Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Relative evaluation of probabilistic methods for spatio-temporal wind forecasting

Lars Ødegaard Bentsen [a,*], Narada Dilp Warakagoda [a], Roy Stenbro [b], Paal Engelstad [a]

[a] *Department of Technology Systems, University of Oslo, P.O. Box 70, Kjeller, 2027, Viken, Norway*
[b] *Institute for Energy Technology, P.O. Box 40, Kjeller, 2027, Viken, Norway*

ARTICLE INFO

Handling Editor: Panos Seferlis

Dataset link: https://aistudio.baidu.com/aistudio/competition/detail/152/

Keywords:
Wind forecasting
Quantile regression
Variational inference
Graph networks
Johnson's SU distribution

ABSTRACT

Short-term wind power forecasting has become a de facto tool to better facilitate the integration of such renewable energy resources into modern power grids. Instead of point predictors, which produce single-value predictions for the expected power, probabilistic forecasts predict probability distributions over the expected power output or associated confidence intervals. In this study, three different parametric and non-parametric methods for uncertainty modelling in wind power forecasting were studied, namely quantile regression (QR), variational inference and a maximum likelihood estimation (MLE) method. Johnson's SU distribution was studied as a novel candidate for modelling wind power, which is a transformed normal distribution that exhibits both skew and heavy tails. This was one of the first studies to provide a thorough investigation of Johnson's SU distribution for uncertainty modelling in a complex deep learning framework for wind forecasting. It was found that Johsnon's SU likelihood and QR-based models significantly outperformed models using Gaussian likelihoods, based on a range of quantitative metrics to evaluate probability distributions and qualitative investigation of produced forecasts. Variational inference models using Johnson's SU likelihoods performed remarkably well, with near-perfect calibration and higher precision than models using any of the other methods for uncertainty modelling, as evaluated through the pinball loss, Average Coverage Error (ACE) and Prediction Interval Coverage Percentage (PICP) metric. With the superior performance of Johnson's SU likelihood models, the study mainly contributes to the literature by introducing another candidate distribution for probabilistic wind forecasting, which is analytical, unbounded and easy to integrate into modern deep learning frameworks.

1. Introduction

With growing pressure on the depletion of fossil fuel-based energy resources, the adoption of wind energy has experienced significant attention and rapid acceleration in recent years. Despite another year marked by the COVID-19 pandemic, 2021 emerged as the second-best year for wind energy, with 94 GW of added capacity globally (GWEC, 2022). Renewable energy resources such as wind and solar are inherently intermittent, which can make their integration into power systems challenging. Therefore, improved weather forecasting emerges as an important topic of research to provide more information on the expected power outputs from these resources in the near future. Physical models, such as numerical weather predictions (NWP), can provide very accurate forecasts for the medium- to long-term horizons, ranging from days to a few weeks ahead (Wang et al., 2021). However, a disadvantage of NWP models is that they require immense computing power to simulate the underlying physics (Bazionis and Georgilakis, 2021), while also being less accurate for the immediate short-term, a few minutes to hours ahead (Wang et al., 2021). Statistical and machine learning methods, which are trained on historical measurement data, have materialised as the most interesting for short-term wind forecasting in recent years (Wang et al., 2021; Aslam et al., 2021).

Even though accurate point forecasts of wind speed and power are necessary for the robust integration of wind power into energy systems, such methods do not provide any confidence intervals or measures of the uncertainty associated with predictions. In order to make more informed decisions using forecasting systems, probabilistic models that can provide some prediction of the intervals that the true values are expected to lie within for some associated confidence level should be developed (Bazionis and Georgilakis, 2021). Probabilistic forecasting models can broadly be categorised into parametric and non-parametric approaches. Parametric approaches assume some fully defined probability distribution for the likelihood of the data, such as a Gaussian or Beta distribution (Zhang et al., 2014; Bazionis et al., 2022).

[*] Corresponding author.
E-mail address: l.o.bentsen@its.uio.no (L.Ø. Bentsen).

https://doi.org/10.1016/j.jclepro.2023.139944
Received 11 August 2023; Received in revised form 12 November 2023; Accepted 26 November 2023
Available online 27 November 2023

Nomenclature

ACE	Average Coverage Error
ARIMA	Autoregressive Integrated Moving Average
BMA	Bayesian Model Averaging
CDF	Cumulative Density Function
CNN	Convolutional Neural Network
CRPS	Continuous Ranked Probability Score
DL	Deep Learning
ELBO	Evidence Lower Bound
FFN	Feed-Forward Network
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
KDE	Kernel Density Estimation
KL	Kullback–Leibler
LayNorm	Layer Normalisation
LSTM	Long Short-Term Memory
LUBE	Lower Upper Bound Estimation
MAE	Mean Absolute Error
MHA	Multi-Head Attention
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MPN	Message Passing Network
MSE	Mean Squared Error
NLL	Negative Log-Likelihood
NWP	Numerical Weather Prediction
PDF	Probability Density Function
PICP	Prediction Interval Coverage Percentage
PIAW	Prediction Interval Average Width
QR	Quantile Regression
RL	Reinforcement Learning
SCADA	Supervisory Control and Data Acquisition
VI	Variational Inference

Having mathematically simple probability density (PDFs) and cumulative distribution functions (CDFs) might make it easier to interpret results, while also providing the opportunity to leverage a long history of research within probability theory. However, the simplicity comes at a cost, where the assumed distributions will impose some bias on the models and might not perfectly reflect the true underlying distribution. Non-parametric approaches do not impose the same constraints on the likelihoods and might therefore be able to more accurately model the underlying true distributions given enough data.

Most studies that focus on probabilistic wind forecasting consider interval prediction, where the models will provide point predictions along with predictions of the expected upper and lower limits for the respective intervals (Bazionis et al., 2022). For instance, a wind speed forecasting model could have five outputs, corresponding to the point prediction for the most likely wind speed and the expected upper and lower wind speeds for the 50 and 95% confidence intervals. Wider intervals can inform the user that the predictions are less certain than for narrow intervals. Such additional information could be crucial to best decide on appropriate downstream actions, such as those related to the unit commitment.

Parametric models include techniques such as variational inference (VI) or maximum likelihood estimation (MLE) methods. Non-parametric approaches include quantile regression (QR), kernel density estimators (KDE) and ensemble methods. Since most studies on wind forecasting consider point estimators and many probabilistic systems have focused on non-parametric approaches, this study instead considers a comparison of different probabilistic models, namely QR, MLE and VI. Furthermore, the study also investigates the effect of a couple of different distributions for the parametric approaches, namely Gaussian and Johnson's SU distribution, of which the latter is yet to be thoroughly investigated for modelling wind power. The main contributions of this work can be summarised as:

- In this study, we investigate the feasibility of Johnson's SU distribution - a transformed Gaussian with four parameters to model skew and heavy tails. The distribution has a fully defined PDF and is unbounded, which makes it easy to integrate into modern DL frameworks. Johnson's SU distribution is yet to be extensively studied for probabilistic wind forecasting. The main novelty of this paper is a thorough analysis of Johnson's SU distribution, both when used in a simple MLE and more complex VI framework, compared to non-parametric QR models and Gaussian likelihoods.
- VI is investigated as a Bayesian and parametric approach that should be able to model both aleatoric and epistemic uncertainty. Such complex parametric methods are less studied for wind forecasting using deep learning. The study therefore aims to provide information on the relative merits of such a complex probabilistic framework, compared to more conventional MLE and QR methods.
- The study provides a detailed analysis and comparison of the performance of QR, VI, MLE and different likelihoods for wind power forecasting. The probabilistic methods are evaluated in a complex spatio-temporal DL framework based on graph neural networks with LSTM and Transformer sequence learners. To the best of the authors' knowledge, this is the first study in the literature to provide such a detailed analysis of the aforementioned probabilistic methods in a complex DL framework.

2. Related works

2.1. Time series forecasting methods

Moving average models, such as the autoregressive integrated moving average (ARIMA) model, represent some of the simplest, yet very effective, statistical forecasting methods used for wind energy (Kavasseri and Seetharaman, 2009). Simple machine learning methods have also been popularly applied to wind forecasting. Ren and Suganthan (2014) forecasted wind speeds using empirical mode decomposition followed by K-nearest neighbour (KNN) models as the principal predictors, while support vector regression (SVR) has also been popularly used as in Santamaría-Bonfil et al. (2016).

