cos}	ΔS_t	ΔS_{t-48}	S_{t-48}	$\sum N_{t-48}^{1up}$	$\sum N_{t-48}^{1down}$	$\sum N_{t-48}^{2up}$	$\sum N_{t-48}^{2down}$	$\overline{\sum N}_{t-48}^{up}$	$\overline{\sum N}_{t-48}^{down}$	$\overline{F}_{t-48}$
H_t^{sin}	1.00	-0.00	-0.01	-0.01	-0.06	-0.05	0.01	-0.09	0.02	-0.06	0.01	-0.02
H_t^{cos}	-0.00	1.00	-0.01	-0.01	0.02	-0.00	-0.07	-0.03	-0.02	-0.01	-0.06	0.05
ΔS_t	-0.01	-0.01	1.00	0.05	0.03	0.02	-0.02	0.02	-0.01	0.02	-0.02	0.03
ΔS_{t-48}	-0.01	-0.01	0.05	1.00	0.58	0.49	-0.44	0.42	-0.30	0.50	-0.44	0.53
S_{t-48}	-0.06	0.02	0.03	0.58	1.00	0.85	-0.77	0.69	-0.53	0.86	-0.77	0.94
$\sum N_{t-48}^{1up}$	-0.05	-0.00	0.02	0.49	0.85	1.00	-0.49	0.70	-0.21	0.98	-0.47	0.86
$\sum N_{t-48}^{1down}$	0.01	-0.07	-0.02	-0.44	-0.77	-0.49	1.00	-0.25	0.61	-0.46	0.99	-0.81
$\sum N_{t-48}^{2up}$	-0.09	-0.03	0.02	0.42	0.69	0.70	-0.25	1.00	-0.09	0.81	-0.24	0.57
$\sum N_{t-48}^{2down}$	0.02	-0.02	-0.01	-0.30	-0.53	-0.21	0.61	-0.09	1.00	-0.19	0.71	-0.45
$\overline{\sum N}_{t-48}^{up}$	-0.06	-0.01	0.02	0.50	0.86	0.98	-0.46	0.81	-0.19	1.00	-0.44	0.84
$\overline{\sum N}_{t-48}^{down}$	0.01	-0.06	-0.02	-0.44	-0.77	-0.47	0.99	-0.24	0.71	-0.44	1.00	-0.80
$\overline{F}_{t-48}$	-0.02	0.05	0.03	0.53	0.94	0.86	-0.81	0.57	-0.45	0.84	-0.80	1.00

Continental Europe

	H_t^{sin}	H_t^{cos}	ΔS_t	ΔS_{t-48}	S_{t-48}	$\sum N_{t-48}^{1up}$	$\sum N_{t-48}^{1down}$	$\sum N_{t-48}^{2up}$	$\sum N_{t-48}^{2down}$	$\overline{\sum N}_{t-48}^{up}$	$\overline{\sum N}_{t-48}^{down}$	$\overline{F}_{t-48}$
H_t^{sin}	1.00	-0.00	-0.04	-0.04	0.04	-0.11	-0.15	-0.07	-0.16	-0.10	-0.17	0.03
H_t^{cos}	-0.00	1.00	0.01	0.01	0.16	0.33	0.01	0.22	0.34	0.32	0.10	0.22
ΔS_t	-0.04	0.01	1.00	0.17	0.08	0.06	-0.06	0.05	-0.07	0.06	-0.07	0.08
ΔS_{t-48}	-0.04	0.01	0.17	1.00	0.52	0.39	-0.35	0.35	-0.28	0.41	-0.37	0.47
S_{t-48}	0.04	0.16	0.08	0.52	1.00	0.80	-0.70	0.58	-0.36	0.79	-0.68	0.97
$\sum N_{t-48}^{1up}$	-0.11	0.33	0.06	0.39	0.80	1.00	-0.26	0.71	0.12	0.98	-0.18	0.80
$\sum N_{t-48}^{1down}$	-0.15	0.01	-0.06	-0.35	-0.70	-0.26	1.00	-0.01	0.56	-0.21	0.97	-0.72
$\sum N_{t-48}^{2up}$	-0.07	0.22	0.05	0.35	0.58	0.71	-0.01	1.00	0.14	0.82	0.03	0.50
$\sum N_{t-48}^{2down}$	-0.16	0.34	-0.07	-0.28	-0.36	0.12	0.56	0.14	1.00	0.13	0.73	-0.27
$\overline{\sum N}_{t-48}^{up}$	-0.10	0.32	0.06	0.41	0.79	0.98	-0.21	0.82	0.13	1.00	-0.14	0.77
$\overline{\sum N}_{t-48}^{down}$	-0.17	0.10	-0.07	-0.37	-0.68	-0.18	0.97	0.03	0.73	-0.14	1.00	-0.67
$\overline{F}_{t-48}$	0.03	0.22	0.08	0.47	0.97	0.80	-0.72	0.50	-0.27	0.77	-0.67	1.00

