LUT University

Assessing the performance of deep learning models for multivariate probabilistic energy forecasting

Mashlakov Aleksei, Kuronen Toni, Lensu Lasse, Kaarna Arto, Honkapuro Samuli

This is a Publisher's version version of a publication

published by Elsevier

in Applied Energy

DOI: 10.1016/j.apenergy.2020.116405

Copyright of the original publication: © 2021 The Authors.

Please cite the publication as follows:

Mashlakov, A., Kuronen, T., Lensu, L., Kaarna, A., Honkapuro, S. (2021). Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. Applied Energy, vol. 285. DOI: 10.1016/j.apenergy.2020.116405

This is a parallel published version of an original publication. This version can differ from the original published article.

Applied Energy 285 (2021) 116405

Contents lists available at ScienceDirect

Applied Energy

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

Assessing the performance of deep learning models for multivariate probabilistic energy forecasting

Aleksei Mashlakov [a,*], Toni Kuronen [b], Lasse Lensu [b], Arto Kaarna [b], Samuli Honkapuro [a]

[a] School of Energy Systems, LUT University, Finland
[b] School of Engineering Science, LUT University, Finland

ARTICLE INFO

Keywords:
Deep learning
Energy forecasting
Multivariate modeling
Performance evaluation
Time series
Uncertainty estimation

ABSTRACT

Deep learning models have the potential to advance the short-term decision-making of electricity market participants and system operators by capturing the complex dependences and uncertainties of power system operation. Currently, however, the adoption of global deep learning models for multivariate energy forecasting in power systems is far behind the developments in the deep learning research field. In this context, the objectives of this study are to review recent developments in the field of probabilistic, multivariate, and multihorizon time series forecasting and empirically evaluate the performance of novel global deep learning models for forecasting wind and solar generation, electricity load, and wholesale electricity price for intraday and day-ahead time horizons. Two forecast types, deterministic and probabilistic forecasts, are studied. The evaluation data consist of real-world datasets with hourly resolution at the levels of an individual customer and regional and national electricity market bidding zones. The model evaluation criteria include achievable levels of forecasting accuracy and uncertainty risks, hyperparameter sensitivity, the effect of exogenous variables and fieldwise dataset split, and run-time efficiency factors, such as memory utilization, simulation time, electricity consumption, and convergence rate. We conclude that the performance of the global models is more beneficial for intraday forecasts of heterogeneous datasets with nonuniform patterns of time series, but can be affected by the hyperparameter sensitivity and hardware limitations with the growth of dataset dimensionality. The results can serve as a reference point for the quantitative evaluation of deep learning models for probabilistic multivariate energy forecasting in power systems.

1. Introduction

The short-term forecasting of energy time series, such as wind and solar energy, electricity load and price, is at the core of electric power system trading and operation. It provides the electricity market participants and system operators with information on the next hours and days to enable cost-efficient market bidding and operating reserve procurement and detecting network congestion. However, the operational predictability of modern power systems is being challenged by intermittency, uncertainty, and stochasticity as the installed capacity of renewable energy sources (RES) increases and new distributed energy resources are introduced into the existing power networks. To adapt to these decarbonization and decentralization trends and support the risk-aware decision-making of power system actors, the present energy forecasting approaches should be improved to estimate the prediction uncertainties and leverage large amounts of data with complex multivariate dependences [1].

Presently, deterministic forecasts are predominantly used in the industry as an input to various power system optimization methods. However, there is a strong trend in the academia and interest in the industry toward the transition from the deterministic forecasts to probabilistic methods with uncertainty quantification [2]. A probabilistic forecast provides a possible range of forecasting errors with the respective probability in contrast to pointwise deterministic forecasts. Moreover, the adoption of a forecasting tool/product enables reducing the operating costs if appropriately applied for solving risk-constrained decision-making problems [3]. The risk management is especially important in the conditions of volatile spot and reserve market prices caused by the large uncertainties of varying electricity load and weather-dependent renewable generation from wind and solar radiation. Therefore, the quantification of uncertainty as a vital part of risk management is critical for truly optimal decision-making in power systems.

Furthermore, a substantial body of research in the probabilistic energy forecasting literature focuses on local (univariate) forecasting

* Corresponding author at: Yliopistonkatu 34, 53850 Lappeenranta, Finland
E-mail address: aleksei.mashlakov@lut.fi (A. Mashlakov).

https://doi.org/10.1016/j.apenergy.2020.116405
Received 12 September 2020; Received in revised form 18 December 2020; Accepted 26 December 2020
Available online 16 January 2021
0306-2619/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

problems [4] where marginal predictive densities are estimated per individual time series assuming (conditional) independence of time series in high-dimensional settings. However, these local approaches exclude the important effects of complex temporal, spatial, and cross-lagged correlations in power systems [5]. The examples of such correlations are the dependence between successive lead times of electricity market price, the parameters of renewable generation (e.g., solar radiation, wind speed) between power plant locations, and the lagged effect of weather parameters on a load profile. Disregarding these dependences with marginal description in power systems-related operational problems with multiple power plants or optimization time periods leads to suboptimal decisions, and hence, is often insufficient forecasting approach [6]. In contrast, the ability to extract and leverage the time-invariant patterns by simultaneously considering several variables can potentially provide more accurate predictions and lower costs [7]. As a result, these factors have provoked an interest in probabilistic forecasting problems with multivariate predictive distributions that can leverage spatio-temporal and cross-lagged correlations with a single global (multivariate) model [8].

The key challenges for accurate and efficient forecasting in the probabilistic multivariate forecasting problems include the following: (a) recognizing short-term and long-term dynamics and noise characteristics of individual time series; (b) discovering nonlinear covariate and latent relationships between the exogenous (i.e., field-independent series or outside influences) and endogenous (i.e., field-dependent) series; (c) sharply (i.e., by minimal variance) and reliably (i.e., by minimal bias) estimating the uncertainty of model predictions; (d) mitigating the effects of a varying time series scale; (e) making predictions in the conditions of data sparsity and "cold start", i.e., new variable or system changes; and (f) being scalable for a large amount of time series [9–11]. Traditionally, statistical multivariate forecasting techniques such as vector autoregression (VAR) and vector autoregressive integrated moving average (VARIMA) [4], linear support vector regression (LSVR) [12], multivariate generalized autoregressive conditional heteroskedasticity (MGARCH) models [13], linear ridge (LRidge) regression [14] and Gaussian processes (GP) [15] have been used for such problems, but they have several limitations related to nonlinearity and scalability [16]. Recently, several novel global deep learning (DL) architectures for multivariate time series forecasting with the capabilities to tackle these challenges have been proposed: autoregressive recurrent networks (DeepAR) [9], deep factor models with random effects (DFM-RF) [10], long- and short-term time series network (LSTNet) [16], temporal pattern attention (TPA) [17], deep temporal convolutional network (DeepTCN) [18], and dual self-attention network (DSANet) [19], to name a few. These models have demonstrated superior accuracy. It is worth noting that any DL method initially designed solely for point forecasts can be extended to provide an estimate of the related uncertainty by applying variational approximation by the Monte Carlo (MC) dropout [20]. Furthermore, progress has been made in explainability and interpretability of DL models also in the context of time series [21] (e.g., through an attention mechanism's weights [22] or saliency maps [23] over time dimensions and features), which gradually changes the general opinion of them as representing fully "black box" models and brings them closer to industry acceptance. These factors along with the wealth of data being collected in power systems all the time and the rapid increase in computing capabilities potentially make them a promising method for probabilistic multivariate energy forecasting.

1.1. About the related work

In the power system domain, there is a plethora of local (univariate) DL-based forecasting models that consider time series independently and, hence, are missing the important interdependence between the series. For instance, the known forecasting applications include electricity price [24], wind [25] and solar [26] power production, total [27] and net [28] load, and battery frequency response [29]. However, there is increasing interest in applying DL models to multivariate forecasting applications to capture both spatio-temporal information and cross-variable dependences in power markets enriched with RES [7], such as solar [30] and especially wind [31]. For example, an improved deep mixture density network (IDMDN) for a short-term wind power probabilistic forecasting of multiple wind farms and the entire region was introduced in [32] to model nonlinear and spatio-temporally coupled uncertainties in wind power prediction. A deep architecture, predictive spatio-temporal network (PSTN), was proposed in [33] for wind speed prediction with the use of spatial features and temporal dependences. A variational Bayesian DL model for probabilistic spatio-temporal forecasting was presented in [34] that predicts the wind speed in a region by exploiting multisite historical information. A novel multifactor spatio-temporal correlation model for wind speed forecasting was proposed in [35]. The combination of spatial and temporal correlations extracted by the DL networks was also used for very short-term solar irradiance forecasting in [36].

Despite the latest developments, the literature in DL-based probabilistic multivariate time series forecasting with application to energy forecasting is still in its infancy and is lagging behind the progress made in the field of computer science that we review in Section 2. Moreover, to the best of the authors' knowledge, a comprehensive and sound empirical evaluation of probabilistic multivariate DL architectures that would assess their multihorizon forecast accuracy and uncertainty for the needs of power systems is not yet available in the literature. However, as stated in a tutorial study in [37] about probabilistic load forecasting, reproducible empirical studies based on public data with sufficient details and unified forecast evaluation are required for research progress in the field. The same requirement is valid for DL-based probabilistic multivariate studies because the literature is heavily occupied with empirical evaluations of DL-based univariate deterministic cases [38] or statistical and physical models [39]. Furthermore, many computing and efficiency details about DL-based forecasting models are often ignored, yet they present valuable practical information if applied for short-term operations.

1.2. Research gaps and scientific contribution

Given that the power system developments of global DL forecasting models are far behind the progress in the DL research, our main aim is to attain the parity between the advances in DL-based multivariate forecasting and power system applicability. This justifies a comprehensive quantitative evaluation of novel data-driven approaches developed in the DL research, which motivates this paper with the focus to address the following research gaps:

(i) Although various global DL models have emerged in recent years, no systematic evaluation of the applicability or suitability of these models to address the energy forecasting problem in power systems has been carried out to date.
(ii) The sensitivity of the global DL models to hyperparameters and exogenous time series has received limited attention in the literature.
(iii) The practical run-time requirements of the global DL models in terms of computing power and energy efficiency continues to be not well covered.

With these research gaps in mind, this study provides the following contributions and novelty:

(i) Empirical evaluation of the advanced global DL models for accuracy and quantile risks on intraday and day-ahead time horizons for the electricity load and price, and wind and solar generation at the levels of individual customer, regional, and national power system.

(ii) Assessment of model sensitivity to common hyperparameters using sequential Bayesian optimization, calendar and time exogenous variables, and fieldwise dataset split.
(iii) Relative estimation of run-time efficiency of the advanced global DL models through total simulation time, the mean and standard deviation of graphics processing unit (GPU) memory, convergence rate, and estimated electricity consumption.

The analysis is performed using two real-world datasets including almost three years' worth of data with hourly resolution and dimensionality of hundreds of time series. An important assumption is that the forecasts are solely based on past data, time and calendar features ignoring meteorological observations and numerical weather prediction. The results of this work can serve as a reference point for DL-based probabilistic multivariate forecasting of electricity load and price and wind and solar production at various power system levels. The broader objective of this study is to comprehend the efficiency of data-driven techniques in power system problems involving data analytics. Application of global DL-based models can help electricity market participants and power system operators in improving their forecasting products and making better decisions on market bidding, setting the operating reserve requirements, and detecting technical problems of grid management. Hence, a DL-based probabilistic multivariate forecasting model that can appropriately accommodate both uncertainties and pattern dependences holds significant potential and is thus of a special interest for electric power systems.

The rest of this paper is organized as follows. Section 2 describes the background of the research in multivariate time series forecasting and reviews the progress in novel global DL architectures. The problem formulation, models examined, benchmarks, data used, evaluation metrics, the setting of hyperparameter optimization, and implementation details of the empirical study on multivariate probabilistic forecasting are addressed in Section 3. Section 4 reports the results of the models on point forecast accuracy and uncertainty estimation, hyperparameter sensitivity, the effect of exogenous variables and fieldwise dataset split, and run-time efficiency. Finally, the results are discussed in Section 5 and conclusions are derived in Section 6.