DL has experienced a surge of use cases in recent years, due to cheaper computing power and large datasets that facilitate the training of such models. Multilayer perceptrons (MLP) are the quintessential architecture that underpins modern DL. Various studies use MLPs as the principal predictor in wind forecasting models, either in isolation (Sfetsos, 2002) or in combination with other methods, such as an ARIMA model (Cadenas and Rivera, 2010) or some signal decomposition (Liu et al., 2013). Convolutional neural networks (CNNs) are primarily used for image analysis, where the data is ordered in a two-dimensional grid. By changing the 2D convolution to a 1D causal convolution along the temporal dimension, CNNs have also become popular for sequence analysis (Oord et al., 2016) and forecasting (Shang et al., 2022). Nevertheless, recurrent neural networks (RNN) are still more popularly employed for forecasting applications. The Long short-term memory (LSTM) cell and gated recurrent unit (GRU) both introduce gating into the vanilla RNN architecture. The gating improves the learning ability and retains information across longer sequences, making them very competitive contenders for use in wind speed and power forecasting (Wang et al., 2021; Salinas et al., 2020). The Transformer alleviates the need for recurrence and has been found to outperform RNN-based methods for a range of different applications (Vaswani et al., 2017), including wind forecasting (Wang et al., 2022a; Qu et al., 2022). The Transformer does not require all past information to be encoded in a single memory vector, which is one of the reasons why the LSTM and GRU networks might struggle for long and complex time series (Vaswani et al., 2017). In recent years, Reinforcement Learning (RL) has also been used as a different learning technique to develop DL-based forecasting systems for wind. Li et al. (2023) used an RL-based algorithm as the basic forecasting model, integrated with federated learning to avoid privacy issues. Similarly, Li et al. (2021) used a deep RL model for multi-period forecasting to be used in optimal energy scheduling.

Spatio-temporal models consider time series recorded at different physical locations to learn global characteristics and correlations, which have been found to improve wind forecasting models (Bazionis and Georgilakis, 2021; Quan et al., 2019). Some studies assign each individual time series to a location in a grid-like structure, similar to pixels in an image, and leverage CNNs to extract spatial correlations (Zhu et al., 2019; Liu et al., 2020). However, such a rigid ordering of the spatial information might not be the best method for representing the spatial locations, as measurement locations might have complex topologies, which do not closely follow a grid-like structure. Instead of CNNs, graph neural networks (GNNs), which operate on graph-structured data with arbitrary ordering, have therefore emerged as the most interesting architectures for modelling spatial correlations in wind forecasting in recent years (Khodayar and Wang, 2018; Liao et al., 2021).

2.2. Probabilistic forecasting methods

There are generally two sources of uncertainty for time series forecasting models, namely aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to the uncertainty and variability in the data that are a result of inherently random effects and cannot be reduced with more data. In the context of wind forecasting, this could for example be uncertainties that stem from poor sensory equipment or random component failures (Billinton and Huang, 2008). Epistemic, or model uncertainty, refers to the model's uncertainty in parameters and can be reduced with more data, information or different model architectures (Der Kiureghian and Ditlevsen, 2009). Methods where point predictors are made probabilistic by adding additional outputs to predict expected distribution parameters or quantiles can generally be argued to only consider aleatoric uncertainty, e.g. in MLE or QR. This is because there is no randomness introduced into the model architecture itself (Valdenegro-Toro and Mori, 2022). On the other hand, Bayesian approaches such as VI aim to learn posterior distributions for all model parameters. In this way, such models are able to capture both epistemic and aleatoric uncertainty, since the distribution over the weights should be able to reflect the epistemic (model) uncertainty.

KDE, QR and lower upper bound estimation (LUBE) are some of the most popularly used non-parametric methods to make machine learning models probabilistic in wind forecasting (Bazionis and Georgilakis, 2021; Bazionis et al., 2022). Various studies propose wind forecasting models based on QR and KDE, where KDE is used to estimate the full PDF given the predicted quantiles (He and Li, 2018; He and Zheng, 2018). Ensemble methods, where the predicted distributions are obtained from aggregating predictions from multiple model candidates, are generally non-parametric, but able to capture both aleatoric and epistemic uncertainty. Sloughter et al. (2010) proposed a probabilistic wind speed forecasting model using ensembles and Bayesian model averaging (BMA), while Liu et al. (2019) performed solar irradiation forecasting using ensemble convolutional GRU. Afrasiabi et al. (2020) proposed a probabilistic DL model based on a CNN and GRU with a special gradient-based loss function. The architecture produced mixture distributions and was found to outperform KDE and Monte Carlo Dropout models on two datasets for hour-ahead wind speed forecasts.

Various studies also train DL models with multiple outputs to predict distribution-specific parameters. For example in MLE, a model can learn to predict the mean and standard deviation for a Gaussian by minimising the negative log-likelihood. The most common and simplest distribution to fit using MLE is the Gaussian distribution, but many studies investigate the performance of other likelihoods, such as Gamma, Beta, Weibull and LogNormal (Bazionis and Georgilakis, 2021; Bazionis et al., 2022; Pobočíková et al., 2017). Pobočíková et al. (2017) compared four probability distributions for wind speed modelling and found a three-parameter Weibull distribution to achieve superior results. Wang et al. (2022b) proposed the AL-MCNN-BiLSTM model, where an asymmetrical Laplace distribution is assumed to characterise the uncertainty in wind power forecasts, which can model skew in either direction. The motivation for using the aforementioned probability distributions is generally that wind speed and power distributions have evident skew and heavy tails which cannot be accurately represented by a standard Gaussian likelihood (Zhang et al., 2014; Bludszuweit et al., 2008). However, a problem with the Gamma, Beta, Weibull and LogNormal distributions is that their PDFs are either bounded or can only model skew in a particular direction. Johnson's SU distribution is a transformed Normal distribution with four parameters to control shift, variance, skewness and heavy tails (Johnson, 1949). The use of Johnson's SU distribution for wind power modelling has been limited, Li et al. (2020) and Zhang et al. (2016). This study has therefore decided to investigate this distribution due to its simple analytical definition and shape versatility. These properties are hypothesised to be able to model the skewness of wind power distributions well, while also being fully defined and applicable in an MLE or VI framework.

VI has been less studied for wind forecasting. Wang et al. (2017) proposed a multi-kernel regression model trained using VI for wind power forecasting. Similarly, Liu et al. (2020) studied a ConvGRU model for spatio-temporal wind speed forecasting trained using VI, achieving superior results compared to Gaussian process regression (GPR) and hidden Markov model (HMM). Despite the large number of research on QR, there lacks substantial research on the comparison of non-parametric methods such as QR against parametric MLE and VI models, where the latter should also be able to explicitly model both aleatoric and epistemic uncertainties. Furthermore, Johnson's SU distribution emerges as an interesting candidate for wind power distribution modelling and should be compared against the Gaussian distribution for a couple of different parametric methods.

Overall, the contributions of this paper with respect to the literature can be summarised in two parts. First, there is a lack of research comparing non-parametric QR against parametric VI and MLE methods in complex spatio-temporal DL forecasting frameworks. In this study, we therefore perform a relative comparison of these probabilistic methods for a couple of different DL architectures, in order to establish distinct characteristics of the different methods. Secondly, the main novelty of this paper will be the investigation into the suitability of Johnson's SU distribution for uncertainty modelling in wind forecasting. This distribution seems to encompass the key properties desirable for modelling wind, namely the ability to model skew and heavy tails. Furthermore, Johnson's SU distribution seems particularly suitable for DL applications since it has analytically defined PDFs and is unbounded. With a lack of research investigating Johnson's SU distribution, this paper should advance the literature by studying a new candidate distribution for modelling uncertainty in wind forecasting.

3. Preliminaries

3.1. Neural architectures

3.1.1. LSTM

The LSTM unit is an alteration of the original RNN, where gating mechanisms enable the network to learn long-term dependencies by

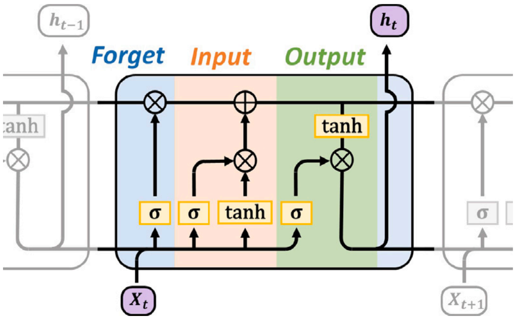

Fig. 1. Illustration of the LSTM gating mechanism.