Northern Europe

	H_t^{sin}	H_t^{cos}	ΔS_t	ΔS_{t-48}	S_{t-48}	$\sum N_{t-48}^{1up}$	$\sum N_{t-48}^{1down}$	$\sum N_{t-48}^{2up}$	$\sum N_{t-48}^{2down}$	$\overline{\sum N}_{t-48}^{up}$	$\overline{\sum N}_{t-48}^{down}$	$\overline{F}_{t-48}$
H_t^{sin}	1.00	-0.00	-0.01	-0.01	-0.08	-0.09	0.06	-0.06	0.02	-0.09	0.06	-0.09
H_t^{cos}	-0.00	1.00	-0.02	-0.02	0.03	0.02	0.00	0.07	0.00	0.04	0.00	0.01
ΔS_t	-0.01	-0.02	1.00	0.12	0.07	0.07	-0.05	0.06	-0.04	0.07	-0.05	0.07
ΔS_{t-48}	-0.01	-0.02	0.12	1.00	0.63	0.49	-0.47	0.44	-0.42	0.50	-0.49	0.55
S_{t-48}	-0.08	0.03	0.07	0.63	1.00	0.80	-0.76	0.68	-0.61	0.81	-0.77	0.92
$\sum N_{t-48}^{1up}$	-0.09	0.02	0.07	0.49	0.80	1.00	-0.50	0.74	-0.25	0.99	-0.47	0.84
$\sum N_{t-48}^{1down}$	0.06	0.00	-0.05	-0.47	-0.76	-0.50	1.00	-0.25	0.72	-0.47	0.99	-0.83
$\sum N_{t-48}^{2up}$	-0.06	0.07	0.06	0.44	0.68	0.74	-0.25	1.00	-0.12	0.84	-0.24	0.56
$\sum N_{t-48}^{2down}$	0.02	0.00	-0.04	-0.42	-0.61	-0.25	0.72	-0.12	1.00	-0.23	0.81	-0.53
$\overline{\sum N}_{t-48}^{up}$	-0.09	0.04	0.07	0.50	0.81	0.99	-0.47	0.84	-0.23	1.00	-0.44	0.82
$\overline{\sum N}_{t-48}^{down}$	0.06	0.00	-0.05	-0.49	-0.77	-0.47	0.99	-0.24	0.81	-0.44	1.00	-0.81
$\overline{F}_{t-48}$	-0.09	0.01	0.07	0.55	0.92	0.84	-0.83	0.56	-0.53	0.82	-0.81	1.00

Fig. 3: Correlation plot of the selected features with target variable ΔS_t for three synchronous areas: Great Britain, Continental Europe, and Northern Europe.

is determined as follows:

$$\mathcal{L}(\mathrm{y}, \mathrm{f}^{\omega}|\tau) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(y_n - f^{\omega}(\mathrm{x}_n)|\tau), \tag{24}$$

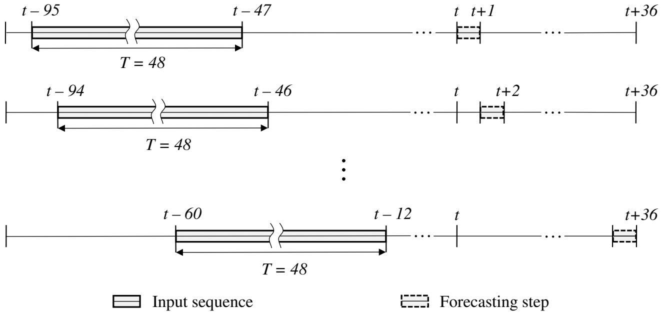


Fig. 4: Many-to-one scheme deployed by the variational MARNN for the forecasting of ΔS_t.

where $\mathcal{L}(y_n - f^{\omega}(\mathrm{x}_n)|\tau) = \mathcal{L}(\varepsilon_n|\tau)$ is the loss of individual data point that is modelled with a pinball loss function as follows:

$$\mathcal{L}(\varepsilon_n|\tau) = \begin{cases} \tau\varepsilon_n, & \text{if } \varepsilon_n \geq 0 \\ (\tau - 1)\varepsilon_n, & \text{if } \varepsilon_n < 0 \end{cases}. \tag{25}$$

D. Hyper-parameter Optimization

The hyper-parameters of the variational MARNN model and corresponding benchmarks are selected using Bayesian optimization with the Tree Parzen Estimator (TPE) hyper-parameter search [53]. The initial condition and the results of hyper-parameter search space are summarized in Table III. This optimization was primarily aimed to tune the structure of attention and decoder layers, find appropriate training settings, and define an proper number of mixtures to fit in the datasets. In the hyper-parameters, the attention length corresponds to the number of past inputs with 24-hour lag that are used to calculate the state vector in the attention layer. The 24-hour lag is selected due to the daily trends in the datasets identified in the previous steps. The number of hidden units specifies the number of units in the fully-connected dense layer of the decoder. The dropout rate is used for the densely-connected parts of attention and decoder layers of the model, as illustrated in Fig. 1. The learning rate is adjusted for the Adam optimization algorithm in the model.

TABLE III: Search space and the results of the MARNN model hyper-parameter optimization

Hyper-parameter	Search space	Distribution	Best trial		
			GB	CE	NE
Batch size	$2^7, 2^8, \ldots, 2^{11}$	Categorical	256	1024	128
Learning rate	$10^{-4} - 10^{-2}$	Loguniform	0.0014	0.0086	0.0086
Attention length	2 – 14	Quniform	5	3	9
Dropout rate	0.0 – 0.5	Uniform	0.087	0.423	0.343
Mixture number	3 – 9	Quniform	3	3	3
Hidden units	3 – 48	Quniform	30	37	30

The optimization process was running for 100 trials on each of the datasets following the sequential automatic hyper-parameter optimization presented in [54]. The inputs of the variational MARNN model were the hyper-parameters selected by the TPE at each trial and training and validation data of the datasets. The history of model loss served as an input for the optimizer and was evaluated with mean squared error of the mixture density network. Each of the trials was restricted to 20 epochs with the early stopping criterion equal to 5. More details about TPE optimizer can be found in [53] and its application to the MARNN model is described in [54].