2. Background and related literature

Statistical methods have been the basis for multivariate forecasting from its origin. The examples consist of parametric autoregressive (AR) models, such as the variants of VAR and VARIMA models [4] and the MGARCH model [13], parametric regression models, such as LSVR [12] and LRidge regression [14], and nonparametric methods, such as GP [15]. The main drawbacks of these approaches are related to the inability to capture long-term and nonlinear relationships between time steps and between multivariate signals, and a high computational cost and model capacity that can increase significantly over the larger window size and the number of time series [16]. The important impact of the seasonalities and causal determinants of the related time series was shown in [40], where joint modeling of related series in a hierarchical Bayesian state-space model (SSM) enabled to achieve sizable accuracy gains. Moreover, a scalability issue was addressed in [41] with temporal matrix factorization models (TMFMs) that support data-driven temporal dependence learning and forecasting. However, these dependences are limited to linear relationships [42].

A large amount of recent research is focused on global DL as a solution to tackle the limitations of statistical methods. We provide a summary of these models in Table 1 and highlight the trends in global DL methods in what follows below. For example, one of the main research directions in global DL methods is the merger of several families of models to use the specific model strengths for the best performance; this approach was implemented, e.g., in LSTNet [16], TPA [17], DeepTCN [18], and DSANet [19]. Moreover, global DL methods are commonly used together with nonparametric models to model forecast uncertainty (e.g., DeepAR [9] and DeepTCN [18]), and with statistical AR methods to improve the accuracy by scale consideration; such combination was applied, e.g., in LSTNet [16], DSANet [19], and memory timeseries network (MTNet) [22]. The other trend is to deploy an attention mechanism that provides a model with the ability to focus on relevant subsets of its inputs to predict the target series without explicitly hard-coding these subsets; the attention mechanism was employed, e.g., in LSTNet [16], TPA [17], and DSANet [19]. This mechanism was designed to solve the vanishing gradient problem of recurrent neural networks (RNNs) [43] when forecasting long-range dependences. A similar function is often performed with skip connections [16] and residual blocks [18]. Furthermore, the adoption of convolutional neural networks (CNNs) [44] that are known for their success in capturing the spatial and temporal dependences in image recognition surpasses the more traditional solutions with RNNs in memorizing short- and long-term invariant patterns and as a basis for the attention mechanism; the CNNs were utilized, e.g., in TPA [17], DeepTCN [18], and DSANet [19]. However, both methods (i.e., RNN and CNN) are dominant in most of the architectures in one way or the other. In addition, the adaptive moment estimation (Adam) optimization algorithm [45] is the first choice for the global DL models.

The challenge of many AR methods in capturing recurrent short- and long-term patterns among multiple time series is addressed in the LSTNet [16]. This model uses the strengths of CNNs to discover the local dependence patterns among multidimensional input variables, RNNs to capture long patterns, and RNN-skip or attention layers to recognize the very long-term periodic patterns of time series. Moreover, in parallel to the nonlinear neural network transformation that is insensitive to the scale variations of inputs, it includes an AR linear model to consider the effects of scale variation in the time series.

A TPA model was developed in [17] based on RNN and CNN modules. The model has resolved several limitations of LSTNet, such as manual tuning of the skip length of the recurrent-skip layer, poor performance on data with a nonperiodic pattern, and averaging of series-specific temporal information in the attention layer. In contrast, the invariant temporal pattern information is automatically extracted from each time series with CNN filters that operate similar to the discrete Fourier transform. Thus, the model can focus on particular time intervals for different time series and extract dynamic interdependences of multivariate data.

A DSANet was proposed in [19]; it highlights the malperformance of the attention mechanism used in LSTNet and TPA when modeling data with dynamic period patterns or nonperiodic patterns. Instead, the nonlinear branch of dual architecture completely dispenses RNNs and applies global and local temporal convolutions to capture the complex mixtures of global and local temporal representations of univariate series. Moreover, a self-attention module is added above these convolutions to identify cross-series dependences. Similarly to LSTNet, the AR linear branch is included in the model to alleviate input scale variations.

A mixture of DL-family models was also proposed in the architecture of MTNet including a large memory network component, three independent encoders, and an AR component. The memory component is used to store the long-term historical data, while encoders equipped with CNN, RNN, and an attention mechanism are used to convert the input data and memory data into their feature representations. The advantages of the model are a high interpretability using post-hoc explanations with the attention mechanism and a capability to focus on a period of time instead of particular timestamps in the past as it is implemented in the LSTNet model.

DeepAR was designed as an AR-based RNN that relies on a global model of related time series [9]. This global model learns the statistical properties of the data by maximizing the log-likelihood of the network outputs conditioned on past observations and covariates that can be item- and time-dependent. DeepAR is claimed to be robust to the effects of the time series with widely varying scales and can use a wide range

Table 1
Summary of deep-learning-based multivariate time series models. *Family of models*: autoregressive (AR), convolutional neural network (CNN), recurrent neural network (RNN), Gaussian processes (GP), feed-forward neural network (FFNN), temporal matrix factorization model (TMFM), likelihood model (LM), memory network (MN), quantile model (QM), innovation state space model (ISSM); *Exogenous variables*: C (calendar), T (time), CT (categorical); *Forecast type*: CM (conditional mean), CD (conditional distribution); *Dataset dimension*: minute (m), hour (h), day (D), week (W), month (M); *Tuning*: grid search (GS), manual (M); (?) Not explicitly stated. *Benchmark models*: The abbreviations are explained in the text.

Model		Families of models	Forecast horizon	Exogenous variables	Optimizer	Time series scaling	Forecast type	Dataset dimension	Benchmark models	Tuning
[16]	LSTNet	AR, CNN, FFNN, RNN	3 to 24	–	Adam	Max per series	CM	10m, 1h, 1D	AR, LRidge, LSVR, TMFM, GP, VAR-FFNN, RNN	GS
[17]	TPA	AR, CNN, FFNN, RNN	3 to 24	–	Adam	Max per series/ data	CM	10m, 1h, 1D	AR, LRidge, LSVR, GP, LSTNet	GS
[19]	DSANet	AR, CNN, FFNN	3 to 24	–	Adam	MinMax	CM	1D	VAR, LRidge, LSVR, GP, RNN, LSTNet, TPA	GS
[22]	MTNet	AR, CNN, RNN, MN, FFNN	3 to 24	C, T	Adam	(?)	CM	10m, 1h, 1D	AR, LRidge, LSVR, GP, VAR-MLP, RNN-GRU, LSTNet	GS
[9]	DeepAR	FFNN, LM, RNN	24 to 52	CT, C, T	Adam	Mean per series	CD	1h, 1W, 1M	TMFM, RNN-LM, AR-LM, ISSM	GS+M
[46]	DSSM	FFNN, ISSM, LM, RNN	8 to 48	C, T	(?)	Mean per series	CD	1h, 1M, 4M	DeepAR, TMFM	(?)
[18]	DeepTCN	CNN, FFNN, LM/QM	12 to 24	CT, C, T	Adam	Standard/ MinMax	CD	1h, 1D, 1M	DeepAR, TMFM, DSSM	M
[10]	DFM-RF	FFNN, GP, ISSM, LM, RNN	24 to 72	CT, C, T	Adam	Automatic per series	CD	1h	DeepAR, Multi Quantile-RNN	–
[42]	DeepGLO	CNN, TMFM	9 to 24	C, T	Adam	Standard/ unscaled	CM	5m, 1h, 1D	RNN, DeepAR, TMFM	M
[47]	ForecastNet	CNN, GP	12 to 24	–	Adam	Standard	CD, CM	1h, 1M	DeepAR, TCN	M

of likelihood functions to better capture the statistical properties of the data. Moreover, the network enables the discovery of time-dependent uncertainty growth and complex patterns, such as seasonal behavior and cross-series dependences.

The Deep state space model (DSSM) proposed in [46] combines a deep RNN and SSM to explicitly incorporate structural assumptions and learn complex patterns from raw time series data. This model computes the joint distribution over the prediction range for each time series analytically based on a globally shared mapping derived from the covariate vectors associated with each time series. The RNN is needed to parametrize the mapping from covariates to the SSM parameters. It is stated that this method allows model interpretability, can exploit assumptions about temporal smoothness, and can be seamlessly scalable to high-dimensional datasets.

A DeepTCN that employs a dilated causal CNN to capture the temporal dependences of multiple related time series was presented in [18]. The novelty of this model is a residual block designed to learn the complex patterns within and across series from past observations and exogenous covariates. The framework includes parametric and nonparametric variants that learn uncertainty estimation by predicting the parameters of hypothetical data distribution or generating quantile forecasts.

DFM-RF represent a local–global method that relies on the combination of individual and generic time series properties for multivariate forecasting [10]. This method adopts deep dynamic factors to extract global nonlinear patterns and probabilistic graphical models to capture local random effects. This model with one factor and AR inputs, without random effects and automatically estimated scales of time series, reduces to DeepAR.

Deep global local forecaster (DeepGLO) is another example of local–global methods [42]. To capture global nonlinear dependences, this method uses temporal convolution network (TCN) for the regularization of the Matrix Factorization model (TCN-MF) that represents each of the original time series as a linear combination of a smaller number of basis time series. The output of TCN-MF is fed as input covariates to another TCN along with the past values of the local time series and associated covariates to make the final prediction. Moreover, a simple initialization scheme for TCN that dispenses a priori normalization was introduced to handle scale variation of high-dimensional time series data.

A time-variant deep feed-forward neural network architecture for multi-step-ahead time-series forecasting (ForecastNet) [47] is a model that addresses the time-invariance problem of RNN and CNN models. The model produced good results against state-of-the-art models, such as, DeepAR and DeepTCN. However, the evaluation was conducted using univariate time series [47].

3. Methodology

3.1. Problem formulation

First, the probabilistic forecasting problem for multivariate time series is formulated. Given a trained model $f^{\hat{W}}(\cdot)$ with fitted parameters $\hat{W}$ and a set of fully observed time series $\boldsymbol{Y} = \{\boldsymbol{y}^{(i)}\}_{i=1}^{N}$ with N univariate series $\boldsymbol{y}^{(i)} \in \mathbb{R}^{D}$ where D is the series dimension, the aim is to predict the conditional distribution of the future time series $\hat{\boldsymbol{y}}^{(i)}_{(t+1):(t+h)}$ for $i = 1, \dots N$:

$$\mathbb{P}\left(\hat{\boldsymbol{y}}^{(i)}_{(t+1):(t+h)} | \boldsymbol{y}^{(i)}_{1:t}, \boldsymbol{X}^{(i)}_{1:t}, \boldsymbol{X}^{(i)}_{(t+1):(t+h)}, \hat{W}\right), \tag{1}$$

where $h \in \mathbb{N}^{+}$ is a forecast horizon, t is a current time stamp, $\boldsymbol{y}^{(i)}_{1:t}$ are historical values of the ith series, and $\boldsymbol{X}^{(i)}_{1:t}$ and $\boldsymbol{X}^{(i)}_{(t+1):(t+h)}$ are the optional associated covariate vectors related to the past or future. The forecast horizons are selected based on the needs of the power system applications for day-ahead dispatch in response to the results of the wholesale market (36 h ahead) and the consequent correction of the system dispatch (3, 6, 12, and 24 h ahead) during the delivery period in the intraday market. Note, however, that the error estimation for the day-ahead forecasts is normally done only for 24 h of the next day, but in our case, the forecast errors at all 36 h are estimated.