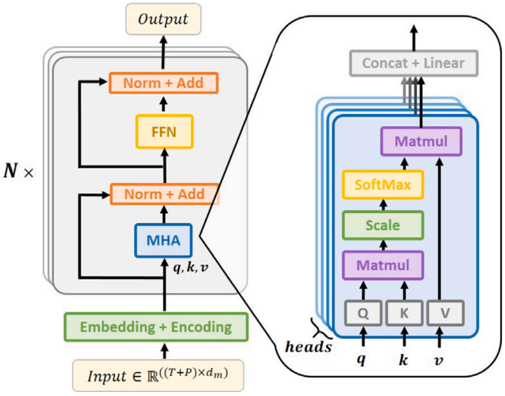

Fig. 2. Transformer encoder with multi-head attention (MHA) and position-wise feed-forward network (FFN).

selecting what information to write to and forget from the memory vector (Hochreiter and Schmidhuber, 1997). An illustration of the LSTM unit is given in Fig. 1 to show the internal workings of the input, forget and output gating mechanisms. The recurrent architecture can process variable length inputs, which are fed sequentially to encode information into the memory vector. To produce wind power forecasts, either an encoder–decoder or direct strategy can be used. In a direct strategy, as will be used in this study, the final output from the last LSTM layer, $\boldsymbol{h}_T \in \mathbb{R}^{1\times d}$, will be fed through an MLP or a single linear transform, $\boldsymbol{W}^{(out)} \in \mathbb{R}^{d\times P}$, as

$$\boldsymbol{h}_T = \text{LSTM}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_T) \tag{1}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{h}_T \boldsymbol{W}^{(out)}, \tag{2}$$

where $\boldsymbol{X}_t$ are the inputs at time t and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{1\times P}$ are the forecasts for P future times $T+1, T+2, \dots, T+P$.

3.1.2. Transformer

The Transformer architecture was proposed as an alternative to RNNs for sequence modelling (Vaswani et al., 2017). Fully reliant on the attention mechanism, the Transformer alleviates the need for recurrence and avoids the problem of vanishing or exploding gradients, making it better at learning complex long-term dependencies. Due to the lack of recurrence, positional encoding has to be introduced into the inputs, often through a sine–cosine embedding as in the original architecture. A single layer is comprised of an attention module, residual connections and position-wise feed-forward networks (FFN). The computation in a single layer, l, can be summarised as

$$\text{Attn}(\boldsymbol{X}^{(l)}) = \text{softmax}\left(\frac{(\boldsymbol{X}^{(l)}\boldsymbol{Q}^{(l)})(\boldsymbol{X}^{(l)}\boldsymbol{K}^{(l)})^T}{\sqrt{d}}\right)\boldsymbol{X}^{(l)}\boldsymbol{V}^{(l)} \tag{3}$$

$$\boldsymbol{Z}^{(l)} = \text{LayNorm}\left(\text{Attn}\left(\boldsymbol{X}^{(l)}\right) + \boldsymbol{X}^{(l)}\right) \tag{4}$$

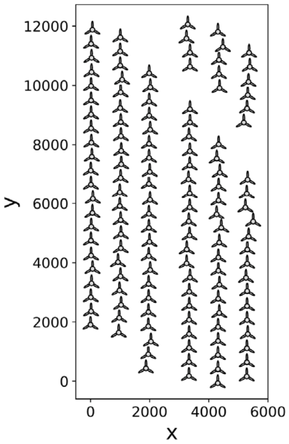

Fig. 3. Illustration of the spatial locations for the 134 turbines in the studied wind farm. The x and y coordinates are in meters.

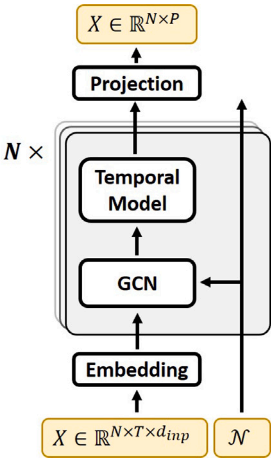
$X \in \mathbb{R}^{N\times P}$
Projection
N ×
Temporal Model
GCN
Embedding
$X \in \mathbb{R}^{N\times T\times d_{inp}}$
$\mathcal{N}$

Fig. 4. Spatio-temporal model architecture. Multiple layers of stacked GCN and temporal layers to extract spatial and temporal correlations, respectively.

$$\boldsymbol{X}^{(l+1)} = \text{LayNorm}\left(\text{FFN}^{(l)}\left(\boldsymbol{Z}^{(l)}\right) + \boldsymbol{Z}^{(l)}\right), \tag{5}$$

where $\boldsymbol{Q}^{(l)}, \boldsymbol{K}^{(l)}$ and $\boldsymbol{V}^{(l)} \in \mathbb{R}^{d\times d_k}$ are linear projections for layer, l, to queries, keys and values, respectively, and $\boldsymbol{X}^{(l+1)}$ the outputs from layer l. Typically, multi-head attention (MHA) is used, where multiple attention operations are performed in each layer before the outputs are concatenated and linearly transformed to produce the latent output, $\boldsymbol{Z}^{(l)} \in \mathbb{R}^{T\times d}$, where T is the sequence length and d the latent dimensionality. LayNorm and FFN are layer normalisation (Ba et al., 2016) and position-wise feed-forward networks, respectively. Multiple layers are typically stacked to increase the model's complexity and a visualisation of the architecture is given in Fig. 2. To make forecasts in an encoder setting, a sequence of length $T+P$ is typically fed to the model for T historical measurements and the P future times for which to forecast. Placeholders such as mean or last recorded values are used for the last P indices in the inputs.

3.1.3. Spatio-temporal forecasting with GNNs

In spatio-temporal forecasting, the aim is to improve prediction performance by leveraging correlated time series information from different physical locations. For our particular application of wind power forecasting, we have separate time series recorded for the 134 wind turbines shown in Fig. 3. An overview of our spatio-temporal framework is given in Fig. 4, where multiple layers are stacked, along with embedding and projection layers which will depend on the particular temporal model used. A single layer is comprised of a graph convolutional layer to learn spatial correlations, followed by a temporal network, such as an LSTM or Transformer layer, to learn temporal correlations as described in Sections 3.1.1 and 3.1.2.

The inputs are graph-structured, $\mathcal{G} = (\boldsymbol{X}, \mathcal{N})$. Input node features, $\boldsymbol{X}^{(0)} \in \mathbb{R}^{(N \times T \times d_{inp})}$, contain d_{inp} number of different features, such as wind speed, power and temperature, recorded for T previous time steps at N different nodes, i.e. 134 different turbine locations. The graph's connectivity is described through the index set, $\mathcal{N}$, where $\mathcal{N}_i$ contains the indices of all nodes sending to node i. For all experiments described in this study, the index set $\mathcal{N}_i$, will contain the 9 closest neighbouring turbines, based on Euclidean distance, as well as self-connections. Considering the simplest graph convolutional update, i.e. a vanilla message-passing network (MPN), a single update can be summarised as:

$$\boldsymbol{x}_i^{(l)} = \phi\left(\boldsymbol{x}_i^{(l-1)}, \bigoplus_{j \in \mathcal{N}_i} \psi(\boldsymbol{x}_i^{(l-1)}, \boldsymbol{x}_j^{(l-1)})\right), \tag{6}$$

where ϕ and ψ are the update and message functions, respectively, such as a linear transform or an MLP, and $\boldsymbol{x}_i^{(l)}$ are the updated features for node, i, after layer, l. To aggregate the neighbourhood information, $\bigoplus$ is a permutation invariant aggregator, such as a sum or mean operation. A range of more advanced graph updates exist, such as graph attention networks (GAT) (Veličković et al., 2017), where an attention mechanism will weigh the influence of neighbouring nodes differently, essentially turning the aggregator, $\bigoplus$, into a learnable weighted average. GATv2 (Brody et al., 2021) is an improved version of the original GAT and is the focus of this study. Here, a single graph layer update becomes:

$$\alpha_{ij} = \frac{\exp\left(\boldsymbol{a}^T \text{LeakyReLU}\left(\boldsymbol{W}^{(r)}\boldsymbol{x}_i^{(l-1)} + \boldsymbol{W}^{(s)}\boldsymbol{x}_j^{(l-1)}\right)\right)}{\sum_{j' \in \mathcal{N}_i} \exp\left(\boldsymbol{a}^T \text{LeakyReLU}\left(\boldsymbol{W}^{(r)}\boldsymbol{x}_i^{(l-1)} + \boldsymbol{W}^{(s)}\boldsymbol{x}_{j'}^{(l-1)}\right)\right)} \tag{7}$$

$$\boldsymbol{x}_i^{(l)} = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \boldsymbol{W}^r \boldsymbol{x}_j), \tag{8}$$

where $\boldsymbol{W}^{(\cdot)} \in \mathbb{R}^{d_k \times d}$, $\boldsymbol{a} \in \mathbb{R}^{d_k}$ and σ a non-linearity. Multiple attention heads are typically also used, where outputs from Eq. (8) will be concatenated and linearly transformed.