Density plots of the TPE optimization trials on the hyper-parameters are illustrated in Fig. 5. From them the effects of these parameters on the model performance can be retrieved. According to the results, some of the hyper-parameters are identical to all the datasets while some are different for each case. For instance, the optimizer demonstrates that three mixtures and more than ten times higher amount of hidden units in the decoder dense layer is enough to reflect the distribution of BESS SOC data despite its diversity in the synchronous areas. In contrast, the dropout rate can vary from 10 % to 40 %, and it does not enable formation of any general conclusions about the best practices. Also, the attention length of the models has diverse patterns for the highest concentration of trials. For the GB area, the attention number was fluctuating around the lowest boundary of the search space, for the NE area, it was just above the average, and CE had trials in both but primarily close to the lowest one. These results can be interpreted with the autocorrelation data in Fig. 2 and summarized as the attention length is negatively proportional to the correlation of attention heads. Moreover, the optimal learning rate resides close to the lowest boundary despite that in the case of CE and NE areas, the best trial was at the higher rate. As for the batch size, the best results for the datasets of comparable length are expected from the categories of 128, 256 or 1048. Besides the above mentioned hyper-parameters, the MARNN model also has the number of inputs, the number of hidden units in the fully-connected layer of attention, and the number of encoder and decoder units. These hyper-parameters were not optimized but chosen as follows: the input is equal to a sequence of 48 data points with arbitrary feature dimension; the number of hidden units in the fully-connected attention layer and the number of encoder units were equal to length of the input sequence; the number of decoder cells is equal to the number of attention heads.

The hyper-parameters of the benchmark models are presented in Table IV. For these models, the loss function during the hyper-parameter optimization was the average quantile loss over the validation set at 0.5 quantile. The results of their optimization are out of scope of this study.

E. Model Implementation Details

The MARNN model was implemented using Keras 2.1.5 high-level neural networks API [55] with Tensorflow 1.14.0 [56] as the backend in Python 3.7 environment. The MARNN model was developed based on [57], the MC dropout was added to the model with astroNN package [58] and MDN layer was built on top of the MARNN with [59]. The fast GRU implementation backed by NVIDIA CUDA Deep Neural Network library (cuDNN) [60] was used for the encoder and decoder RNNs.

TABLE IV: Search space for hyper-parameter optimization of the benchmark models

Hyper-parameter	Search space	Distribution	Benchmark model		
			QRF	QGB	QRNN
Max depth	10 – 110	Quniform	-	X	-
Max leaf nodes	10 – 110	Quniform	-	X	-
Min samples split	2 – 10	Quniform	-	X	-
Min samples leaf	1 – 10	Quniform	-	X	-
Estimators	100 – 1000	Quniform	X	X	-
Bootstrap	True, False	Categorical	X	-	-
Learning rate 1	$10^{-4} - 3 \cdot 10^{-1}$	Loguniform	-	X	-
Batch size	$2^7, 2^8, \ldots, 2^{11}$	Categorical	-	-	X
Hidden units	36 – 512	Quniform	-	-	X
Learning rate 2	$10^{-4} - 10^{-2}$	Loguniform	-	-	X

The target was chosen as a difference between the consecutive hours ΔS_t in order to remove the hourly autocorrelation that generally led to a persistence model. Moreover, MinMax scaling with the range from 0 to 1 was utilized for the datasets. Finally, the model validation was carried out using hold-out method, in which the dataset was split for training, parameter regulation, and performance evaluation in proportions of 50 %, 25 % and 25 %, which is approximately equal to two years of training and one year for validation and testing, respectively. The model was trained for 100 epochs with 20 as the early stopping criterion. The number of MC dropout iterations was limited to 200.

In this study, LQR model was implemented with QuantReg function of statsmodel package [61]. Also, a variance inflation factor was utilized to remove collinear features from the datasets prior the LQR modeling. QRF and QGB models were created using Random Forest and Gradient Boosting Regressors from the scikit-learn library [62]. In order to obtain QRF from Random Forest Regressor, the minimum leaf node hyper-parameter was set to one. The QGB was modelled with explicit quantile prediction. The QRNN was developed with Keras library using sequential model architecture. Examples of the benchmark models are available in [63].

The automatic hyper-parameter optimization of the models was built with Hyperopt library [64]. Also, Hyperas package [65] that is a wrapper over Hyperopt library was utilized for the variational MARNN model hyper-parameter optimization.

F. Prediction Interval Evaluation Indexes

In this study, the performance of the probabilistic forecasting is quantitatively evaluated based on the resolution and the reliability of the constructed PIs, as described in [66]. The reliability of the PI for N samples is illustrated by the prediction interval coverage probability (PICP) that is defined as follows:

$$PICP = \frac{1}{N} \sum_{n=1}^{N} I_n^{\delta}, \tag{26}$$

where I_n^{δ} is the PICPs index:

$$I_n^{\delta} = \begin{cases} 1, & \hat{y}_n \in [L_n^{\delta}(\mathrm{x}_n), U_n^{\delta}(\mathrm{x}_n)] \\ 0, & \hat{y}_n \notin [L_n^{\delta}(\mathrm{x}_n), U_n^{\delta}(\mathrm{x}_n)] \end{cases}. \tag{27}$$