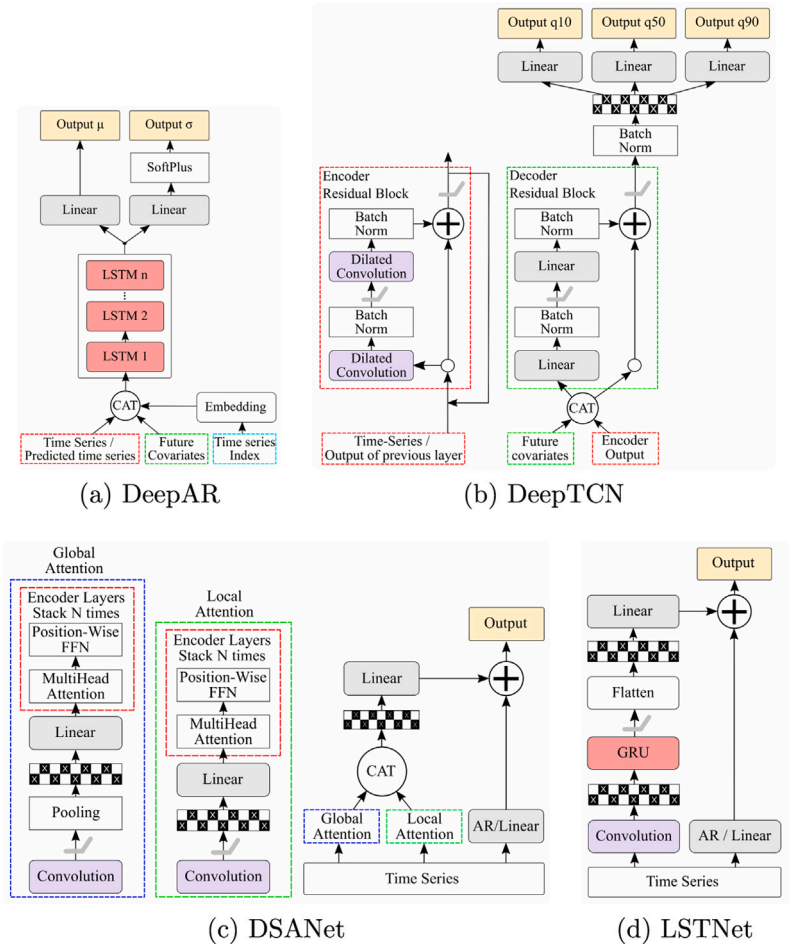

Fig. 1. Visualization of the structures of the selected methods: DeepAR (a), DeepTCN (b), DSANet (c), and LSTNet (d). ReLU activation functions are indicated by a gray curve symbol and the dropout layers by the checkerboard pattern. The abbreviations are explained in the text.

3.2. Methods

The DeepAR, DeepTCN, LSTNet, and DSANet models were selected for the empirical study to examine the recent advances in global DL models that have achieved notable results in the forecasting of multivariate time series. These models were selected because they include most of the trends in multivariate forecasting, and at the same time, are diverse for potential combinations of approaches. In particular, the LSTNet and DSANet models belong to the class of many-to-one models that predict one value that is at the horizon distance from an input sequence, whereas DeepAR and DeepTCN are many-to-many models predicting a sequence of a horizon length ahead given the input sequence. Moreover, the use of exogenous variables is also different: the DeepAR and DeepTCN models have categorical, calendar, and time variables, whereas LSTNet and DSANet do not use them by default.

Fig. 1 presents high-level schemes of the layers used in the models. The same colors are used to indicate similar layers and their components in the different models. DeepAR consists of multiple layers of long short-term memory (LSTM) components where the outputs are fed into two linear layers, one for determining the mean and the other for determining the standard deviation of the time series. In the case of standard deviation, the linear layer is fed into a SoftPlus layer to ensure positive values. The means and standard deviations are then the input to a Gaussian likelihood model that is used to generate samples. DeepAR with three LSTM layers is shown in Fig. 1a, where the output layer can be seen as the input for the likelihood model. The inputs for the model are time series values until $t-1$, the covariates at time t, and the time series index that is fed into the embedding layer. The output of the embedding layer is then concatenated (CAT) with the time series values and covariates and fed into LSTM layers.

Fig. 1b consists of two residual blocks, encoder and decoder, and batch normalization, dropout, and linear layers for determining quantile outputs q10, q50, and q90 (i.e., for 0.1, 0.5, and 0.9 quantiles) of the DeepTCN model. The encoder residual block consists of dilated convolution layers whose output is fed into the batch normalization layers, where from the first batch normalization layer the output is fed through the ReLU activation function. The batch normalization layers are aimed at providing a stable distribution of activation values during the training [48]. This enables faster convergence and shortens the training process of the model. The output of the residual block is fed into the next residual block or to the decoder residual block in the case of the last encoder residual block. The decoder residual block takes two inputs, the future covariates and the output from the encoder residual block. It consists of linear layers and batch normalization layers, where the output from the first batch normalization layer is fed through the ReLU activation function. The output of the batch normalization layer is summed with the concatenated future covariates and the encoder output and fed through ReLU. The output from the decoder residual block after the ReLU is fed into batch normalization, dropout, and finally linear layers, which then provide the quantile outputs of the model.

The LSTNet model is presented in Fig. 1d. Time series until $t-1$ are fed into the convolutional and linear layers. The convolution layer is followed by the dropout and gated recurrent unit (GRU) layers. The output from GRU goes through the ReLU activation followed by a flattening operation, after which dropout is done. The dropped-out

Table 2
The details of the datasets.

Dataset	Electricity	Open power system
Number of series	321	183
Length	26,304	25,560
Domain	$\mathbb{R}^+$	$\mathbb{R}$
Granularity	Hourly	Hourly

results are then fed into the linear layer. The final output of the model is calculated as a sum of the output from the linear layer and the output of the AR/linear layer. DSANet can be seen as an evolution of the LSTNet model, and the similarities between LSTNet and DSANet can be seen in Figs. 1c and 1d. GRU with ReLU and the flatten layers of the LSTNet layers are replaced with linear layers and encoder layers that consist of self-attention and positionwise feed-forward neural network (FFNN) layers in the DSANet model. Moreover, the DSANet model uses global and local temporal convolutions, which are aimed to capture both long- and short-term temporal patterns. Furthermore, global and local convolutions are fed into the self-attention layers, indicated as encoder layers in Fig. 1c, which aim to identify the dependences between different series. Both LSTNet and DSANet take only time series values as inputs without including any future covariates by default. Unlike the output of the probabilistic many-to-many models DeepAR and DeepTCN, the outputs of these models are single point values at a horizon away from the time t.

The LSTNet and DSANet models are not designed to produce probabilistic forecasts. Therefore, a probabilistic interpretation of dropout is applied to obtain an approximate variational distribution representing uncertainty estimates [20]. For the DeepTCN, a nonparametric model that predicts the quantiles is used, and it is referred to as TCN-quantile in [18].

3.2.1. Benchmarks

We use two similar-day techniques as the trivial benchmark methods and denote these benchmarks by Naïve-1 and Naïve-2:

$$\hat{y}^{(i),Naive-1}_{(t+h)} = y^{(i)}_{(t+h-d)}, \tag{2}$$

$$\hat{y}^{(i),Naive-2}_{(t+h)} = \begin{cases} y^{(i)}_{(t+h-d)}, & \text{if Tue } (h \le 24), \text{ Wed, Thu, and Fri,} \\ y^{(i)}_{(t+h-d)}, & \text{if Sat, Sun, Mon or Tue } (h > 24), \end{cases} \tag{3}$$

where offset $d = 24$ if $h \le 24$ and $d = 48$ if $h > 24$. We assume that the first model is more relevant for solar and wind time series, whereas the second is primarily for load and price time series. The forecast uncertainty of the persistence models is obtained using a historical *(post-hoc)* simulation that consists of computing sample quantiles of the empirical distribution of model residuals [49]. Here, we use a weekly $(t-168)$ rolling window of naïve model residuals prior to the prediction time t, i.e, $\mathbf{y}^{(i)}_{(t-168):(t)} - \hat{\mathbf{y}}^{(i)}_{(t-168):(t)}$.

A deep FFNN modeled in GluonTS [51] serves as a benchmark method for univariate DL forecasting. The model consists of two ReLU layers with 40 hidden neurons in each and is trained with a batch size of 32 to fit a Gaussian distribution. The probabilistic forecasts are obtained by sampling the quantiles of the output distribution.

3.2.2. Experiment details

All DL models are trained for the horizons with the early stopping criterion being equal to 25 epochs and the maximum number of epochs being set to 500. The selected epoch values correspond to the smallest and largest epoch numbers used in the global DL models under examination.

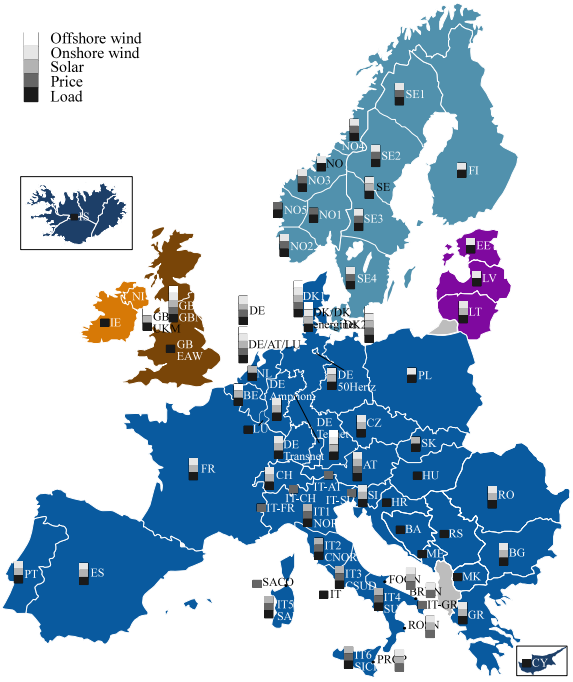

Fig. 2. Geolocations of the variables in the preprocessed `open power system` dataset. The geolocations are specified with ISO 3166 area code or name of control area or bidding zone [50].

3.3. Data

In this study, two real-world datasets related to different granularities of power systems and consisting of homogeneous and heterogeneous time series were taken for the comparison. The statistics of these datasets are presented in Table 2. The datasets and the tested DL models are publicly available.[1]

The `electricity` dataset has already been used in most of the models under examination, except DSANet. However, it is impossible to compare the performance of the models on this dataset because of the different preprocessing of the data and error metric. Here, the modification of this dataset[2] used in [16] is employed that has a reduced dimensionality as a result of the removal of the time range and the customers with a significant share of zero values. This version has hourly consumption values in kWh for 321 customers for the time period from 2012 to 2014. As in the initial models, the same split principles were followed for the dataset, such as 60% for training, 20% for validation, and 20% (5256 samples per series) for out-of-sample (OOS) testing. The testing samples have a sufficient size covering several seasonal, monthly, and diurnal patterns for the time period from summer to wintertime.

The `open power system` dataset represents the data originating from the European market bidding zones. This dataset contains a diverse mix of time series, namely electricity consumption, market prices, and wind and solar power generation with hourly resolution from January 2015 to November 2017 [50]. The consumption and generation variables are given in MW, and the prices in euro or pound sterling. The split percentages for this dataset are 70%, 15%, and 15% (3816 samples per series). The initial dataset was preprocessed by removing the capacity, forecast, and profile data, as well as the series whose percentage of missing values exceeds 5% for the defined time period. As a result, the data consist of 183 variables, where 59 are related to load, 31 to price, 57 to onshore and offshore wind, and 36 to solar. The geolocations of the variables in the dataset are illustrated in Fig. 2. The illustration

[1] https://github.com/aleksei-mashlakov/multivariate-deep-learning.
[2] https://github.com/laiguokun/multivariate-time-series-data.

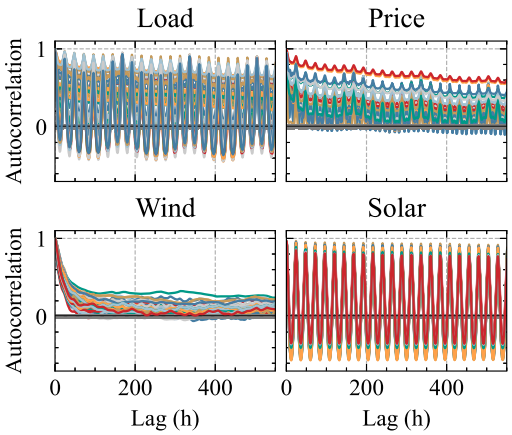

Fig. 3. Autocorrelations of the time series in the `open power system` dataset. A presence (absence) of repetitive pattern indicates seasonality (randomness) in the series.