3.2. Uncertainty modelling

In this study, we chose to focus on three different methods for uncertainty modelling of wind power forecasts, namely quantile regression (QR), maximum likelihood estimation (MLE) and variation inference (VI) techniques. QR is non-parametric and can model any underlying distribution. VI introduces randomness into the model parameters and is the only technique that explicitly incorporates epistemic (model) uncertainty.

3.2.1. Quantile regression

In QR, the model should predict the conditional q-quantile of the dependant variable (Koenker and Hallock, 2001). For instance, a QR model can provide an estimate for $y^{(q)}$, with $q \in (0, 1)$, for which there is a probability q that the true value is smaller than $y^{(q)}$. To enable quantile predictions for a model f, we will have $|Q|$ outputs that correspond to the $|Q|$ different quantiles we want to predict, e.g. $Q = (0.1, 0.25, 0.5, 0.75, 0.9)$. Our QR model then becomes:

$$\hat{\boldsymbol{Y}} = f(\boldsymbol{X}, \theta), \tag{9}$$

where, $\hat{\boldsymbol{Y}} \in \mathbb{R}^{N \times P \times |Q|}$ contains the predicted quantile levels for the P future time steps at N spatial locations, $\boldsymbol{X} \in \mathbb{R}^{N \times T \times d_{inp}}$, θ are the model parameters and f a model such as the spatio-temporal architecture outlined in Fig. 4. Apart from additional outputs to predict the different quantile levels, nothing else changes in terms of the actual model architecture, compared to a point-prediction setting. However, since the true quantiles are not known, a special loss function is used to train a QR model, namely the pinball loss function:

$$L^{\text{pinball}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{q \in Q} \varphi_q(Y_{i,t} - \hat{Y}_{i,t,q}), \tag{10}$$

where

$$\varphi_q(u) = \begin{cases} qu & \text{when } u \geq 0 \\ (q-1)u & \text{when } u < 0. \end{cases} \tag{11}$$

3.2.2. Maximum likelihood estimation

Similar to the QR methodology, the only architectural changes for MLE estimation are additional outputs. However, instead of predicting quantile levels, an MLE model will predict distribution-specific parameters to fit a particular distribution to the data. For a Gaussian likelihood, the model will simply have two outputs to predict the expected mean and standard deviation of the forecasts, i.e. $\hat{\boldsymbol{Y}} \in \mathbb{R}^{N \times P \times 2}$ in Eq. (9). To train an MLE model, we minimise the negative log-likelihood (NLL). For a Gaussian likelihood, where $\hat{\boldsymbol{Y}}_{i,t,0}$ and $\hat{\boldsymbol{Y}}_{i,t,1}$ are the predicted means and standard deviations, respectively, for location, i, at a time, t, the NLL loss function becomes:

$$L^{NLL} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} \left(\frac{Y_{i,t} - \hat{Y}_{i,t,0}}{\hat{Y}_{i,t,1}} \right)^2 + \ln \hat{Y}_{i,t,1} \sqrt{2\pi}. \tag{12}$$

Since L^{NLL} computes the log probabilities, a requirement is that the distribution that we want to fit the data has a fully defined PDF that can be easily computed.

3.2.3. Variational inference

Bayesian methods are different to QR and MLE in that they propose prior distributions for the model parameters and update the belief about these as more observations are available. The updated distributions serve as the model's best guess for the posterior, $p(\theta|\mathcal{D})$. Since the true posterior is intractable, variational inference (VI) aims to find a distribution, $q(\theta)$, which is a good approximation of the posterior. The proposed distribution, also referred to as the guide, needs to be well-defined, meaning that it should be easy to sample from and that the PDF can be analytically computed. In VI, the distribution $q(\theta)$ is obtained by minimising the Kullback–Leibler (KL) divergence between $q(\theta)$ and the true posterior $p(\theta|\mathcal{D})$ (Blei et al., 2017):

$$q^*(\theta) = \arg\min_{q(\theta)} KL(q(\theta) \parallel p(\theta|\mathcal{D})). \tag{13}$$

However, since the posterior is intractable, we cannot obtain the optimal, $q^*(\theta)$, directly. Instead, we rearrange Eq. (13) to

$$\ln p(\mathcal{D}) = \mathbb{E}_{q(\theta)} \ln \frac{p(\theta, \mathcal{D})}{q(\theta)} + KL(q(\theta) \parallel p(\theta|\mathcal{D})). \tag{14}$$

Here, the evidence, $p(\mathcal{D})$, is difficult to compute but is constant. Therefore, the KL term in Eq. (14), can be minimised indirectly, by maximising the expectation. In VI we maximise this expectation, known as the evidence lower bound (ELBO), through gradient ascent, which can also be written as:

$$\text{ELBO} = \mathbb{E}_{q(\theta)} \ln p(\mathcal{D}|\theta) - KL(q(\theta) \parallel p(\theta)). \tag{15}$$

During training, the KL term in Eq. (15), can be quite a strong regulariser, resulting in KL vanishing and the model parameters converging to the prior distributions, known as posterior collapse. Training might therefore be improved by reducing the importance of the KL term or using different annealing schedules (Bowman et al., 2015). In this study, we experimented with a few different schedules, such as cyclic, step, linear, exponential and constant annealing. For the final results, all VI-based models used a constant schedule, multiplying the KL-term by 0.1, as this gave similar results to models using the other more complex annealing schedules. Without any annealing, training was more challenging, with some models resulting in posterior collapse, i.e. simply learning the prior distribution, p, for the guide, q.

Gaussian distributions are most commonly used for the prior and proposal distributions, $p(\theta)$ and $q(\theta)$, respectively, while the likelihood, $p(\mathcal{D}|\theta)$, is more domain specific. For Gaussian likelihoods, a model will have two outputs, as for the MLE method described in Section 3.2.2. Unlike the QR and MLE methods, VI is a true probabilistic method, since the aim is to learn distributions for all the parameters in the model, instead of scalar values. This means that the model will also learn the uncertainty associated with its parameters, and VI is, therefore, capable of explicitly modelling epistemic uncertainty.

3.3. Homoscedastic and heteroscedastic likelihoods

When the data noise is constant across the input domain, the noise can be classified as homoscedastic, while for heteroscedasticity, the noise will be input-dependent (Rogers et al., 2020). In our analysis, we will investigate both homo- and heteroscedastic likelihoods for the MLE and VI models. In the heteroscedastic setting, the models will have additional outputs to predict the input-dependent variance, as shown in Eq. (12). For homoscedastic likelihoods, the variance will not depend on the input. Considering a homoscedastic Gaussian likelihood, the models will only have one output to predict the mean and a single additional learnable parameter to represent the constant variance. The different methods for homo- and heteroscedastic settings with Gaussian likelihoods are summarised in Eqs. (16) and (17), respectively.

$$\text{Homoscedastic: } \hat{\boldsymbol{Y}}_{i,t} \sim \mathcal{N}\left(f(\boldsymbol{X},\theta)_{i,t}, \gamma\right) \tag{16}$$

$$\text{Heteroscedastic: } \hat{\boldsymbol{Y}}_{i,t} \sim \mathcal{N}\left(f(\boldsymbol{X},\theta)_{i,t,0}, f(\boldsymbol{X},\theta)_{i,t,1}\right), \tag{17}$$

where f is an arbitrary model, with either one or two outputs for the two settings, respectively, and γ is a learnable parameter to represent homoscedastic noise.