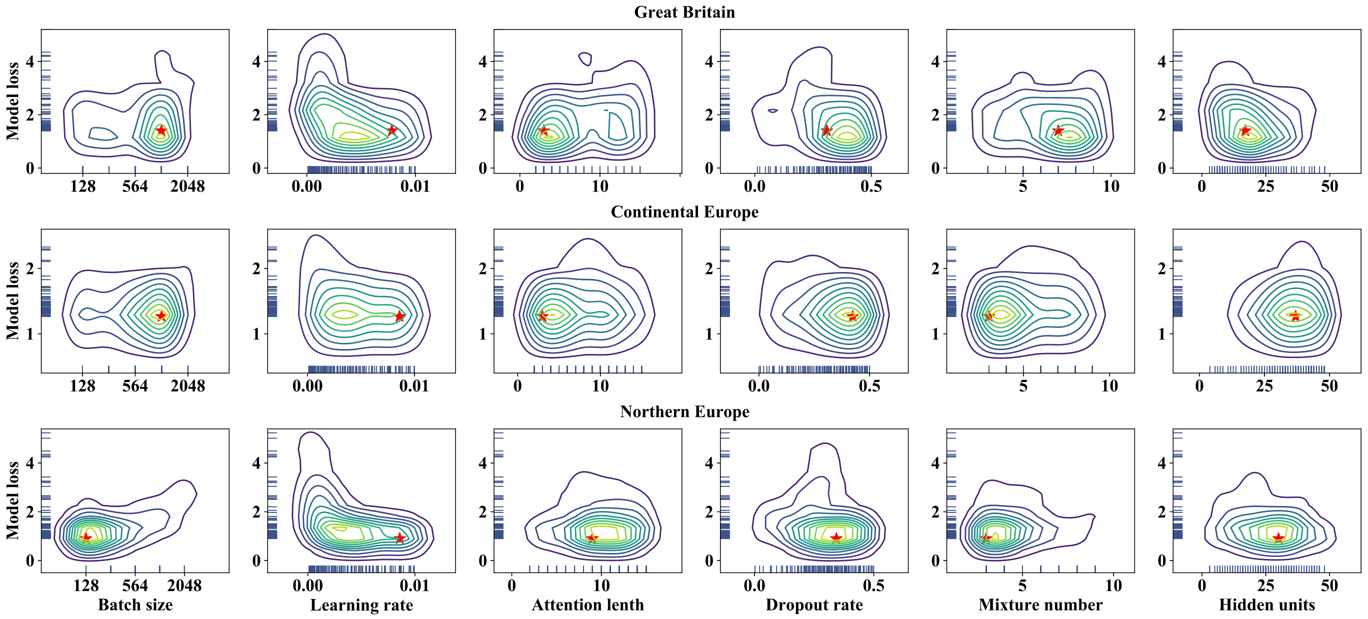


Fig. 5: The effects of hyper-parameter optimization on the variational MARNN model loss: the highest concentration of the optimization trials is marked with light color tint and the lowest with dark color tint. The best trial is marked with a red star. The values of the hyper-parameters that led to extreme model loss are excluded from the visualization.

The purpose of the resolution metric is to evaluate the interval sharpness to restrict the possibility to obtain a high reliability PIs with increased width. This metric is expressed in the PI normalized average width (PINAW) defined as follows:

$$PINAW = \frac{1}{R \cdot N}\sum_{n=1}^{N}\left[U_n^{\delta}(\mathrm{x}_n) - L_n^{\delta}(\mathrm{x}_n)\right], \tag{28}$$

where $R = \max(\mathrm{Y}) - \min(\mathrm{Y})$ is the difference between the bounds of the main targets. A narrower PIs correspond to a smaller PINAW and, hence, demonstrates better sharpness.

To jointly assess the coverage and interval width, a coverage width-based criterion (CWC) is applied. For a confidential probability $100(1-\delta)$ % denoted as ξ, it is defined as

$$CWC = PINAW(1 + \nu e^{-\lambda(PICP-\xi)}) \tag{29}$$

$$\nu = \begin{cases} 0, & PICP \geq \xi \\ 1, & PICP < \xi \end{cases}, \tag{30}$$

where ν represents a forecast score used for penalizing if the PI coverage is lower than the required confidence level. The penalty coefficient is defined by λ and set to 10 in this study.

V. Results and Discussion

The results of the performance evaluation are shown in Fig. 6 and Tables V, VI, VII. Each sub-plot of the figure contains the real and forecasted BESS SOC information related to a random 36-hour forecast interval from the test data. The forecasting results are represented in the form of PIs with confidence levels from 10 % to 95 %. The PIs are consecutively illustrated for LQR,QRF, QGB, QRNN, and variational MARNN for each of the datasets. The best results for the evaluation indexes are marked in bold in the Tables. The logic behind the best index evaluation is the following: the best PICP is the minimum PICP index that is above the required confidence level or, if this condition is not satisfied, the closest to the required coverage; The best PINAW and CWC indexes are chosen as the lowest from those models whose PICP is above the required confidence level, or otherwise from the models whose PICP is above or equal to the previous required coverage level(s).

In the GB synchronous area, the variational MARNN model demonstrated superiority over the other models despite the lowest correlation of hourly lag values among the investigated areas and, hence, the lowest expectations for the MARNN model performance. This predominance can be explained by generally too narrow PIs of the quantile regression models and wide coverage of the MARNN model. For example, none of the quantile regression models were able to provide the required coverage for the PIs, and, hence, a penalty score was applied in the CWC index that raised the index values several times compared to PINAW. In contrast, the coverage of the variational MARNN was exceeding the required level by 5.5 % on average. However, visual representation of the model performances also suggests that the LQR model had good sharpness around the true mean of BESS SOC values, while the QRNN had passive variation and low prediction capabilities. Among the ensemble models, QRF had slightly better results than the QGB model.