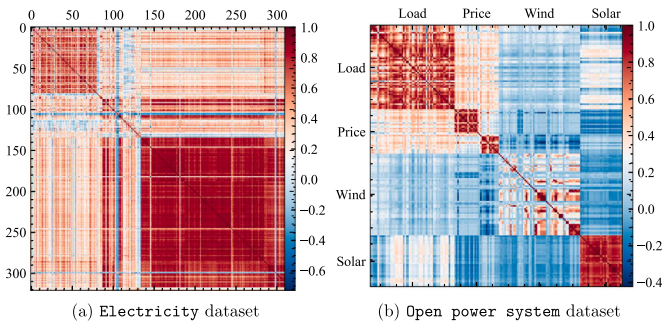

Fig. 4. Correlation matrix of time series in the `electricity` (a) and `open power system` (b) datasets. Red to blue colors represent the highest to the lowest correlations.

suggests that the dataset provides good grounds for the spatial load dependences represented in almost all the bidding zones. Similarly, the price variables are densely located in the Italian bidding zones and in the Nordic countries, and solar variables are common in Southern Europe and offshore wind in the north.

In order to examine the time-varying properties of variables in the `open power system` dataset, autocorrelation graphs are shown for the related time series fields in Fig. 3. The autocorrelation graph of the `electricity` dataset is available in [16] and shows clear short-term daily (24 h) and long-term weekly (168 h) seasonality. In the `open power system` dataset, a clear serial correlation with daily and weekly patterns can be observed in the Load and Price variables. However, only a daily pattern with a steadily decreasing trend can be seen in Solar and no repetitive patterns in Wind, which suggests about the high randomness of the wind data with profound short-term dependences. These observations can indicate possible challenges in wind forecasting and an expectation of better results for the other variables and are revised again in Section 4.

To support the idea of the existence of a spatial dependence of time series at different zones in power systems, the empirical covariance matrices of the datasets are visualized in Fig. 4. According to Fig. 4a, there are mostly positive correlations between the consumption patterns of households in the `electricity` dataset. In particular, at least two large groups of the first 80 and the last 190 households have significant correlations. For the `open power system` dataset shown in Fig. 4b with random variables from different subfields, the correlations vary within and between the subfields. It is to be noted that there are correlations between the Load and Price variables, whereas such correlations are absent between Price and renewable generation. This fact demonstrates the higher influence of Load on the electricity market clearing price and yet, low levels of renewables in the generation mix of the European bidding zones. Moreover, the red squares within Wind and Price show the probable spatial closeness of several bidding zones.

The selected datasets cover a wide range of use cases for the potential application of probabilistic multivariate forecasting methods by system operators and electricity market participants. For example, the `electricity` dataset serves the needs of system operators and retailers. The system operators can produce probabilistic load forecasting at multiple distribution levels and grid nodes and detect technical problems, such as congestion in the electrical grid in advance. For retailers, the multivariate time series forecasting can provide more accurate descriptions of the behavioral patterns on thousands of customers and make better tariff proposals. In contrast, the `open power system` dataset is a good reference for operations at lower granularity levels in power systems. For the business of aggregators, these methods are useful to precisely take into account the production of the renewables-based virtual power plants with multiple sources in the day-ahead and intraday energy markets. Furthermore, such a forecasting procedure can be used to balance renewables by leveraging the use of cross-border interconnections with suitable flexible power plants.

3.4. Evaluation metric

The model forecasts are assessed in accordance with the recommended practices of renewable energy forecasting [52], i.e., using point and probabilistic forecast numerical metrics and formal statistical tests. The point forecast metrics evaluate the predictive accuracy of forecasting the conditional mean of the series per horizon, whereas the probabilistic forecast metrics estimate the model performance in compliance with the paradigm of 'maximizing sharpness subject to reliability' [11]. Finally, the formal statistical tests are applied to verify the statistical consistency between the performance difference in the results of point and probabilistic forecasting.

3.4.1. Point forecasting

The evaluation metric for point forecasting is normalized by the sum of actual time series values to enable a fair comparison of multiple series. It consists of ND and NRMSE. For the given set of time series $\boldsymbol{Y}$ and the corresponding predictions $\hat{\boldsymbol{Y}}$, the metric is defined as follows:

$$\begin{aligned} \text{ND} &= \frac{\sum_{i,t} |y_t^{(i)} - \hat{y}_t^{(i)}|}{\sum_{i,t} |y_t^{(i)}|}, \\ \text{NRMSE} &= \frac{\sqrt{\frac{1}{N}\sum_{i,t}(y_t^{(i)} - \hat{y}_t^{(i)})^2}}{\frac{1}{N}\sum_{i,t}|y_t^{(i)}|}, \end{aligned} \tag{4}$$

where $y_t^{(i)}$ is the true value of the series i at the time step t, $\hat{y}_t^{(i)}$ is the corresponding prediction value, and N is the number of all points in the testing periods.

To evaluate a statistical difference in the accuracy of model point forecasts from each other, we conduct a Diebold–Mariano test [53]. Given that $L_t(y_t^{(i)}, \hat{y}_t^{(i)})$ is an arbitrary forecast loss function of the observation and prediction at time t, the idea of the test is to compute the difference between the forecast scores of the pair of models A and B:

$$\Delta_t^{A,B} = L_t^A(y_t^{(i)}, \hat{y}_t^{(i)}) - L_t^B(y_t^{(i)}, \hat{y}_t^{(i)}), \tag{5}$$

and to perform an asymptotic z-test for the null hypothesis that the expected forecast error is equal and the mean of differential loss series is zero $\mathbb{E}(\Delta_t^{A,B}) = 0\ \forall t$, i.e., that there is no statistically significant difference in the accuracy of two competing forecasts. Then, under the assumptions of covariance stationarity of loss differential series, the statistic of the test is deduced from the asymptotically standard normal distribution as follows:

$$\text{DM}^{A,B} = \sqrt{M}\frac{\hat{\mu}_{\Delta^{A,B}}}{\hat{\sigma}_{\Delta^{A,B}}}, \tag{6}$$

where $\hat{\mu}_{\Delta^{A,B}}$ and $\hat{\sigma}_{\Delta^{A,B}}$ are the sample mean and the standard deviation of $\Delta^{A,B}$, and M is the length of OOS period.

Two one-sided DM tests are conducted at the 5% significance level p for the forecast series of each horizon: (i) a standard test with the null hypothesis $H_0 : \mathbb{E}(\Delta_t^{A,B}) < 0$, i.e., the forecasts of the model B are more accurate than those by the model A, and (ii) the reverse null $H_0^R : \mathbb{E}(\Delta_t^{A,B}) \geq 0$, i.e., the forecasts of the model B are less accurate than those by the model A. If the p-value of the test is lower than the significance threshold, we reject the null hypothesis in favor of the alternative. As loss functions $L_t(\cdot)$, we use the above-mentioned ND and NRMSE metrics leading to the *multivariate* variant of standard DM tests and assume that the loss differential series is covariance stationary.

3.4.2. Probabilistic forecasting

The performance of the probabilistic forecasting is quantitatively evaluated based on the *reliability* and *sharpness* criteria. The *reliability* (also called *calibration* or *unbiasedness*) validates the statistical consistency between the distributional forecasts and the observations, whereas the *sharpness* measures the concentration of the predictive distributions.

The *reliability* is typically assessed by the coverage rate of the PI using a 'hit and violation' indicator $I_t^{(i)}$:

$$I_t^{(i)} = \begin{cases} 1, & y_t^{(i)} \in [\hat{L}_t^{(i)}, \hat{U}_t^{(i)}] \rightarrow \text{'hit'}, \\ 0, & y_t^{(i)} \notin [\hat{L}_t^{(i)}, \hat{U}_t^{(i)}] \rightarrow \text{'miss' (violation)}, \end{cases} \tag{7}$$

where $\hat{L}_t^{(i)}$ and $\hat{U}_t^{(i)}$ are the lower and upper bounds of PI for the series i at time t. These bounds of PI with a nominal coverage rate c and confidence level α can be described by the lower (i.e, $\underline{\tau} = \alpha/2$) and upper (i.e, $\overline{\tau} = 1-\alpha/2$) quantile. That means that for the PI, the nominal coverage should be equal to $\mathbb{P}(y_t^{(i)} \in [\hat{y}_t^{\underline{\tau},(i)}, \hat{y}_t^{\overline{\tau},(i)}]) = 1 - \alpha = c$, i.e., the realized value is expected to be within the predicted range $100 \cdot c$ % of the time. For one-sided PI specified by an individual quantile τ, i.e., $[-\infty, \hat{y}_t^{\tau,(i)}]$, the realized value is anticipated to be lower than the predicted quantile τ in $100 \cdot \tau$ % of the cases.

Given the sequence of $\{I_t^{(i)}\}_{t=1}^{T}$, the prediction interval coverage probability (PICP) is found as follows:

$$\text{PICP} = \frac{1}{N} \sum_{i,t} I_t^{(i)}, \tag{8}$$

whereas the mismatch of the empirical coverage with the prediction interval nominal coverage (PINC) is evaluated using the average coverage error (ACE):

$$\text{ACE} = \text{PICP} - \text{PINC}. \tag{9}$$

Generally, the closer the empirical coverage is to the nominal rate, the better.

The coverage rate is formally validated by conditional coverage (CC) Christoffersen tests [54] that assess an unconditional coverage (UC) and independence hypothesis by LR evaluation procedures.

The unconditional coverage test, initially developed by Kupiec [55], tests the null hypothesis that the nominal coverage rate is equal to the empirical $H_0 : \mathbb{E}(c) = \mathbb{E}(\pi)$ based on the total number of violations and ignoring the order of the 'hits' and 'misses':

$$\text{LR}_{\text{UC}} = -2 \log\left\{ \frac{(1-c)^{n_0}(c)^{n_1}}{(1-\pi)^{n_0}(\pi)^{n_1}} \right\} \overset{\text{asymp.}}{\sim} \chi^2(1), \tag{10}$$

where $\pi = n_1/(n_0 + n_1)$ is the percentage of hits, n_0 and n_1 being the number of ones and zeros in the indicator $I_t^{(i)}$ series. The null hypothesis is rejected if the actual fraction of PICP violations is statistically different than α.

The independence test validates the hypothesis of an absence of the first-order dependence in the violation sequence, i.e., the violations must be distributed independently, based on the estimation of the first-order Markov chain model, and it is defined as follows:

$$\text{LR}_{\text{IND}} = -2 \log\left\{ \frac{(1-\pi_2)^{n_{00}+n_{10}}(\pi_2)^{n_{01}+n_{11}}}{(1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}} \right\} \overset{\text{asymp.}}{\sim} \chi^2(1), \tag{11}$$

Table 3
Details of the hyperparameter search space.

Parameter	Search space	Stochastic expression
Hidden units	2^5, 2^6, 2^7, 2^8 $5^{2\cdot}$ [2^1, 2^2, 2^3]	Quantized uniform
Batch size	2^6, 2^7, 2^8, 2^9, 2^{10}	Quantized uniform
Dropout rate	0.1, 0.2, 0.3, 0.4, 0.5	Quantized uniform
Learning rate	10^{-4}, 10^{-3} , 10^{-2} $5\cdot$ [10^{-4}, 10^{-3}]	Quantized uniform

where $\pi_2 = (n_{01} + n_{11})/(n_{00} + n_{01} + n_{10} + n_{11})$, n_{ij} is the number of observations with value i followed by j and $n_{ij} = \mathbb{P}(I_t^{(i)} = j | I_{t-1}^{(i)} = i)$. The conditional coverage test is then numerically related with the LR test statistics of UC and the independence tests as their sum $\text{LR}_{\text{CC}} = \text{LR}_{\text{UC}} + \text{LR}_{\text{IND}} \overset{\text{asymp.}}{\sim} \chi^2(2)$, if we condition on the first observation.

We conduct the Christoffersen tests for 80% PI (i.e., $c = 0.8$, $\alpha = 0.2$, $\underline{\tau} = 0.1$, $\overline{\tau} = 0.9$) of the last 24 h of the day-ahead forecast separately for each hour and the univariate time series. Then, the results of the LR statistics are presented as mean values per a set of multivariate time series.