3.4. Johnson's SU distribution

Johnson's SU distribution is a four-parameter distribution with skew and heavy tails (Johnson, 1949). The distribution is unbounded, which makes it easy to work with since it can be applied to any data range, including negative values. Furthermore, the distribution is a transformed normal distribution and has analytically defined PDFs which makes it easy to compute the likelihoods required for MLE and VI models. Random variables, x, of Johnson's SU distribution can be generated as

$$\mathrm{x} = \xi + \lambda \sinh\left(\frac{\Phi^{-1}(\mathrm{u}) - \gamma}{\delta}\right), \tag{18}$$

where Φ^{-1} is the inverse CDF of a normal distribution and u a random variable uniformly distributed on [0, 1]. The distribution shift (analogous to mean of a standard normal), is controlled by ξ, $\lambda > 0$ controls the spread, γ the amount and direction of skew, while $\delta > 0$ is a shape parameter for which larger values result in heavier tails. The PDF is analytically defined as:

$$p(x) = \frac{\delta}{\lambda\sqrt{2\pi}} \frac{1}{\sqrt{1+z^2}} e^{-\frac{1}{2}(\gamma+\delta\sinh^{-1}(z))^2}, \quad \text{where } z = \frac{x-\xi}{\lambda}. \tag{19}$$

Examples of Johnson's SU PDFs, $\mathcal{J}_{SU}(\xi, \lambda, \gamma, \delta)$, for different skew and shape parameters are given in Fig. 5, along with a standard Gaussian with zero mean and 0.5 standard deviation, $\mathcal{N}(0.0, 0.5)$. Johnson's SU distribution was selected as an additional distribution to study for probabilistic wind power forecasting, due to its versatility in modelling a range of underlying distributions, with both skew and heavy tails. Since the distribution is unbounded, it is also easier to implement with modern DL techniques, as it does not impose any constraints on the value space, such as strictly positive values or upper and lower bounds. Finally, the lack of research applying Johnson's SU distribution for probabilistic wind forecasting also makes it a particularly interesting candidate to further investigate. Studying the Johnson's SU distribution could therefore potentially bring significant contributions to the literature, by proposing a flexible distribution to model uncertainty in wind forecasting, which is also easy to integrate into modern DL frameworks.

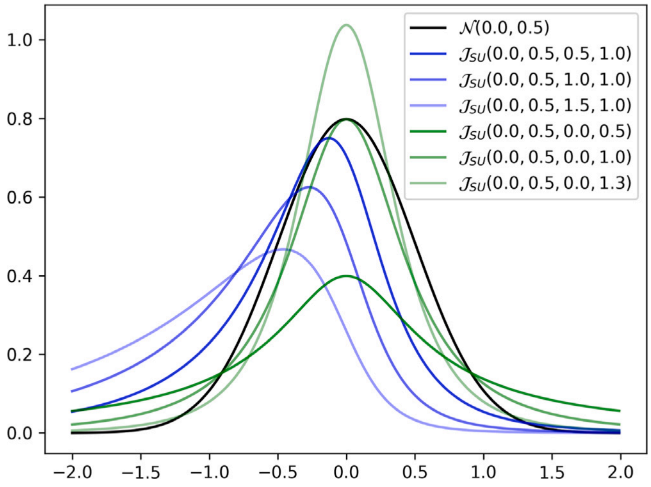

Fig. 5. Example of Johnson's SU PDF for different levels of skew (blue) and shape (green) parameters, along with a Gaussian PDF (black).

4. Experiments

4.1. Dataset

The dataset used for the study was taken from the KDD Cup22 forecasting competition (Zhou et al., 2022). The dataset contains 10-minute averaged Supervisory Control and Data Acquisition (SCADA) measurements on wind speed and direction, external and internal temperature, blade pitch, yaw and power outputs, recorded at the 134 different wind turbine locations shown in Fig. 3 for 245 days. First, corrupt measurements were removed according to the specifications (Zhou et al., 2022):

$$Pab \geq 90° \tag{20}$$

$$Wdir \notin [-180°, 180°] \tag{21}$$

$$Ndir \notin [-720°, 720°] \tag{22}$$

$$Patv < 0, \tag{23}$$

where Pab are blade pitch angles, $Wdir$ the wind direction, $Ndir$ the yaw angles, and $Patv$ the active turbine power output. The directional features, i.e. blade pitch, wind direction and yaw angle, were all decomposed into sine and cosine components to capture the circular characteristics. All features were scaled using a standard scaler to have zero mean and unit variance. The dataset was then split into training and validation sets, with a 90%–10% split, respectively. A separate test dataset was provided, used for the final test phase in the actual competition. Additional information and download links to the datasets can be found here.[1]

Different to the 48-hour (288 steps) ahead forecasts produced for the KDD Cup22 competition (Zhou et al., 2022), the main focus of this study will be on short-term forecasts, one hour ahead (6 steps). Forecasts will be produced for the expected power output from every turbine over the next hour in 10-minute resolution. The reason for choosing a shorter forecasting horizon was that this can be critical for

[1] https://aistudio.baidu.com/aistudio/competition/detail/152.

grid stability and short-term decision-making. Furthermore, it is for the shorter horizons that machine learning methods have seen the greatest improvements over NWP models.

4.2. Experimental setting

All models followed the spatio-temporal architecture in Fig. 4, with the GATv2 (Brody et al., 2021) as the GCN update in each layer. In order to conclude on general characteristics of the different probabilistic methods, two different temporal update functions were studied for each method, namely, a single LSTM or Transformer encoder layer. For the LSTM-based models, a direct strategy was used, as described in Section 3.1.1, with inputs $X^{(0)} \in \mathbb{R}^{N\times T\times d_{inp}}$, where d_{inp} are the number of input features and T the number of previous recordings used to produce forecasts for the future $P = 6$ time steps. The Transformer-based models on the other hand had inputs $X^{(0)} \in \mathbb{R}^{N\times (T+P)\times d_{inp}}$, where the first T vectors are recorded values for the last T time steps. We concatenate the last recorded values P times, which just act as a set of placeholders, with the above sequence of T vectors to form the input $X^{(0)}$.

Both spatio-temporal architectures, using either LSTM or Transformer updates as sequence learners, were tested for the different uncertainty modelling techniques described in Section 3.2. For QR, the models had 11 outputs, corresponding to the quantiles; Q = [0.05, 0.15, 0.25, 0.35, 0.45, 0.5, 0.55, 0.65, 0.75, 0.85, 0.95] in Eq. (10). Three different MLE models were tested for both LSTM and Transformer updates. The first two had heteroscedastic likelihoods using either the Gaussian or Johnson's SU distribution, resulting in 2 and 4 model outputs, respectively. The third MLE model had homoscedastic likelihoods, where the models predict the mean of a Gaussian distribution, with an independent learnable parameter to model the constant noise, as in Eq. (16). Exactly the same likelihoods were investigated for the VI models, i.e. for $p(\mathcal{D}|\theta)$ in Eq. (15), resulting in three different models for each of the sequence learners. All VI models had Gaussian proposal and prior distributions, i.e. for $q(\theta)$ and $p(\theta)$, respectively, in Eq. (15). To ensure distributional stability, a softplus activation was applied to the predicted standard deviations and λ for the Gaussian and Johnson's SU likelihoods, respectively. Values for the predicted skew and shape parameters, γ and δ, for the Johnson's SU likelihoods were also constrained using $\tanh(\hat{\gamma})$ and $\tanh(\hat{\delta})/2 + 1$. This was found to give better results as the distributions were constrained to sensible levels. Furthermore, at the beginning of training, the parameters for Johnson's SU distribution were initialised to 0.2, 0.0 and 1.0 for λ, γ and δ, respectively. This was done by multiplying the predicted outputs for the three parameters by a trainable parameter, initialised to zero. Such an approach has not been used in the current literature but was found to significantly stabilise training and reduce convergence time, because the models were initialised to a sensible starting point. Deterministic point predictors that do not model uncertainty were also implemented to be able to better evaluate the point prediction accuracy of the probabilistic models.

All models described so far used embedding layers, as shown in Fig. 4:

$$\boldsymbol{X}^{(0)} = \boldsymbol{X}_{inp}\boldsymbol{W}^{(e)} + \text{TempEnc}(\boldsymbol{ts}) + \text{PosEnc}(T), \tag{24}$$

where $\boldsymbol{W}^{(e)} \in \mathbb{R}^{d_{inp}\times d}$, $\boldsymbol{X}_{inp} \in \mathbb{R}^{N\times S\times d_{inp}}$ are the d_{inp} recorded features described in Section 4.1 for S time steps and N nodes. The TempEnc encodes temporal information using the available time-stamp information taken as the minute-of-hour and hour-of-day. Each time-stamp feature was decomposed into sine and cosine components to capture the circular characteristic, similar to what was done for the directional features described in Section 4.1, but with different frequencies. Considering a recording obtained at 18:20, the input feature to the TempEnc module would become:

$$\begin{bmatrix} -1 & 0 & 0.866 & -0.5 \end{bmatrix} = \begin{bmatrix} \sin(18\cdot(2\pi/24)) \\ \cos(18\cdot(2\pi/24)) \\ \sin(20\cdot(2\pi/60)) \\ \cos(20\cdot(2\pi/60)) \end{bmatrix}^T \tag{25}$$

The decomposed time-stamp information, $\boldsymbol{ts} \in \mathbb{R}^{N\times S\times 4}$, would then be fed to the TempEnc:

$$\text{TempEnc}(\boldsymbol{ts}) = \boldsymbol{ts}\boldsymbol{W}^{(ts)}, \tag{26}$$

where $\boldsymbol{W}^{(ts)} \in \mathbb{R}^{4\times d}$. The PosEnc in Eq. (24) was used only for the Transformer-based models, due to the lack of recurrence, and followed the same sine–cosine positional encoding proposed in the vanilla Transformer (Vaswani et al., 2017).