The performances of all the models were at the high level in the CE area. These results can be explained by good correlation of the BESS SOC values seen in Fig. 2 and Fig. 3. Almost optimal indexes were shown by the LQR model with average PICP error of only 0.4 %. However, even in this scenario, there were several cases when the QGB and QRNN models were better. Moreover, the QRF model was the only to achieve the required coverage in all of the intervals. The

TABLE V: Performance evaluation of the prediction interval forecasts for Great Britain

$\mathbb{E}$,	LQR			QRF			QGB			QRNN			Variational MARNN		
[%]	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC
10	0.076	0.021	0.049	0.081	0.022	0.049	0.071	0.019	0.045	0.024	0.006	0.019	**0.127**	**0.033**	**0.033**
20	0.146	0.042	0.115	0.167	0.045	0.108	0.145	0.039	0.106	0.160	0.043	0.108	**0.251**	**0.068**	**0.068**
30	0.228	0.065	0.200	0.260	0.070	0.174	0.224	0.060	0.189	0.192	0.052	0.206	**0.372**	**0.104**	**0.104**
40	0.303	0.090	0.327	0.350	0.096	0.255	0.308	0.083	0.292	0.301	0.082	0.304	**0.482**	**0.141**	**0.141**
50	0.387	0.118	0.484	0.440	0.126	0.357	0.389	0.109	0.437	0.312	0.085	0.642	**0.590**	**0.181**	**0.181**
60	0.478	0.150	0.658	0.535	0.161	0.468	0.477	0.138	0.612	0.442	0.127	0.744	**0.692**	**0.226**	**0.226**
70	0.578	0.188	0.821	0.640	0.203	0.574	0.573	0.173	0.791	0.554	0.168	0.894	**0.779**	**0.279**	**0.279**
80	0.684	0.238	1.000	0.747	0.258	0.699	0.686	0.222	0.916	0.630	0.198	1.284	**0.862**	**0.344**	**0.344**
90	0.812	0.320	1.084	0.857	0.345	0.878	0.809	0.296	1.033	0.810	0.298	1.033	**0.934**	**0.442**	**0.442**
95	0.888	0.397	1.137	0.917	0.423	1.012	0.885	0.365	1.068	0.875	0.356	1.109	**0.966**	**0.527**	**0.527**

TABLE VI: Performance evaluation of the prediction interval forecasts for Continental Europe

$\mathbb{E}$,	LQR			QRF			QGB			QRNN			Variational MARNN		
[%]	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC
10	0.099	0.018	0.036	0.110	0.020	0.020	**0.101**	**0.018**	**0.018**	0.035	0.007	0.021	0.136	0.024	0.024
20	**0.203**	**0.037**	**0.037**	0.216	0.040	0.040	0.195	0.035	0.073	0.127	0.023	0.070	0.262	0.048	0.048
30	**0.305**	**0.056**	**0.056**	0.326	0.060	0.060	0.294	0.053	0.110	0.327	0.061	0.061	0.376	0.073	0.073
40	**0.403**	**0.078**	**0.078**	0.424	0.083	0.083	0.390	0.073	0.154	0.352	0.069	0.181	0.490	0.099	0.099
50	**0.503**	**0.101**	**0.101**	0.531	0.108	0.108	0.490	0.095	0.199	0.438	0.085	0.244	0.594	0.127	0.127
60	**0.606**	**0.129**	**0.129**	0.635	0.138	0.138	0.581	0.119	0.263	0.608	0.130	0.130	0.687	0.158	0.158
70	0.705	0.164	0.164	0.735	0.175	0.175	0.685	0.151	0.325	**0.701**	**0.163**	**0.163**	0.774	0.195	0.195
80	**0.805**	**0.213**	**0.213**	0.832	0.225	0.225	0.777	0.191	0.432	0.810	0.218	0.218	0.845	0.241	0.241
90	0.908	0.297	0.297	0.919	0.307	0.307	0.886	0.259	0.558	**0.906**	**0.291**	**0.291**	0.913	0.309	0.309
95	**0.955**	**0.377**	**0.377**	0.961	0.387	0.387	0.941	0.332	0.693	0.965	0.408	0.408	0.948	0.369	0.745

TABLE VII: Performance evaluation of the prediction interval forecasts for Northern Europe

$\mathbb{E}$,	LQR			QRF			QGB			QRNN			Variational MARNN		
[%]	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC
10	0.094	0.011	0.022	0.109	0.011	0.011	**0.106**	**0.010**	**0.010**	0.059	0.006	0.015	0.112	0.017	0.017
20	0.189	0.022	0.047	0.220	0.024	0.024	**0.208**	**0.021**	**0.021**	0.232	0.025	0.025	0.217	0.034	0.034
30	0.285	0.037	0.079	0.334	0.039	0.039	**0.312**	**0.034**	**0.034**	0.206	0.022	0.079	0.321	0.052	0.052
40	0.385	0.054	0.118	0.439	0.056	0.056	**0.407**	**0.049**	**0.049**	0.453	0.059	0.059	0.410	0.071	0.071
50	0.481	0.075	0.167	0.535	0.077	0.077	**0.505**	**0.068**	**0.068**	0.516	0.073	0.073	0.493	0.091	0.188
60	0.586	0.102	0.219	0.640	0.103	0.103	**0.604**	**0.091**	**0.091**	0.651	0.108	0.108	0.575	0.113	0.259
70	0.690	0.136	0.286	0.732	0.136	0.136	**0.702**	**0.121**	**0.121**	0.715	0.129	0.129	0.654	0.139	0.359
80	0.789	0.183	0.386	**0.827**	**0.181**	**0.181**	0.799	0.163	0.328	0.775	0.157	0.358	0.723	0.172	0.544
90	0.899	0.262	0.527	**0.912**	**0.252**	**0.252**	0.896	0.231	0.472	0.889	0.227	0.480	0.813	0.221	0.750
95	**0.950**	0.340	0.340	0.953	**0.321**	**0.321**	0.946	0.296	0.604	0.930	0.279	0.619	0.863	0.263	0.893