The *sharpness* is evaluated with a bidirectional wQL, $\tau \in (0,1)$, denoted as quantile risk in [9]:

$$\text{wQL}_\tau(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = 2 \frac{\sum_{i,t} \text{P}_\tau(y_t^{(i)}, \hat{y}_t^{(i)})}{\sum_{i,t} |y_t^{(i)}|}, \tag{12}$$

where the quantile loss per time index is defined as:

$$\text{P}_\tau(y_t^{(i)}, \hat{y}_t^{(i)}) = \begin{cases} \tau(y_t^{(i)} - \hat{y}_t^{(i)}), & \text{if } y_t^{(i)} > \hat{y}_t^{(i)} \\ (1-\tau)(\hat{y}_t^{(i)} - y_t^{(i)}), & \text{otherwise.} \end{cases} \tag{13}$$

For the statistical validation of the wQL metric, we apply Diebold–Mariano tests with the arrangements described for the accuracy evaluation (see Section 3.4.1).

For more details about the evaluation of probabilistic forecasting, the reader is referred to [56]. In this study, only 0.1 and 0.9 quantiles were used for the wQL evaluation as it is done in the reference models [9,18]. These metrics were selected because they are common in DL research and can provide full-stack evaluation of the model generalization abilities for deterministic and probabilistic multivariate forecasts. For all the error metrics, lower values indicate a better model performance.

3.5. Hyperparameter optimization

In the selected models, the optimal values for hyperparameters were chosen based on a grid or manual search, but no information about the effects of these parameters on the model performance was given. The hyperparameter optimization in this study aims to reveal the sensitivity of the DL models to the most common hyperparameters and investigate their optimal values in a day-ahead forecasting scenario. This experiment was conducted by sequential model-based optimization with the Tree Parzen Estimator algorithm that selects the next hyperparameters based on Bayesian reasoning [57].

The search space of the selected hyperparameters is presented in Table 3 and based on the outer intersection of the common hyperparameters initially used for the grid search or manual tuning of the models. The hyperparameters include the number of hidden units in the recurrent or dense layers, batch size, dropout rate, and learning rate. The number of hidden units is controlled in the dense layers with DeepTCN and DSANet and in the recurrent layers in the case of DeepAR and LSTNet. Moreover, the maximum batch size was limited to 2^8 for the DSANet model owing to GPU memory issues, but batch sizes 2^4 and 2^5 were added to the search space in order to keep the search space more consistent with the other models. The effect of input sequence length on prediction accuracy was ignored using the look-back window of 168 h for all the models. The rest of the parameters

were not fitted and used as default values for the datasets with the same resolution and properties in the corresponding models to preserve the model architecture.

The number of iterations for the sequential optimization was limited to 100 because of its high computational requirements [58]. Compared with the training conditions, the maximum number of epochs and early stopping criteria were decreased by a factor of five, i.e., to 100 and 5. Moreover, the dataset was reduced to 11% for training and about 4% for validation and testing. The loss function for the sequential optimization is defined by the best mean absolute error obtained on the validation set. The number of epochs, dataset length, and stopping criteria were decreased with the intention to obtain close-to-optimal values under lower computation requirements. The models during the hyperparameter optimization were examined from the viewpoints of run-time efficiency by total simulation time, mean and standard deviation of the GPU memory, convergence rate, and estimated energy consumption. The simulation time measures the wall-clock time of dataset preprocessing and model training and evaluation for all iterations. The convergence rate is equal to the average number of epochs required to find the best solution, after which the early stopping criterion is reached. The energy consumption is estimated as the maximum power of the peripheral component interconnect express (PCIe) card (300 W) multiplied by the simulation time (h) and the proportion of mean GPU memory from the maximum GPU memory (32 GiB).

3.6. Model sensitivity

Besides the hyperparameter tuning, we also studied the model sensitivity to exogenous variables and fieldwise split based on the day-ahead forecasting problem with the `open power system` dataset. The former experiment consisted of removing or adding the calendar and time exogenous variables from and to the models. For the DeepAR and DeepTCN models with these variables used by default, we left only categorical variables, but removed the calendar and time features. For LSTNet and DSANet, the exogenous variables were added as the input sequences that were then retrieved from the error analysis. The idea of fieldwise split experiment was to validate the hypothesis that the DL models can extract the correlations of the time series from the related fields of power system operations, e.g., the total load and price from the same market bidding area. This experiment was conducted by splitting the `open power system` dataset in a fieldwise manner (i.e., load, price, wind, and solar fields), and training and validating the models separately on each of the obtained subsets of data.

3.7. Experiment environment

The experiments were executed on a node of an Intel Xeon processor running at 2.1 GHz (Xeon Gold 6230) and with 4 NVIDIA Tesla V100-SXM2 GPUs containing 32 GiB of memory each.

4. Empirical results

4.1. Accuracy, reliability, and sharpness assessment

The test results for the assessment of the average model accuracy and sharpness of the `electricity` and `open power system` datasets are shown in Figs. 5a–5b with ND and NRMSE and weighted quantile loss (wQL10 and wQL90)[3]. The observations suggest that the best score in ND with the `electricity` dataset is achieved by the local FFNN model. The other models (e.g., DSANet, DeepAR, and naïve benchmarks) follow closely by, whereas LSTNet and DeepTCN constitute a group with the largest ND. The score distribution for the NRMSE is more volatile with different winning methods for specific horizons but still distinguishes two main groups with different performance. In contrast to the ND metric, the Naïve-2 model has shifted to the worst performing group, the DeepAR results have worsened, and the rest generally remain unchanged. The sharpness evaluation of the model forecasts repeated the ranks of model performances on NRMSE with the exception of the DSANet model that demonstrated low-quality sharpness for the lower (wQL10) and upper (wQL90) quantiles. Overall, the models have more difficulties to forecasts the upper quantile of the datsets.

With the `open power system` dataset, the best accuracy and sharpness are demonstrated by DeepAR. Together with DeepTCN, these models outperform all benchmarks, and in contrast to the previous dataset, all DL models outperform the naïve benchmarks except in a few cases for the 36 h ahead forecast horizon. The error variation and wQL are weakly dependent on the time horizon in the `electricity` dataset, whereas they are more strongly correlated with the `open power system` dataset. For this dataset, the NRMSE values tended to be lower than in the `electricity` dataset, which can be explained by the different granularity levels of the datasets, where individual electricity metering data have more rapid peak values. Overall, the poor LSTNet and DSANet results in wQL in both datasets suggest that the risk of uncertainty estimation obtained with the MC approximation tends to be higher than with the other methods. In relation to the benchmarks for both datasets, the larger was the forecast horizon, the closer was the performance of the naïve models to the best performing models, and the local FFNN performed well ranking first and third for the datasets, respectively.

The test results were also examined separately for each of the variable types in the `open power system` data in Figs. 5c–5f. The Load time series were predicted with the highest accuracy and smallest quantile risks at both quantiles. For example, DeepAR, as the best performing model, achieved ND and wQL less than 4% and 2%, respectively, for the 36 h ahead forecasts. Interestingly, the Naïve-2 model performed generally very well and close to the DeepAR results for the 36 h forecast horizon. For the Price series, the error amplitude was higher with the best levels of accuracy and the quantile risk of 8%–10% and 4%–8%, respectively, along the horizon. The level of forecast accuracy and sharpness evaluation significantly deteriorated for renewable source variables compared with the Price and Load series. The best point forecast accuracy and sharpness for Solar time series forecasts was shown by the DeepAR and DeepTCN models followed by FFNN. The level of ND, and wQL varied between 8%–20% and 5%–13%. The stochasticity of the Wind time series causes a rapid deterioration of the forecast quality along the horizons, which can be the reason for the higher error dependence with the forecast horizon. In particular, the wQL varied from 6% to 28%, whereas ND varied from 8% to 35% along the forecast horizon. As it was assumed in Section 3.2.1, the Naïve-1 benchmark performed better than Naïve-2 for renewable generation forecasts. In general, the DL models demonstrated superior performance for intraday forecasts of renewable generation over the benchmark models. On the other hand, the advantages of DL models for the Load and Price series were marginal. In addition, DeepTCN and DSANet showed unstable results for the Price and Solar time series.

The results of empirical coverage for one-sided prediction intervals of 0.1 and 0.9 quantiles on the `electricity` and `open power system` datasets are illustrated in Fig. 6 with the average coverage error (ACE, see Section 3.4.2) for all forecast horizons. In the figure, the error is indicated by red color with respect to the 0.1 and 0.9 quantiles (marked with a dashed line) that cover 80% PI (marked in gray). For one-sided intervals, ACE is negative when the quantile forecasts are biased lower than the required quantile. In this case, the red area is located to the left from the corresponding quantile dashed line and to the right for the positive ACE.

Nearly all the models yield a small ACE for the whole `electricity` and `open power system` datasets except the LSTNet and DSANet models

[3] Numerical data of the experiment results is available in Appendix A.

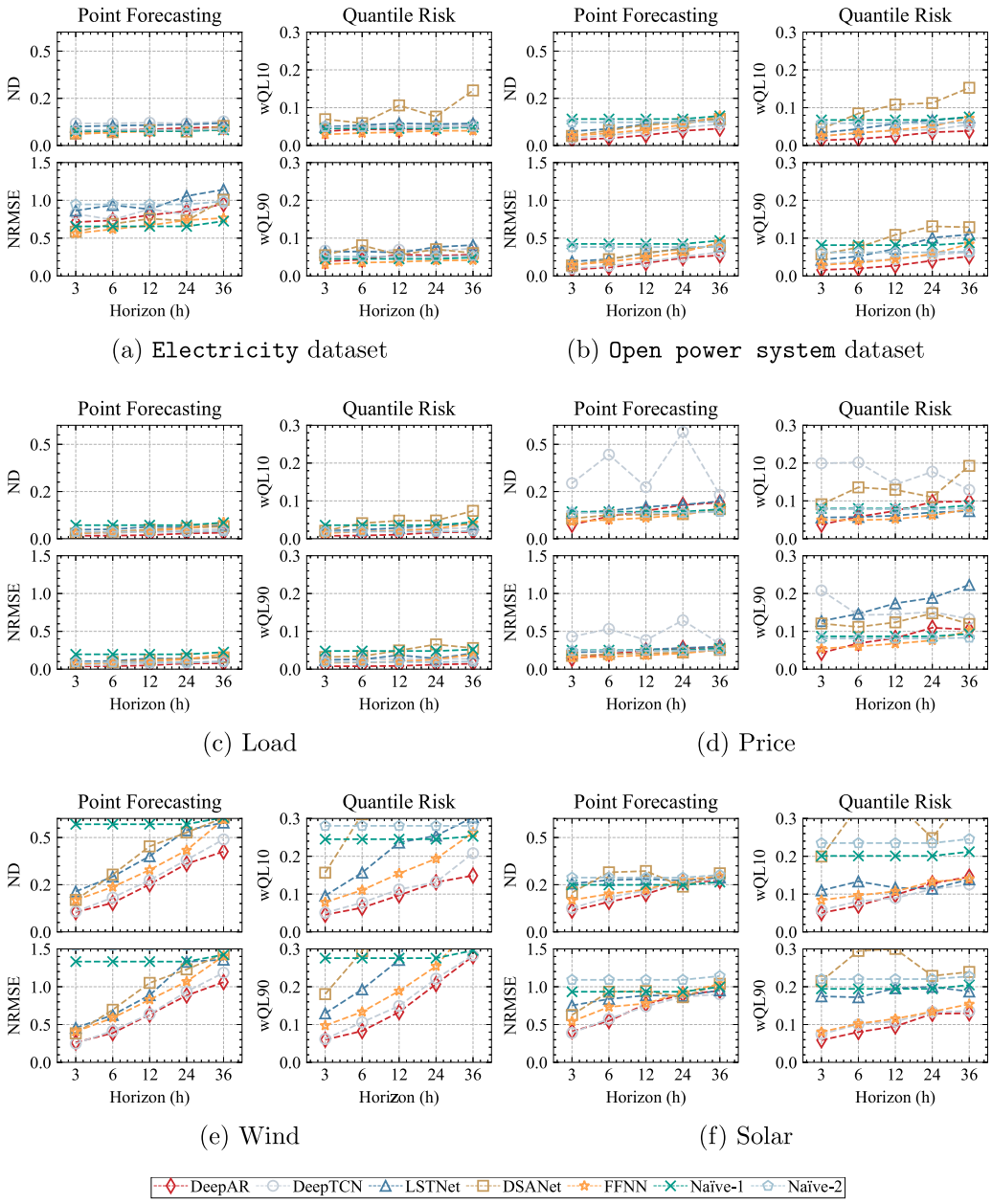

Fig. 5. Best point forecast losses and quantile τ-risk achieved by the models for the `electricity` (a) and `open power system` (b) datasets as well as for the Load (c), Price (d), Wind (e), and Solar (f) related fields of the `open power system` dataset. ND: normalized deviation, NRMSE: normalized root mean square error. QL10 and QL90: quantile risks at 0.1 and 0.9 quantiles, respectively. The lower metric values indicate a better model performance.

with narrow and overly narrow PIs, respectively, which were obtained by the MC dropout. In Figs. 6c–6f, we can observe the model reliability per a particular variable, which excludes complementary coverage compensation by forecasts of different variables. For example, the quantile forecasts of FFNN have wider coverage levels than the nominal for Load, but narrower than the nominal for Price. The LSTNet model has generally narrow prediction intervals and, especially, difficulties in capturing peak values of Price and Load, which leads to a large negative ACE for the upper quantile of these variables. The coverage of DSANet has also a very large ACE, and, together with LSTNet, these models underestimate the uncertainty and the least reliable for probabilistic forecasting. DeepTCN has generally a small ACE, but the sign of this error varies per variable and horizon supporting the observations above about the quantile loss. Surprisingly, the naïve benchmarks have the lowest ACE among all the models, but DeepAR and FFNN follow closely by.