Inspired by Zeng et al. (2023), which argue that a simple linear transform of the inputs can be a competitive model, outperforming advanced Transformer architectures for long-term forecasting, we also implement a Linear baseline model as:

$$\hat{\boldsymbol{Y}}_{T+1}, \cdots, \hat{\boldsymbol{Y}}_{T+P} = \boldsymbol{Q}\boldsymbol{X}\cdot, \tag{27}$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times T\times d_{inp}}$ and $\boldsymbol{Q} \in \mathbb{R}^{P\times(T\cdot d_{inp})}$. The Linear model does not take into account any spatial correlations but was implemented for all the deterministic, QR, MLE and VI settings.

The input sequence length, T, for all models, was 48, i.e. features observed for the last 8 h. All models were iteratively tuned using a random grid search. First, a wide search space was used, before it was narrowed to obtain the final model parameters. With reference to Fig. 4, all models had $N = 3$ layers and a latent dimensionality, $d = 32$. A batch size of 32 was used and models were trained for 25 epochs. All models, along with model-specific parameters obtained from tuning, are summarised in Table 1.

4.3. Evaluation metrics

In order to ensure that the overall findings were robust and to reduce potential bias, a range of different metrics were used to evaluate the different models. The first metrics to evaluate prediction performance were the mean squared (MSE) and absolute errors (MAE), which only consider point predictions, $\hat{\boldsymbol{Y}}$:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)^2 \tag{28}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i|, \tag{29}$$

where $\boldsymbol{Y}$ are the observed measurements. Since the probabilistic models do not directly provide point predictions, but instead distributions or intervals, different methods are required to obtain point predictions from these models. For all models, point predictions were taken as the median value predictions, i.e. those associated with the quantile level, $q = 0.5$, where 50% of the data is expected to fall above and below the predicted values.

Since the VI models, unlike QR and MLE models, were not deterministic, we had to sample multiple predictions from these models to obtain a single estimate for the distribution-specific parameters. To combine aleatoric and epistemic uncertainty for VI models with Gaussian likelihoods, we used a weighted average approach based on the predicted noise (Ritter and Karaletsos, 2022; Kendall and Gal, 2017). The process can be summarised as

$$\hat{\mu}_{i,t} = \frac{\sum_{j=1}^{k}(\hat{\mu}_{i,t,j}/\hat{\sigma}^2_{i,t,j})}{\sum_{j=1}^{k}(1/\hat{\sigma}^2_{i,t,j})} \tag{30}$$

$$\hat{\sigma}_{i,t} = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(\hat{\sigma}^2_{i,t,j}) + Var(\hat{\mu}_{i,t})}, \tag{31}$$

where k is the number of samples drawn from the model, here 300, and $Var(\hat{\mu}_{i,t})$ the variance of predicted means across samples. For VI with Johnson's SU likelihoods, the models predict four parameters, for which the shift and spread parameters, ξ and λ, respectively, do not directly correspond to the mean and standard deviation. Therefore, we here

Table 1
Model parameters.

Model	GCN	Temporal model	Outputs	Loss	LR	Norm	FFN
Det-Linear	–	–	1	MSE	1e−04	–	–
Det-LSTM	GATv2	LSTM	1	MSE	1e−04	–	–
Det-Trans	GATv2	MHA	1	MSE	1e−04	ReZero	Yes
MLE-Gauss-Hetero-Linear	–	–	2	NLL	1e−04	–	–
MLE-Gauss-Hetero-LSTM	GATv2	LSTM	2	NLL	1e−04	–	–
MLE-Gauss-Hetero-Trans	GATv2	MHA	2	NLL	1e−04	ReZero	Yes
MLE-Gauss-Homo-Linear	–	–	1	NLL	1e−04	–	–
MLE-Gauss-Homo-LSTM	GATv2	LSTM	1	NLL	1e−04	–	–
MLE-Gauss-Homo-Trans	GATv2	MHA	1	NLL	1e−04	ReZero	Yes
MLE-Johns-Hetero-Linear	–	–	4	NLL	1e−04	–	–
MLE-Johns-Hetero-LSTM	GATv2	LSTM	4	NLL	1e−04	–	–
MLE-Johns-Hetero-Trans	GATv2	MHA	4	NLL	1e−04	ReZero	Yes
QR-Linear	–	–	11	pinball	1e−04	–	–
QR-LSTM	GATv2	LSTM	11	pinball	1e−04	–	–
QR-Trans	GATv2	MHA	11	pinball	1e−04	ReZero	Yes
VI-Gauss-Hetero-Linear	–	–	2	ELBO	1e−03	–	–
VI-Gauss-Hetero-LSTM	GATv2	LSTM	2	ELBO	1e−03	–	–
VI-Gauss-Hetero-Trans	GATv2	MHA	2	ELBO	1e−03	ReZero	Yes
VI-Gauss-Homo-Linear	–	–	1	ELBO	1e−03	–	–
VI-Gauss-Homo-LSTM	GATv2	LSTM	1	ELBO	1e−03	–	–
VI-Gauss-Homo-Trans	GATv2	MHA	1	ELBO	1e−03	ReZero	Yes
VI-Johns-Hetero-Linear	–	–	4	ELBO	1e−03	–	–
VI-Johns-Hetero-LSTM	GATv2	LSTM	4	ELBO	1e−03	–	–
VI-Johns-Hetero-Trans	GATv2	MHA	4	ELBO	1e−03	ReZero	Yes

instead used Bayesian model averaging (BMA) to estimate the predicted parameters from multiple samples:

$$\hat{\xi}_{i,t} = \frac{\sum_{j=1}^{k} \hat{\xi}_{i,t,j} q(\theta_j)}{\sum_{j=1}^{k} q(\theta_j)} \tag{32}$$

$$\hat{\lambda}_{i,t} = \frac{\sum_{j=1}^{k} \hat{\lambda}_{i,t,j} q(\theta_j)}{\sum_{j=1}^{k} q(\theta_j)} \tag{33}$$

$$\hat{\delta}_{i,t} = \frac{\sum_{j=1}^{k} \hat{\delta}_{i,t,j} q(\theta_j)}{\sum_{j=1}^{k} q(\theta_j)} \tag{34}$$

$$\hat{\gamma}_{i,t} = \frac{\sum_{j=1}^{k} \hat{\gamma}_{i,t,j} q(\theta_j)}{\sum_{j=1}^{k} q(\theta_j)}. \tag{35}$$

BMA was also tested for the Gaussian likelihoods, but the weighted average approach in Eqs. (30) and (31) seemed to yield slightly better results.

To evaluate the probabilistic models' ability to model the underlying uncertainty, the pinball loss function in Eqs. (10) and (11) was used as a measure to compare the precision of the different quantile levels. For the parametric MLE and VI methods, quantiles could be easily obtained from the predicted Johnson's SU and Gaussian distributions with analytical PDFs and CDFs.

In probabilistic forecasting, it is also desirable to have narrow prediction intervals. As a measure of interval sharpness, the prediction interval average width (PIAW) calculates the average interval width for different confidence intervals:

$$PIAW_\alpha = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{I}_{i,0.5+\frac{\alpha}{2}} - \hat{I}_{i,0.5-\frac{\alpha}{2}} \right), \tag{36}$$

where $\hat{I}_{0.5+\frac{\alpha}{2}}$ and $\hat{I}_{0.5-\frac{\alpha}{2}}$ are the predicted lower and upper bounds for the α interval. The PIAW is calculated for a range of different intervals, $\alpha \in (0,1)$, where smaller values for PIAW are desirable, indicating narrower prediction intervals.