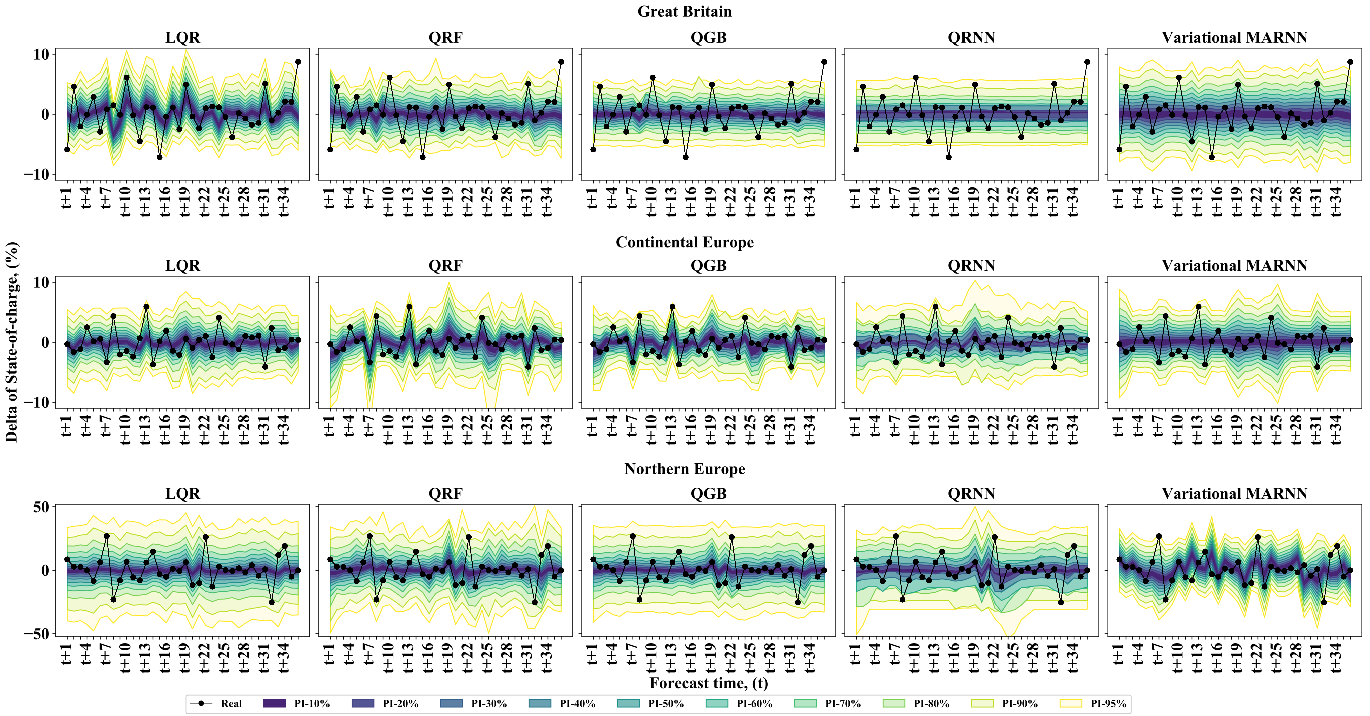


Fig. 6: An example day-ahead forecast of BESS SOC delta with prediction intervals (PIs): LQR – Linear Quantile Regression, QRF – Quantile Random Forests, QGB – Quantile Gradient Boosting, QRNN – Quantile Regression Neural Network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2019.2952195, IEEE Journal on Selected Areas in Communications

variational MARNN model provided the required coverage for all the PIs except 95 % interval where it had shown 94.8 % coverage. In the successful cases, the average exceeding of the PICP index by the MARNN model was at the level 5.8 %. The visual perspective illustrates that the LQR, QRF, and QGB models had identified well the conditional mean of the BESS SOC data.

In the case of NE area, the best indexes were demonstrated by the QGB model even though its PICP index was slightly lower the required coverage for the highest intervals (80, 90, and 95 %). In these intervals, the QRF was more robust to provide the required coverage. The indexes of LQR were generally lower the required levels, and the QRNN had variable success in coverage of the data. The MARNN model was able to provide the required coverage for only less than half of the intervals having trouble to catch the data points in the highest intervals. The maximum decrease of the MARNN PICP in comparison with the nominal confidence level is within 8.7 %. The reason of such indexes can be seen in high sharpness and narrowness of the MARNN intervals that is demonstrated in Fig. 6. Also, the MARNN model was superior in predicting the conditional mean of the BESS SOC data.

In general, the MARNN model has shown good performance in respect to the coverage probability for the different regulatory environments and if compared to the performance of quantile regression algorithms. In particular, according to the PICP indexes, its coverage probability achieved the required confidence levels in most of the intervals for the GB and CE areas, but was relatively low for the NE area. However, in the CE and GB areas, the MARNN model follows the true mean regression worse that in the NE are, and this can be seen in Fig. 6. Nevertheless, this approach gives better PICP and CWC in the shortcoming areas of the quantile regression algorithms compared.

VI. Conclusion and Future Work

In this study, a hybrid probabilistic model with combined MDNs and MC dropout over MARNN was presented and deployed in order to forecast BESS SOC under the PFC. This approach allows rigorously quantifying the overall forecasting uncertainty related to the inherent data noise and model estimation error in the form of PIs with a particular confidence level. Moreover, the hybrid MDN and MC dropout approach is extremely generic and can be easily applicable to any existing neural networks. The performance of the model was evaluated for three different regulatory environments corresponding to the PFC by BESS in CE, NE, and GB European synchronous areas and compared with state-of-the-art quantile regression algorithms in probabilistic energy forecasting such as the LQR, QRF, QGB, and QRNN. According to the case studies, the proposed variational MARNN model has satisfactory performance and good generalization capabilities for prediction interval forecasting of the BESS SOC despite the diversity of droop curve parameters and, hence, different frequency distributions in the case areas. Therefore, the proposed approach can potentially provide an efficient and meaningful tool to hedge not only against uncertainties and risks of the BESS PFC, but it can also be leveraged in other smart grid applications to assist in the related decision making activities.

The potential future research questions include:

(1) Feasibility of the economic benefits that this hybrid forecast model may achieve in comparison with other optimization methods.