4.2. Tests of statistical significance

In Fig. 7, we summarize the results of the DM tests for point (ND and NRMSE) and probabilistic (wQL) forecast metrics with a binary heat map. The map is divided into model areas, where each area contains binary indicators per a particular horizon. A red (blue) square in the map indicates that forecasts of a model on the x-axis are significantly better (worse) than the forecasts of a model on the y-axis for a particular horizon. In other words, if in a given area each diagonal square is red, then the forecasts of models on the x-axis are significantly better than those of the model on the y-axis for all horizons. For example, in Fig. 7b, the DeepAR model has columns with areas consisting of diagonal red squares, which indicates that the model is significantly better than any other model for each metric. In contrast, the models with blue columns are significantly worse, e.g., DeepTCN for ND in Fig. 7a. If the square is absent in the area, it indicates that the forecasts are not significantly

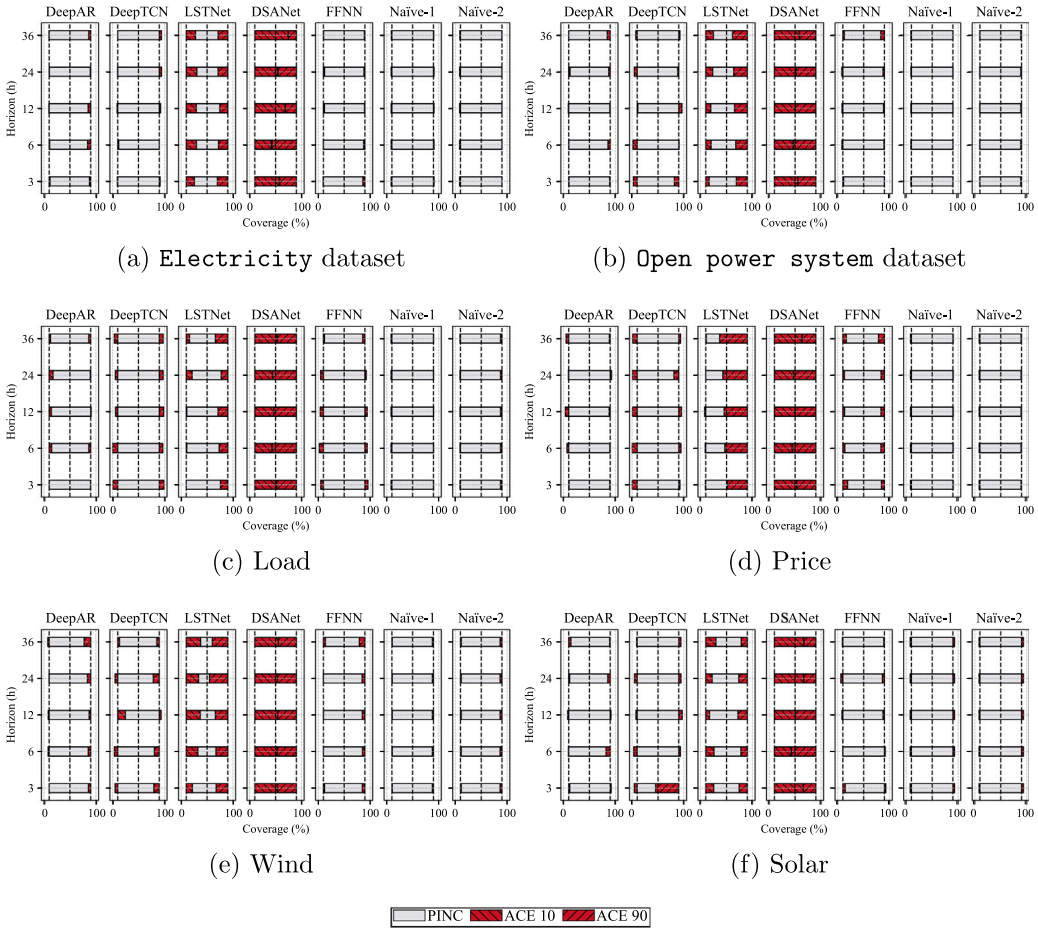

Fig. 6. Results of empirical coverage on the `electricity` (a) and `open power system` (b) datasets per particular horizons (from bottom to top: 3, 6, 12, 24, and 36 h). The red area corresponds to the average coverage error (ACE) from 0.1 and 0.9 quantiles, whereas the gray area corresponds to the nominal coverage of the 80% prediction interval (PI) covered by these quantiles.

different for a particular horizon. Similarly, the diagonal line is empty because it concerns the same model on both axes. The lower and upper triangles show the opposite performance as a result of the standard and complementary hypothesis testing.

For the `electricity` dataset, the DM statistic suggests that the forecasts of the FFNN model are significantly better than the forecasts of all the other models for the wQL metric, whereas they have comparable results for ND and NRMSE with the forecasts of the DSANet and naïve models. Among the global DL models, the forecasts of DSANet are ranked the first based on the ND and NRMSE metric, but they perform significantly worse than the other model forecasts for wQL loss. Interestingly, the forecasts produced by the FFNN and Naïve-1 benchmark models demonstrate higher point forecast accuracy and a lower quantile risk than the forecasts of the global DL models with a few exceptions in relation to the DeepAR and DSANet forecasts. The DM results for the `open power system` dataset are more in favor of the forecasts by the DL models. In particular, the DeepAR forecasts are significantly better than the others in all cases, whereas the DeepTCN forecasts are ranked second based on the accuracy metric, but have a comparable quantile loss with the forecasts of the FFNN model. The forecasts of DSANet and LSTNet are better than the naïve benchmarks for intraday horizons, but fall short of the FFNN forecasts. Overall, the results of statistical significance are in line with the assessment of accuracy and sharpness in Section 4.1.

The results of the Christoffersen test on unconditional (LR_{UC}) and conditional (LR_{CC}) LR statistics (see Section 3.4.2) for 80% PI of the `electricity` and `open power system` datasets are presented in Fig. 8. We first conducted the tests separately for each variable in the dataset and for each of the 24 h of the day-ahead (36 h ahead) forecast and then presented the results as the mean of the LR statistics for all variables. In the `electricity` dataset, the naïve benchmarks yield the best unconditional coverage; see the green crosses and the light blue pentagons in the top plot of Fig. 8a, by passing both the 1% and 5% significance levels for all hours. DeepAR and FFNN demonstrate close performance, but only FFNN passes the 1% level, whereas the mean of DeepAR forecasts is slightly above this level. The rest of the DL models (i.e., DeepTCN, LSTNet, and DSANet) fall short to reject the null hypothesis. For the conditional coverage, the situation is mostly similar with a minor difference in the performance of the DeepAR and FFNN models that pass the 1% and 5% tests for certain horizons. Overall, the LR statistics does not reveal any pattern along the hours. For the `open power system` dataset, the naïve models remained the clear leaders in both unconditional and conditional coverages. However, in contrast to the previous dataset, none of the DL models managed to pass the significance test for the UC. For the CC, DeepAR reached 1% significance only for few hours, but DeepTCN and FFNN were close to this level. The performance of LSTNet and DSANet remained the worst among the models, whereas the DeepTCN model showed an improvement in both statistics and came close to DeepAR and FFNN for the CC. In contrast to the `electricity` dataset, the UC and CC of several models showed the coverage pattern with a dip during the daily hours.

4.3. Hyperparameter sensitivity

The search of sequential hyperparameter optimization for DeepAR, DeepTCN, LSTNet, and DSANet is illustrated in Fig. 9 with kernel

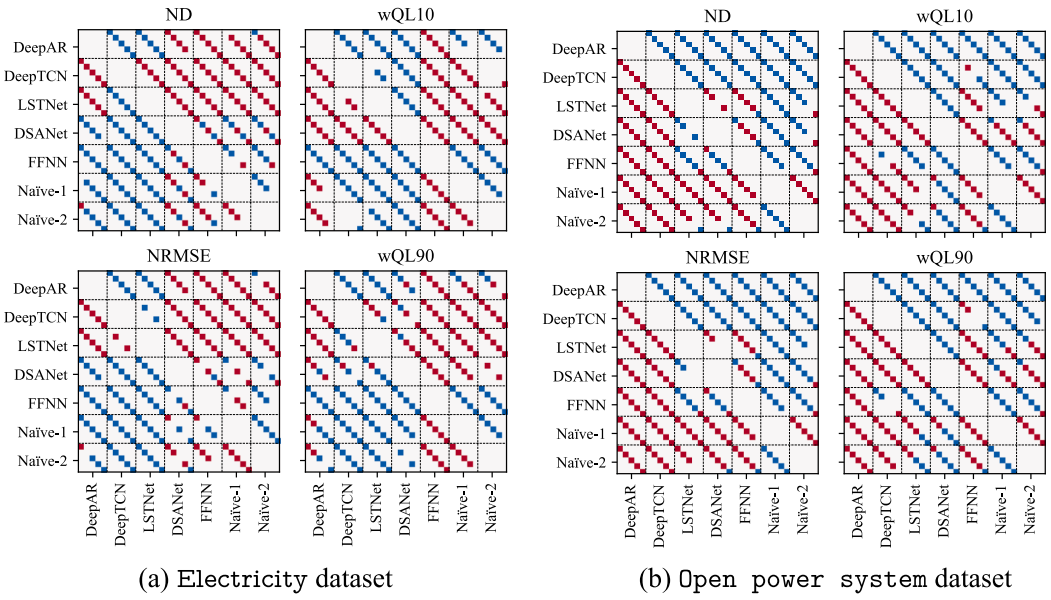

Fig. 7. Results of the conducted one-sided Diebold–Mariano tests at the 5% significance levels on the `electricity` (a) and `open power system` (b) datasets for point accuracy (ND and NRMSE) and quantile risk (wQL) of 0.1 and 0.9 quantiles per particular (from left to right by diagonal: 3, 6, 12, 24, and 36 h) horizons. A red (blue) square indicates that the forecasts of a model on the x-axis are significantly better (worse) than the forecasts of a model on the y-axis for a particular horizon, whereas an absence of square indicates that the forecasts are not significantly different for a particular horizon.