The PIAW measure in Eq. (36) does not evaluate the reliability of the probabilistic forecasts. The prediction interval coverage percentage (PICP) indicates the ability of the models to cover the targets under certain confidence intervals, α:

$$PICP_\alpha = \frac{1}{N} \sum_{i=1}^{N} c_i, \quad c_i = \begin{cases} 1, y_i \in [\hat{I}_{0.5-\frac{\alpha}{2}}, \hat{I}_{0.5+\frac{\alpha}{2}}] \\ 0, y_i \notin [\hat{I}_{0.5-\frac{\alpha}{2}}, \hat{I}_{0.5+\frac{\alpha}{2}}], \end{cases} \tag{37}$$

where y_i are the true values. Since the PICP computes the proportion of true values that fall within a certain confidence interval, i.e. the calibration, an ideal model will have $PICP_\alpha$ as close to α as possible. To obtain a single measure for the overall calibration, the average coverage error (ACE) can also be computed:

$$ACE = \frac{1}{|\mathbb{A}|} \sum_{\alpha \in \mathbb{A}} |\alpha - PICP_\alpha|, \tag{38}$$

where $\mathbb{A}$ is a set containing the different confidence intervals, α, that are considered, e.g. $\mathbb{A} = [0.25, 0.5, 0.75]$. Since ACE computes the average absolute difference between PICP and the desired confidence interval, smaller values for ACE is better.

The final probabilistic evaluation metric used was the continuous ranked probability score (CRPS). CRPS quantifies the difference between the CDF for the observed and predicted data:

$$CRPS = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} \left(\hat{F}_i(x) - \mathbf{1}_{x \geq y_i} \right)^2 dx, \tag{39}$$

where $\hat{F}$ is the CDF for the predicted distribution and $\mathbf{1}_{x \geq y_i}$ is a cumulative-probability step function, which jumps from zero to one at the point where x equals the observation y_i. The CRPS is negatively oriented, meaning that smaller values are better, and has a close relation to the MAE, since for point predictions, the CRPS will equal the MAE.

4.4. Results and discussion

All results are the computed mean from five independent training iterations, with the standard deviation values in Tables 2 and 3 being the standard deviation for the five different seeds. Similarly, shaded regions in Fig. 6 are also ±1 standard deviation from the mean.

4.4.1. Point prediction performance

First, point prediction performance of the different models was evaluated using the MAE and MSE metrics, with the results given in Table 2. Considering the different neural architectures, the models with LSTMs as temporal update functions in Fig. 4, generally achieved the best performance in terms of both MAE and MSE. Models with Transformer-based temporal functions achieved similar performance to the LSTM-based models, while the architectures using only a simple Linear transform yielded higher point prediction errors. Such findings

Table 2
Point prediction errors in kW for the different models.

Model	MSE ± std			MAE ± std		
	LSTM	Transformer	Linear	LSTM	Transformer	Linear
Deterministic	33,048 ± 199	34,211 ± 450	48,355 ± 2410	114.1 ± 0.8	118.9 ± 2.3	148.2 ± 5.2
QR	34,741 ± 340	34,188 ± 293	**40,459 ± 840**	112.2 ± 0.6	**111.4 ± 1.0**	**122.1 ± 2.3**
MLE-Gauss-Hetero	34,316 ± 1,951	40,878 ± 9184	100,129 ± 19,084	116.5 ± 4 3.9	134.5 ± 20.3	209.8 ± 23.0
MLE-Gauss-Homo	32,456 ± 241	34,066 ± 435	67,861 ± 16,811	113.0 ± 0.8	119.4 ± 2.4	177.8 ± 26.0
MLE-Johns-Hetero	35,133 ± 552	36,368 ± 638	41,532 ± 800	111.6 ± 1.3	114.5 ± 0.7	126.3 ± 1.6
VI-Gauss-Hetero	31,737 ± 182	33,337 ± 228	56,545 ± 1447	110.3 ± 0.5	113.5 ± 0.4	162.9 ± 2.2
VI-Gauss-Homo	**31,533 ± 86**	**32,790 ± 31**	51,401 ± 1751	111.1 ± 0.3	114.1 ± 0.3	155.2 ± 3.3
VI-Johns-Hetero	34,329 ± 375	37,531 ± 668	47,956 ± 1963	**107.8 ± 0.8**	113.7 ± 1.0	142.5 ± 3.9

Table 3
Probabilistic prediction evaluation for the different models.

Model	ACE ± std			Pinball ± std			CRPS ± std		
	LSTM	Transformer	Linear	LSTM	Transformer	Linear	LSTM	Transformer	Linear
QR	0.046 ± 0.011	0.038 ± 0.005	0.089 ± 0.007	6288 ± 106	6566 ± 342	6809 ± 111	75.2 ± 0.4	75.8 ± 3.0	**69.0 ± 0.8**
MLE-Gauss-Hetero	0.054 ± 0.009	0.088 ± 0.046	0.140 ± 0.029	16,724 ± 253	20,840 ± 4.670	36,436 ± 899	75.0 ± 2.8	92.9 ± 20.8	163.1 ± 8.5
MLE-Gauss-Homo	0.210 ± 0.003	0.209 ± 0.009	0.207 ± 0.053	23,241 ± 74	24,301 ± 1,075	33,568 ± 1,462	89.8 ± 0.2	94.7 ± 3.6	143.9 ± 8.6
MLE-Johns-Hetero	0.058 ± 0.016	0.065 ± 0.014	**0.033 ± 0.001**	6276 ± 308	6325 ± 605	6416 ± 156	76.8 ± 0.9	76.3 ± 2.8	83.7 ± 2.7
VI-Gauss-Hetero	0.085 ± 0.004	0.089 ± 0.003	0.149 ± 0.002	16,345 ± 73	16,834 ± 75	27,232 ± 220	70.9 ± 0.3	73.2 ± 0.3	121.8 ± 1.1
VI-Gauss-Homo	0.156 ± 0.001	0.155 ± 0.001	0.160 ± 0.002	20,020 ± 30	20,363 ± 27	26,350 ± 526	78.6 ± 0.2	80.7 ± 0.1	114.6 ± 2.8
VI-Johns-Hetero	**0.022 ± 0.003**	**0.018 ± 0.007**	0.157 ± 0.008	**5313 ± 355**	**5629 ± 111**	**5353 ± 201**	**69.0 ± 0.2**	**70.9 ± 1.4**	73.4 ± 2.3

were unsurprising, as the LSTM and Transformer learners are more complex and specifically designed to extract temporal features, as well as having GCNs to leverage spatial correlations with the GATv2 updates. There are a few reasons why the LSTM networks might have achieved superior results compared to the Transformer updates. First, the sequence and prediction lengths of 48 historical and 6 future time steps, respectively, are not very long. The advantage of Transformers for extracting long-term dependencies, where the LSTM network can tend to forget information for long sequences, might not be as relevant for the particular application of short-term wind power forecasting. Furthermore, the MHA in the Transformer is insensitive to local context and considers time steps independently when querying. This might in part explain why the LSTM performed better since the autoregressive architecture considers points sequentially, being more sensitive to local context. Finally, Transformers are known to require considerably longer training and large datasets, which might also explain the inferior performance compared to LSTMs for the fairly limited dataset size used in this study.

Now, considering the different probabilistic models, given by the different rows in Table 2, it was found that the VI models with Gaussian likelihoods, ‘VI-Gauss-Hetero’ and ‘VI-Gauss-Homo’, achieved superior results for the LSTM- and Transformer-based models in terms of MSE. However, looking at the MAEs, the ‘VI-Johns-Hetero-LSTM’ model outperformed all other models. Generally, it was found that the relative performance of the models with Johnson’s SU likelihoods, i.e. ‘MLE-Johns-Hetero’ and ‘VI-Johns-Hetero’ in Table 2, was much better in terms of MAEs, where these models showed some of the best performance, compared to some of the worst in terms of MSEs. Higher relative MSEs than MAEs indicate that the QR models and those with Johnson’s SU likelihoods had a few point predictions that were far away from the observed values, since MSE penalise larger errors more heavily, while on average most predictions were close to the target. The more flexible distributions produced using Johnson’s SU likelihood or QR might enable the models to have sharper probability distributions peaking closer to the exact values, while still covering the less likely values with a small probability. These less likely observations might be further away from the point predictions and therefore result in higher MSEs for these models, even though they might be within the larger confidence intervals which are not considered when evaluating point prediction accuracy. Overall, it was found that models with LSTM sequence learners generally outperformed Transformer-based models. Furthermore, it was found that probabilistic models capable of producing more flexible distributions, i.e. QR and models with Johnson’s SU distribution, had higher relative MSEs than MAEs. This meant that these models had some point predictions that were far away from the observed values, while on average having more predictions that were very close to the observed values.