(2) Extension of the model to multivariate and simultaneous probabilistic forecasting with cross-dependency of the BESS SOC and the market forecasts as well as consideration of the BESS degradation under PFC in the BESS SOC forecast.

References

[1] R. H. Byrne, T. A. Nguyen, D. A. Copp, B. R. Chalamala, and I. Gyuk, "Energy management and optimization methods for grid energy storage systems," *IEEE Access*, vol. 6, pp. 13 231–13 260, 2017.

[2] B. Nykvist and M. Nilsson, "Rapidly falling costs of battery packs for electric vehicles," *Nature climate change*, vol. 5, no. 4, p. 329, 2015.

[3] P. Ralon, M. Taylor, A. Ilas, H. Diaz-Bone, and K. Kairies, "Electricity storage and renewables: Costs and markets to 2030," *International Renewable Energy Agency: Abu Dhabi, United Arab Emirates*, 2017.

[4] E. Namor, F. Sossan, R. Cherkaoui, and M. Paolone, "Control of battery storage systems for the simultaneous provision of multiple services," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2799 – 2808, 2018.

[5] C. Wan, Z. Xu, Y. Wang, Z. Y. Dong, and K. P. Wong, "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 463–470, 2013.

[6] 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 Transactions on Power Systems*, vol. 34, no. 2, pp. 1203–1215, 2019.

[7] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[8] R. Tahmasebifar, M. K. Sheikh-El-Eslami, and R. Kheirollahi, "Point and interval forecasting of real-time and day-ahead electricity prices by a novel hybrid approach," *IET Generation, Transmission & Distribution*, vol. 11, no. 9, pp. 2173–2183, 2017.

[9] W. Xie, P. Zhang, R. Chen, and Z. Zhou, "A nonparametric Bayesian framework for short-term wind power probabilistic forecast," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 371–379, 2019.

[10] Y.-K. Wu, P.-E. Su, T.-Y. Wu, J.-S. Hong, and M. Y. Hassan, "Probabilistic wind-power forecasting using weather ensemble models," *IEEE Transactions on Industry Applications*, vol. 54, no. 6, pp. 5609–5620, 2018.

[11] F. Golestaneh, P. Pinson, and H. B. Gooi, "Very short-term nonparametric probabilistic forecasting of renewable energy generation – With application to solar energy," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3850–3863, 2016.

[12] M. Rafiei, T. Niknam, J. Aghaei, M. Shafie-Khah, and J. P. Catalão, "Probabilistic load forecasting using an improved wavelet neural network trained by generalized extreme learning machine," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6961–6971, 2018.

[13] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic load forecasting via quantile regression averaging on sister forecasts," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 730–737, 2017.

[14] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[15] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1548–1568, 2018.

[16] D. W. Van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1484–1512, 2018.

[17] T. Mercier, "Storage-based frequency control and grid-frequency deviations forecasting," *Revue E tijdschrift*, vol. 2016, p. 1, 2016.

[18] Y. Tang, H. Cui, and Q. Wang, "Prediction model of the power system frequency using a cross-entropy ensemble algorithm," *Entropy*, vol. 19, no. 10, p. 552, 2017.

[19] National Grid ESO, "Historic frequency data," 2019, data retrieved from National Grid ESO, https://www.nationalgrideso.com/balancing-services/frequency-response-services/historic-frequency-data.

[20] RTE France, "Network frequency," 2019, data retrieved from RTE France, https://clients.rte-france.com/lang/an/visiteurs/vie/vie_frequence.jsp.

[21] Fingrid, "Frequency - historical data," 2019, data retrieved from Fingrid, https://data.fingrid.fi/en/dataset/frequency-historical-data.

[22] A. Ulbig, T. S. Borsche, and G. Andersson, "Impact of low rotational inertia on power system stability and operation," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 7290–7297, 2014.

[23] Y. Mu, J. Wu, J. Ekanayake, N. Jenkins, and H. Jia, "Primary frequency response from electric vehicles in the Great Britain power system," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1142–1150, 2012.

[24] Z. A. Obaid, L. M. Cipcigan, L. Abrahim, and M. T. Muhssin, "Frequency control of future power systems: reviewing and evaluating challenges and new control methods," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 1, pp. 9–25, 2019.

[25] S. Chen, T. Zhang, H. B. Gooi, R. D. Masiello, and W. Katzenstein, "Penetration rate and effectiveness studies of aggregated BESS for frequency regulation," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 167–177, 2016.

[26] H. T. Nguyen, "Frequency characterization and control for future low inertia systems," Ph.D. dissertation, Technical University of Denmarik, 2018.

[27] M. Świerczyński, D. I. Stroe, A.-I. Stan, R. Teodorescu, and D. U. Sauer, "Selection and performance-degradation modeling of $\text{LiMo}_2/\text{Li}_4\text{Ti}_5\text{O}_{12}$ and LiFePo_4/C battery cells as suitable energy storage systems for grid integration with wind power plants: an example for the primary frequency regulation service," *IEEE transactions on Sustainable Energy*, vol. 5, no. 1, pp. 90–101, 2013.

[28] R. Moreno, R. Moreira, and G. Strbac, "A MILP model for optimising multi-service portfolios of distributed energy storage," *Applied Energy*, vol. 137, pp. 554–566, 2015.

[29] R. Hollinger, A. M. Cortes, and T. Erge, "Fast frequency response with BESS: A comparative analysis of Germany, Great Britain and Sweden," in *2018 15th International Conference on the European Energy Market (EEM)*. IEEE, Jun. 2018.

[30] O. Mégel, J. L. Mathieu, and G. Andersson, "Scheduling distributed energy storage units to provide multiple services under forecast error," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 48–57, 2015.