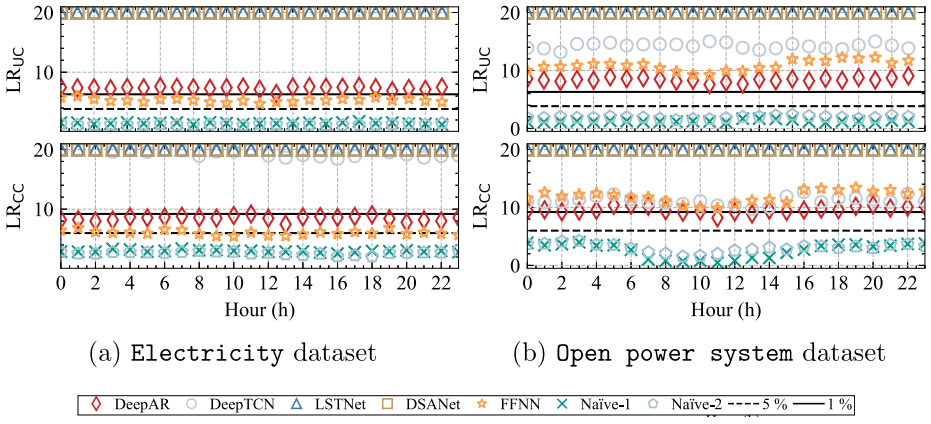

Fig. 8. Results of the Christoffersen tests on unconditional (LR_{UC}) and conditional (LR_{CC}) LR statistics for the `electricity` (a) and `open power system` (b) datasets. The statistics is obtained for day-ahead (36 h ahead) forecast of 80% prediction interval (PI) separately for each 24 h of the next day. The solid (dashed) horizontal lines represent the 1% (5%) significance level of the appropriate χ^2 distribution. All the test values exceeding 20 are set to 20.

density estimation and linear regression. The gray areas represent the estimated density of model hyperparameters based on the observed hyperparameter samples, and they are complemented by average values of these samples indicated by a dashed line. The number of hidden units (HU) has generally an increasing tendency with a few exceptions (e.g., DeepTCN and LSTNet in the `electricity` and `open power system` datasets, respectively), which is correlated with a decreasing dropout rate (DR) in many cases (except DSANet). However, one would expect positive correlation, i.e., more hidden units-higher dropout rate, as a way to prevent overfitting. The learning rate has mostly a positive correlation with the batch size, which can be seen as a measure to balance the gradient update distance of different batch sizes. Overall, the NDs demonstrated by the regression plots continuously decrease along the iterations, validating the correctness of the selected hyperparameters during the search.

The sensitivity of the hyperparameters to the ND of the models is described in Table 4. The results suggest that the DSANet model demonstrated more stable predictions compared with the other models, whereas DeepTCN had the highest sensitivity to the hyperparameters. The results may also indicate that there are not enough examples in the data for DeepAR, DeepTCN, and LSTNet for convergence. However, the results of DSANet are more in line with the ones presented in Fig. 5b.

Table 4
Error sensitivity of the models.

Dataset	DeepAR	DeepTCN	LSTNet	DSANet
`Electricity`	0.26 ± 0.107	0.38 ± 0.260	0.21 ± 0.142	0.14 ± 0.020
`Open power system`	0.24 ± 0.099	0.32 ± 0.281	0.33 ± 0.141	0.16 ± 0.070

4.4. Run-time efficiency

The results of the run-time efficiencies of the models including wall-clock simulation time, GPU memory usage, convergence rate, and electricity consumption are presented in Tables 5 and 6. The dimensionality of the dataset has almost a linear effect on the simulation

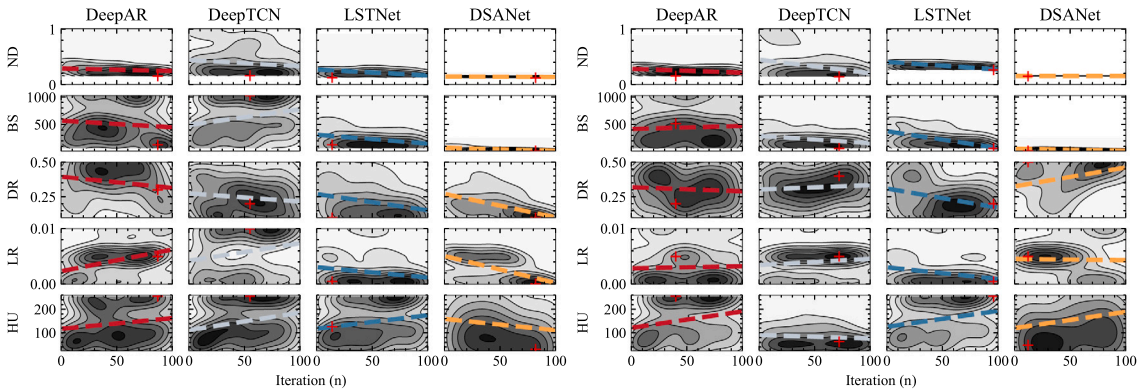

Fig. 9. Results of sequential model hyperparameter optimization for the `electricity` (a) and `open power system` (b) datasets. The darker the gray area, the higher the estimated kernel density of the model hyperparameters based on the observed samples, whose average is represented by the dashed line. ND: normalized deviation, BS: batch size, DR: dropout rate, LR: learning rate, HU: hidden units, +: the best result.

Table 5
Run-time efficiency of models during hyperparameter optimization for the `electricity` dataset.

Parameter	DeepAR	DeepTCN	LSTNet	DSANet
Simulation time, h:min	150:16	17:51	07:25	18:56
GPU memory, GiB	4.71	2.22	15.23	14.22
GPU memory, stdDev	1.06	0.22	0.98	2.24
Convergence rate, $\frac{\text{epochs}}{100}$	0.10	0.56	0.42	0.31
Energy consumption, Wh	6635	372	1059	2524

Table 6
Run-time efficiency of models during hyperparameter optimization for the `open power system` dataset.

Parameter	DeepAR	DeepTCN	LSTNet	DSANet
Simulation time, h:min	80:55	09:04	05:30	11:05
GPU memory, GiB	4.49	2.01	9.76	8.22
GPU memory, stdDev	1.02	0.14	0.97	6.4
Convergence rate, $\frac{\text{epochs}}{100}$	0.09	0.47	0.44	0.17
Energy consumption, Wh	3406	171	503	854

time for the DeepAR and DeepTCN models and on the GPU usage of the LSTNet model, whereas with DSANet there appears to be a linear effect with both of them. DeepAR with a likelihood model has a significantly reduced convergence rate compared with the quantile- and conditional-mean-based DeepTCN, LSTNet, and DSANet models. Interestingly, the most energy-consuming model consumes approximately up to 20 times more energy than the least consuming model.

4.5. Effect of exogenous variables

The effect of exogenous variables (i.e., calendar and time) on the accuracy and sharpness of forecasts in the `open power system` dataset is visualized in Fig. 10a. DeepTCN appeared to be the most reliant on exogenous variables, and the reliance was the most evident in the case of solar power generation data. However, the results of DeepAR, LSTNet, and DSANet were not affected that much. The reason for the low influence on the results of at least the LSTNet and DSANet models can be a method of integration of these calendar and time variables through input time series. A higher influence on the forecasting performance can be achieved if these variables are more comprehensively merged into the model architectures.

4.6. Cross-field dependence

The benefit of fieldwise prediction instead of simultaneously predicting the whole `open power system` data is visualized in Fig. 10b. All the models benefited slightly from the fieldwise split in both the point accuracy and the quantile loss, whereas DSANet had a substantial gain in NRMSE but a loss in wQL10. Overall, the results indicated that it is more beneficial to use the fieldwise split instead of predicting with the whole data. This is in conflict with the initial hypothesis that several time series from related fields can improve forecasting when using cross-series dependences. However, there are a few exceptions: for example, DeepAR would achieve better results for wind generation and slightly better results for load prediction when using the whole dataset. The results can be further verified if using the dataset from an isolated power system with more closely coupled processes.

5. Discussion

5.1. Model performance

5.1.1. Accuracy

An analysis of the DL model point accuracy on the presented datasets suggests that the performance of the studied architectures can be data- or context-specific. In particular, the global DL models can perform well on the homogeneous `electricity` dataset, e.g., DeepAR and DSANet, but their feasible superiority is more evident on the heterogeneous `open power system` dataset, e.g., DeepAR and DeepTCN. We assume that the latter is possible if the global model truly enables cross-learning of the dependences across multiple time series of the dataset.

These results of global model performances in comparison with the local benchmark model (FFNN) are in line with previous empirical evaluations [59], suggesting that global models are superior over local ones in forecasts for heterogeneous datasets, i.e., the global models can generalize better for unrelated simple patterns, but the local models can retrieve more complex patterns per similar series. However, the splitting experiment also slightly supports the more traditional assumption that the global models offer benefits over local methods only when the time series involve similar or related patterns.

Regarding the multihorizon forecast approaches, we investigated two types of models, i.e., many-to-one (DSANet and DeepAR) and many-to-many (DeepAR and DeepTCN). The results suggest that the forecast models with the many-to-one approach generally perform worse than the other. For example, DSANet has one of the lowest

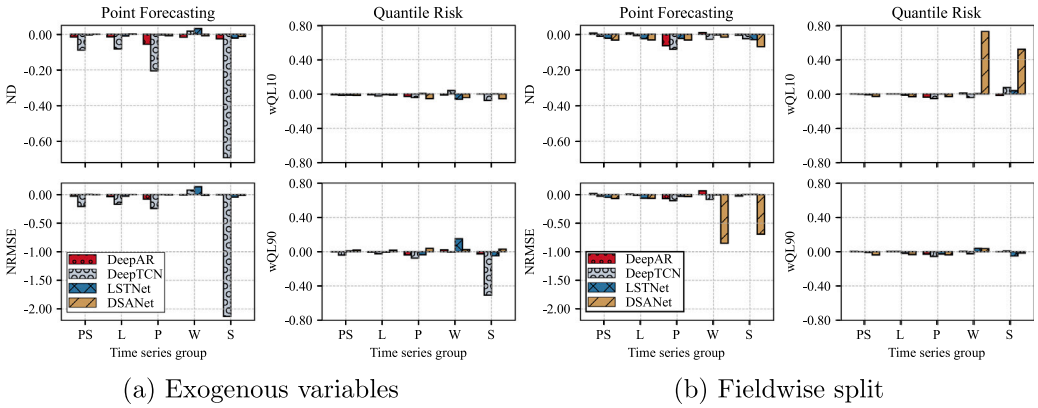

Fig. 10. Effect of time- and calendar-related exogenous variables (a) and fieldwise split (b) on losses in the `open power system` dataset. Negative values indicate that the results are better with exogenous variables or fieldwise split, and positive values the opposite. PS: whole dataset, L: Load, P: Price, W: Wind, S: Solar.

accuracies for the `open power system` dataset but one of the best for the `electricity` dataset, whereas LSTNet is the least accurate for both (see Fig. 5). In theory, this difference in the performances of the many-to-one models can be explained by the absence of a recursive strategy in DSANet and its presence in LSTNet. The recursive strategy can accumulate errors of previous steps along the forecast horizon and make the overall accuracy worse. However, it is unlikely the case for the experiments carried out in this study because the initial error levels of LSTNet and DSANet in the `open power system` dataset are higher than the best-performing competitive models. Therefore, potentially other problems in the model architecture affect the performance and should be studied in the future.

5.1.2. Reliability and sharpness

In the study, four methods of generating probabilistic forecasts by the DL model were investigated: producing selected quantiles directly by the DL model (DeepTCN), fitting the DL parameters of probability distribution (DeepAR and FFNN), and applying "bootstrapping"-based MC dropout during the testing (LSTNet and DSANet). Moreover, a post-hoc residual simulation based on the point forecast was conducted for two naïve models.

The results suggest that the probability distribution and residual simulation methods provide a statistically significant performance improvement in the reliability and sharpness of forecasts. On the other hand, MC dropout probabilistic forecasts produce too narrow PIs, whereas quantile forecasts are unable to reliably capture the distribution of several groups of series. Therefore, further work should be done to investigate the reasons for these phenomena.

5.1.3. Run-time efficiency

The scalability of the DL models was questioned, and a linear dependence of the dataset dimensionality with the simulation time and the GPU resource usage was observed in some cases. Moreover, a hypothesis of a higher computational time needed for more accurate results can be partly supported with the results. Therefore, the choice of the model is mostly affected by the application requirements for users' specific needs and data dimensionality, and the decision can be prioritized based on the forecast accuracy, uncertainty risks, hardware limitations, or a trade-off between these conditions.

5.2. Limitations and implications

The trustworthiness of the research results can be questioned by several factors. For instance, although the length of OOS testing covered half a year with summer–autumn–winter periods for both datasets, it is less than the one-year OOS period recommended in the literature [52]. Moreover, the number of DL models and the number of datasets provide a sufficient view to the DL performance of the investigated models for energy forecasting but cannot necessarily be generalized to other models. Furthermore, the hyperparameter tuning was implemented based on a fraction of the datasets and only for a 36 h forecast horizon. This could potentially lead to suboptimal model training and forecasting results.