[31] M. Kazemi, H. Zareipour, N. Amjady, W. D. Rosehart, and M. Ehsan, "Operation scheduling of battery storage systems in joint energy and ancillary services markets," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 4, pp. 1726–1735, 2017.

[32] Y. Shi, B. Xu, D. Wang, and B. Zhang, "Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 2882–2894, 2018.

[33] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[35] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting − a novel pooling deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271–5280, 2018.

[36] B. Chang, M. Chen, E. Haber, and E. H. Chi, "Antisymmetricrnn: A dynamical system view on recurrent neural networks," *arXiv preprint arXiv:1902.09689*, 2019.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[38] C. M. Bishop, "Mixture density networks," Citeseer, Tech. Rep., 1994.

[39] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York, 1988, vol. 84.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[42] ——, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019–1027.

[43] H. Hesse, M. Schimpe, D. Kucevic, and A. Jossen, "Lithium-ion battery storage for the grid a review of stationary battery storage system design tailored for applications in modern power grids," *Energies*, vol. 10, no. 12, p. 2107, 2017.

[44] A. Mashlakov, S. Honkapuro, V. Tikka, A. Kaarna, and L. Lensu, "Multi-Timescale Forecasting of Battery Energy Storage State-of-Charge under Frequency Containment Reserve for Normal Operation," in *2019 16th International Conference on the European Energy Market (EEM)*. IEEE, Sep. 2019.

[45] A. Mashlakov, "BESS SOC forecasting," https://github.com/aleksei-mashlakov/BESS-SOC-forecasting/tree/master/JSAC, 2019.

[46] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[47] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 983–999, 2006.

[48] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[49] S. Zheng, "Boosting based conditional quantile estimation for regression and binary classification," in *Mexican International Conference on Artificial Intelligence*. Springer, 2010, pp. 67–79.

[50] J. W. Taylor, "A quantile regression neural network approach to estimating the conditional density of multiperiod returns," *Journal of Forecasting*, vol. 19, no. 4, pp. 299–311, 2000.

[51] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," 2016.

[52] T. Hong, J. Xie, and J. Black, "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting," *International Journal of Forecasting*, 2019.

[53] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

[54] A. Mashlakov, V. Tikka, L. Lensu, A. Romanenko, and S. Honkapuro, "Hyper-parameter optimization of multi-attention recurrent neural network for battery state-of-charge forecasting," in *EPIA Conference on Artificial Intelligence*. Springer, 2019, pp. 482–494.

[55] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[56] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[57] M. Ratsimbazafy, "Mckinsey smartcities traffic prediction," https://github.com/mratsim/McKinsey-SmartCities-Traffic-Prediction, 2018.

[58] H. W. Leung and J. Bovy, "Deep learning of multi-element abundances from high-resolution spectroscopic data," *Monthly Notices of the Royal Astronomical Society*, vol. 483, no. 3, pp. 3255–3277, 2018.

[59] C. Martin, "Keras mixture density network layer," https://github.com/cpmpercussion/keras-mdn-layer, 2019.

[60] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

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

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[63] M. Ghenis, "Quantile regression from OLS to TensorFlow." 2019–, [Online; accessed 14-09-2019]. [Online]. Available: https://colab.research.google.com/drive/1nXOlrmVHqCHiixqiMF6H8LSciz583_W2#scrollTo=PQdJbWCS9N3G

[64] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in science conference*. Citeseer, 2013, pp. 13–20.

[65] M. Pumperla, "Hyperas: A very simple convenience wrapper around hyperopt for fast prototyping with keras models." 2017–, [Online; accessed 30-04-2019]. [Online]. Available: http://maxpumperla.com/hyperas/

[66] A. Khosravi, S. Nahavandi, D. Srinivasan, and R. Khosravi, "Constructing optimal prediction intervals by using neural networks and bootstrap method," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 8, pp. 1810–1815, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2019.2952195, IEEE Journal on Selected Areas in Communications

Aleksei Mashlakov received the double M.Sc. (Tech.) degree in electrical engineering from National Research University "Moscow Power Engineering Institute", Moscow, Russia and Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland in 2017.

He is currently pursuing the D.Sc. (Tech.) degree in energy market and solar economy at LUT University. His research focuses on proactive analytics of aggregated flexibility of distributed energy resources for provision of network management and system balancing services.

Lasse Lensu received the D.Sc. (Tech.) degree in computer science and engineering from Department of Information Technology of Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland in 2002.

He is currently a professor of machine vision and data analysis at LUT University. His research interests include machine/computer vision, pattern recognition with machine learning and data analysis.

Prof. Lensu is the head of the Department of Computational and Process Engineering of LUT University, and he has contributed to technology transfer to spin-off companies from the university.

Arto Kaarna received the M.Sc. (Tech.) degree in mechanical engineering, the Lic.Sc. (Tech.) degree in information technology, and the D.Sc. (Tech.) degree in information technology from Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland in 1980, 1990, and 2000, respectively.

He is currently an associate professor (tutkijaopettaja) at LUT School of Engineering Science. His main research interests are in digital image processing and in colour science.

Assoc. Prof. Kaarna has published 100+ articles in scientific conferences and journals. He is a member of Pattern Recognition Society of Finland (member of IAPR).

Ville Tikka received the M.Sc (Tech.) degree in electrical engineering from Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland in 2010.

Since then he has been a Doctoral Student and project manager in several projects at LUT University. His main area of interest is modeling of the grid effects of electric mobility and active resources in smart grids.

Samuli Honkapuro received the M.Sc. (Tech.) and the D.Sc. (Tech.) degrees in electrical engineering from Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland in 2002 and 2008, respectively.

He is currently an associate professor (tenure track) of energy markets at LUT University. His present research interests are related to business and market models for integration of distributed energy resources in energy markets.

Assoc. Prof. Honkapuro has been active in academic research related to electricity distribution business and electricity markets for over 15 years.