However, with respect to the present developments in the forecasting domain in general, and energy forecasting in particular, this study follows most of the state-of-the-art practices. For example, the study meets several requirements introduced by the recent well-known M5 forecasting competition [60] and energy forecasting literature [52]: a clearly defined application domain, assessing forecast uncertainty along with the point forecast accuracy, well-recognized evaluation methods with the assessment of their statistical significance, datasets with high-frequency hourly data, and existing cross-correlations between the multivariate series.

Moreover, the final results of the competition suggest that the advanced global DL models can become mainstream methods dealing with large datasets and motivate further research in this field [60]. Keeping in mind the data-specific performance of the investigated DL models, more automated testing of global models is required to progress the research, e.g., using open source libraries available in the DL community, such as the modeling toolkit GluonTS [51]. To support the reproducibility of the research, we share the code for the modifications and experiment arrangements of the applied and already publicly available DL models as well as publish the preprocessed `open power system` dataset.

6. Conclusion

This study bridges the gap between the adoption of novel global deep-learning-based models for probabilistic multivariate forecasting in the deep learning community and the applicability of these methods for energy forecasting. In particular, this work provides insights for the academia, industry specialists, and practitioners into plausible levels of forecast accuracy and uncertainty that can be achieved in this context with the use of novel global deep learning models. Moreover, this study provides a numerical quantification of challenges that such methods have in terms of computational efficiency, hyperparameter sensitivity, and influence of exogenous and field-dependent variables.

In summary, the results suggest that a satisfactory level of accuracy and uncertainty risk can be achieved by global deep learning models for the forecasting of load, price, and solar time series exclusively based on historical data, but additional exogenous information is required to minimize the forecast loss of the more stochastic wind time series. Moreover, in comparison with the local models, the empirical results seem especially favorable for the applicability of these global models

for intraday forecasts and heterogeneous datasets. Furthermore, the results also indicate that with a large dataset where the fieldwise split of the data is feasible, the results can be slightly improved by splitting. In general, the addition of calendar and time exogenous variables has a minor but mainly positive effect on the model performances. A hyperparameter sensitivity test indicated that careful tuning or search of good hyperparameters should be carried out when selecting the model to be used to achieve the best results. These findings motivate further exploration of global deep learning models for probabilistic multivariate energy forecasting.

Interesting future research directions include integration of multivariate forecasts into decision-making under uncertainty risks, improvement of the deep learning models with privacy-preserving federated learning, and merger of numerical weather predictions into the model architectures.

CRediT authorship contribution statement

Aleksei Mashlakov: Conceptualization, Methodology, Software, Visualization, Data curation, Investigation, Formal analysis, Writing - original draft, Writing - review and editing. **Toni Kuronen:** Methodology, Software, Investigation, Formal analysis, Writing - review and editing. **Lasse Lensu:** Writing - review & editing, Funding acquisition, Resources, Supervision. **Arto Kaarna:** Writing - review & editing. **Samuli Honkapuro:** Writing - review and editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to acknowledge the support by the LUT Research Platform on Smart Services for Digitalisation (DIGI-USER), Finland, and CSC – IT Center for Science, Finland, for computing resources. Furthermore, we would like to thank Hanna Niemelä for proofreading this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.apenergy.2020.116405.

References

[1] Sweeney C, Bessa RJ, Browell J, Pinson P. The future of forecasting for renewable energy. Wiley Interdiscip Rev: Energy Environ 2020;9(2):e365.
[2] Bessa R, Moreira C, Silva B, Matos M. Handling renewable energy variability and uncertainty in power systems operation. Wiley Interdiscip Rev: Energy Environ 2014;3(2):156–78.
[3] Bessa RJ, Möhrlen C, Fundel V, Siefert M, Browell J, Haglund El Gaidi S, et al. Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry. Energies 2017;10(9):1402.
[4] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecasting 2006;22(3):443–73.
[5] Lenzi A, Steinsland I, Pinson P. Benefits of spatiotemporal modeling for short-term wind power forecasting at both individual and aggregated levels. Environmetrics 2018;29(3):e2493.
[6] Golestaneh F, Gooi HB, Pinson P. Generation and evaluation of space–time trajectories of photovoltaic power. Appl Energy 2016;176:80–91.
[7] Toubeau J-F, Bottieau J, Vallée F, De Grève Z. Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets. IEEE Trans Power Syst 2018;34(2):1203–15.
[8] Chakraborty K, Mehrotra K, Mohan CK, Ranka S. Forecasting the behavior of multivariate time series using neural networks. Neural Netw 1992;5(6):961–70.
[9] Salinas D, Flunkert V, Gasthaus J, Januschowski T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. Int J Forecast 2019.
[10] Wang Y, Smola A, Maddix D, Gasthaus J, Foster D, Januschowski T. Deep factors for forecasting. In: International conference on machine learning. 2019, p. 6607–17.
[11] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B Stat Methodol 2007;69(2):243–68.
[12] Cao L-J, Tay FEH. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans Neural Netw 2003;14(6):1506–18.
[13] Bauwens L, Laurent S, Rombouts JV. Multivariate GARCH models: a survey. J Appl Econom 2006;21(1):79–109.
[14] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55–67.
[15] Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S. Gaussian processes for time-series modelling. Phil Trans R Soc A 2013;371(1984):20110550.
[16] Lai G, Chang W-C, Yang Y, Liu H. Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st international acm sigir conference on research & development in information retrieval. ACM; 2018, p. 95–104.
[17] Shih S-Y, Sun F-K, Lee H-y. Temporal pattern attention for multivariate time series forecasting. Mach Learn 2019;108(8–9):1421–41.
[18] Chen Y, Kang Y, Chen Y, Wang Z. Probabilistic forecasting with temporal convolutional neural network. Neurocomputing 2020.
[19] Huang S, Wang D, Wu X, Tang A. DSANet: Dual self-attention network for multivariate time series forecasting. In: Proceedings of the 28th ACM international conference on information and knowledge management. 2019, p. 2129-32.
[20] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. 2016, p. 1050–9.
[21] Lim B, Zohren S. Time series forecasting with deep learning: A survey. 2020, arXiv preprint arXiv:2004.13408.
[22] Chang Y-Y, Sun F-Y, Wu Y-H, Lin S-D. A memory-network based solution for multivariate time-series forecasting. 2018, arXiv preprint arXiv:1809.02105.
[23] Assaf R, Schumann A. Explainable deep neural networks for multivariate time series predictions. In: IJCAI. 2019, p. 6488–90.
[24] Brusaferri A, Matteucci M, Portolani P, Vitali A. Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices. Appl Energy 2019;250:1158–75.
[25] Wang H-z, Li G-q, Wang G-b, Peng J-c, Jiang H, Liu Y-t. Deep learning based ensemble approach for probabilistic wind power forecasting. Appl Energy 2017;188:56–70.
[26] Wang K, Qi X, Liu H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. Appl Energy 2019;251:113315.
[27] Yang Y, Hong W, Li S. Deep ensemble learning based probabilistic load forecasting in smart grids. Energy 2019;189:116324.
[28] Sun M, Zhang T, Wang Y, Strbac G, Kang C. Using Bayesian deep learning to capture uncertainty for residential net load forecasting. IEEE Trans Power Syst 2019;35(1):188–201.
[29] Mashlakov A, Lensu L, Kaarna A, Tikka V, Honkapuro S. Probabilistic forecasting of battery energy storage state-of-charge under primary frequency control. IEEE J Sel Areas Commun 2019;38(1):96–109.
[30] Khodayar M, Mohammadi S, Khodayar ME, Wang J, Liu G. Convolutional graph autoencoder: A generative deep neural network for probabilistic spatio-temporal solar irradiance forecasting. IEEE Trans Sustain Energy 2019;11(2):571–83.
[31] Khodayar M, Wang J. Spatio-temporal graph deep neural network for short-term wind speed forecasting. IEEE Trans Sustain Energy 2018;10(2):670–81.
[32] Zhang H, Liu Y, Yan J, Han S, Li L, Long Q. Improved deep mixture density network for regional wind power probabilistic forecasting. IEEE Trans Power Syst 2020.
[33] Zhu Q, Chen J, Shi D, Zhu L, Bai X, Duan X, Liu Y. Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. IEEE Trans Sustain Energy 2019;11(1):509–23.
[34] Liu Y, Qin H, Zhang Z, Pei S, Jiang Z, Feng Z, et al. Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model. Appl Energy 2020;260:114259.
[35] Chen Y, Zhang S, Zhang W, Peng J, Cai Y. Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting. Energy Convers Manage 2019;185:783–99.
[36] Zang H, Liu L, Sun L, Cheng L, Wei Z, Sun G. Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. Renew Energy 2020;160:26–41.
[37] Hong T, Fan S. Probabilistic electric load forecasting: A tutorial review. Int J Forecast 2016;32(3):914–38.
[38] Wang H, Lei Z, Zhang X, Zhou B, Peng J. A review of deep learning for renewable energy forecasting. Energy Convers Manage 2019;198:111799.
[39] Van der Meer DW, Widén J, Munkhammar J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew Sustain Energy Rev 2018;81:1484–512.
[40] Chapados N. Effective Bayesian modeling of groups of related count time series. In: International conference on machine learning. 2014, p. 1395–403.

[41] Yu H-F, Rao N, Dhillon IS. Temporal regularized matrix factorization for high-dimensional time series prediction. In: Advances in neural information processing systems. 2016, p. 847–55.
[42] Sen R, Yu H-F, Dhillon IS. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In: Advances in neural information processing systems. 2019, p. 4838–47.
[43] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.
[44] LeCun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series. Handb Brain Theory Neural Netw 1995;3361(10):1995.
[45] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.
[46] Rangapuram SS, Seeger MW, Gasthaus J, Stella L, Wang Y, Januschowski T. Deep state space models for time series forecasting. In: Advances in neural information processing systems. 2018, p. 7785–94.
[47] Dabrowski JJ, Zhang Y, Rahman A. Forecastnet: A time-variant deep feed-forward neural network architecture for multi-step-ahead time-series forecasting. In: International Conference on Neural Information Processing. Springer; 2020, p. 579–91.
[48] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, editors. In: Proceedings of machine learning research, vol. 37, Lille, France: PMLR; 2015, p. 448–56.
[49] Alexander C. Market risk analysis, value at risk models, Vol. 4. John Wiley & Sons; 2009.
[50] Wiese F, Schlecht I, Bunke W-D, Gerbaulet C, Hirth L, Jahn M, et al. Open power system data–frictionless data for electricity system modelling. Appl Energy 2019;236:401–9.
[51] Alexandrov A, Benidis K, Bohlke-Schneider M, Flunkert V, Gasthaus J, Januschowski T, et al. GluonTS: Probabilistic and neural time series modeling in Python. J Mach Learn Res 2020;21(116):1–6, URL http://jmlr.org/papers/v21/19-820.html.
[52] Croonenbroeck C, Stadtmann G. Renewable generation forecast studies–review and good practice guidance. Renew Sustain Energy Rev 2019;108:312–22.
[53] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat 2002;20(1):134–44.
[54] Christoffersen PF. Evaluating interval forecasts. Int Econ Rev 1998;841–62.
[55] Kupiec P. Techniques for verifying the accuracy of risk measurement models. J Derivatives 1995;3(2).
[56] Nowotarski J, Weron R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renew Sustain Energy Rev 2018;81:1548–68.
[57] Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems. 2011, p. 2546–54.
[58] Mashlakov A, Tikka V, Lensu L, Romanenko A, Honkapuro S. Hyper-parameter optimization of multi-attention recurrent neural network for battery state-of-charge forecasting. In: EPIA conference on artificial intelligence. Springer; 2019, p. 482–94.
[59] Montero-Manso P, Hyndman RJ. Principles and algorithms for forecasting groups of time series: Locality and globality. 2020, arXiv preprint arXiv:2008.00444.
[60] Makridakis S, Spiliotis E, Assimakopoulos V. The M5 accuracy competition: Results, findings and conclusions. Int J Forecast 2020.