4.4.2. Probabilistic prediction performance

For assessing the models’ probabilistic forecasting performance, the ACE, pinball loss and CRPS were computed and provided in Table 3. In general, the Transformer and LSTM-based architectures gave similar results, outperforming the linear models for most settings. The linear models mostly showed inferior results but had a few surprising instances of well-calibrated and precise probabilistic predictions, such as for the QR-Linear model. The VI-Johns-Hetero models performed remarkably well, with high calibration, as seen by the small ACE values of 0.022 and 0.018 in Table 3, low pinball losses which indicate high precision and narrow intervals, along with accurate overall CDFs seen by low CRPS values.

Gaussian likelihoods showed good point-prediction accuracies, along with the heteroscedastic Gaussians also having fairly good CRPS values. However, Gaussian likelihood models were not as good at producing accurate probability distributions, as seen by the higher ACE and pinball loss values, which indicate poor calibration. CRPS is a holistic metric that accounts for both distribution calibration and prediction accuracy. The acceptable CRPS values for Gauss-Hetero models in Table 3, might therefore in large part be due to good point prediction accuracies for these models, while the pinball metrics indicate that the models were not able to generate very precise overall distributions. Similarly, for the models with Johnson’s SU likelihoods, the relative performance of these models in terms of CRPS, was not as superior as for the ACE and pinball metrics, which might be due to these models having relatively higher MSEs.

Considering the VI- and MLE-Gauss-Hetero models, they achieved decent results in terms of ACE, with values ranging from 0.054–0.089. ACE is an absolute accuracy metric that does not consider the precision of predicted intervals, only calibration, while the pinball loss uses asymmetric weighting and captures the distance of observed values away from the predicted quantiles. In terms of pinball loss, the VI- and MLE-Gauss-Hetero models showed significantly worse performance, with values in the range of 16,000–20,000 kW. QR and Johnson’s SU likelihood models on the other hand, yielded pinball loss values in the range 5300–6600 kW. Low precision for the Gaussian likelihoods, as seen through the poor pinball loss values, could be because even though

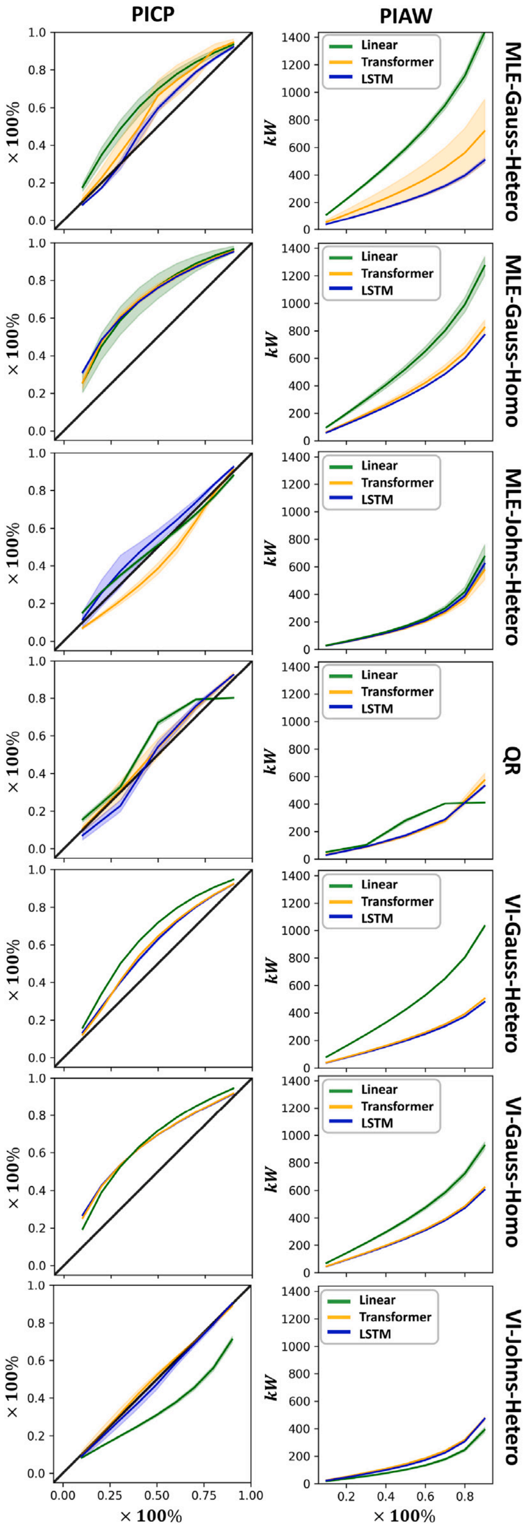

Fig. 6. Left and right columns show the PICP and PIAW values for the different models, respectively, where straight lines in the left-hand column show ideal calibration. Each row corresponds to a different method for uncertainty modelling, indicated by the text to the right of the PIAW plots. Shaded regions represent the ±1 standard deviation for five different training seeds.

such simple symmetric distributions might overall capture the correct proportion of observations within the respective intervals, they might have too wide intervals or quantile levels that are placed too far in either direction. For instance, considering the 50% confidence interval, these models might correctly have predicted the 25% quantile level, but have the 75% quantile level too far away due to the constraint of a symmetric Gaussian. In this scenario, nearly 50% of the data could be between the 25 and 50% quantiles, with hardly any points within the 50%–75% interval. Such low precision would not be captured by the ACE metric. The pinball loss on the other hand, considers quantiles independently and uses the observed distances of points away from the predicted quantiles, which would give a much higher loss for the too-high 75% quantile. With reference to the PIAW plots in Fig. 6, Gaussian likelihoods generally had wider intervals than the QR- and Johns-based models. From the PICP values in Fig. 6, it was found how Gaussian likelihoods tended to have too conservative predictions that overestimate the confidence intervals (i.e. curves lie over the straight line representing ideal coverage).

For the PICP values from the QR and Johns-Hetero models in Fig. 6, it was found that the models did not systematically overestimate intervals, but generally had very accurate interval predictions. Furthermore, considering the PICP values, the VI-Johns-Hetero models achieved remarkably well-calibrated predictions, which aligned closely with the superior results discussed for Table 3, while also having slightly smaller PIAWs meaning higher interval sharpness. Such findings were interesting, as the results in Table 3 indicate that the simple Gaussian distributions were not complex enough to model the wind power distributions, compared to QR and Johns-Hetero models which yielded much better results.

Even though the VI-Johns-Hetero models achieved superior results across all probabilistic evaluation metrics in Table 3, their relative improvements compared to other models were lower for pinball loss than ACE values. Since Johnson's SU distribution is parametric, quantile levels are not determined independently of each other. The slightly worse performance in terms of pinball loss might indicate that Johnson's SU distribution is also not perfect for modelling wind power and could benefit from potentially slightly narrower intervals around certain regions of highly concentrated observations. However, it was difficult to conclude on such an observation, since the VI-Johns models achieved even lower pinball losses than the QR-based models, which should be able to place quantile levels precisely at the correct locations, while also having very well-calibrated predictions according to the PICP plots in Fig. 6.

Homoscedastic likelihoods had poor calibration and precision, as seen by the higher ACE, pinball and CRPS values in Table 3. This was unsurprising, as it was expected that the uncertainty associated with a prediction would be input-dependant. For instance, the uncertainty is intuitively higher for powers around 1000 kW than for around 0 kW, as the power outputs are proportional to the wind speed cubed. Furthermore, power values close to rating or zero are also expected to remain in these regions for the immediate short-term, potentially reducing the uncertainty. Comparing VI and MLE, VI-based models seemed to achieve slightly more favourable results, indicating that these models benefited from having uncertainty inherently incorporated into the models.

To summarise, it was found that parametric MLE and VI models with Johnson's SU likelihoods and QR models produced well calibrated and precise interval predictions. VI-Johns-Hetero models outperformed all other models in terms of every evaluation metric, indicating accurate CDFs, high precision and calibration. Both MLE and VI models with Johnson's SU likelihood had very accurate PICPs, with less erratic behaviour across intervals compared to the more flexible QR-models, proving the suitability of Johnson's SU distribution in modelling uncertainty for wind forecasting. Gaussian likelihood models showed decent calibration in terms of ACE values but were found to consistently overestimate prediction intervals and had low precision with higher pinball losses. From the results, the VI-Johns-Hetero models seemed the most suitable for probabilistic wind power forecasting, with better performance across all metrics, while QR models achieved slightly inferior results, but with the added benefit of being somewhat easier to implement and faster.

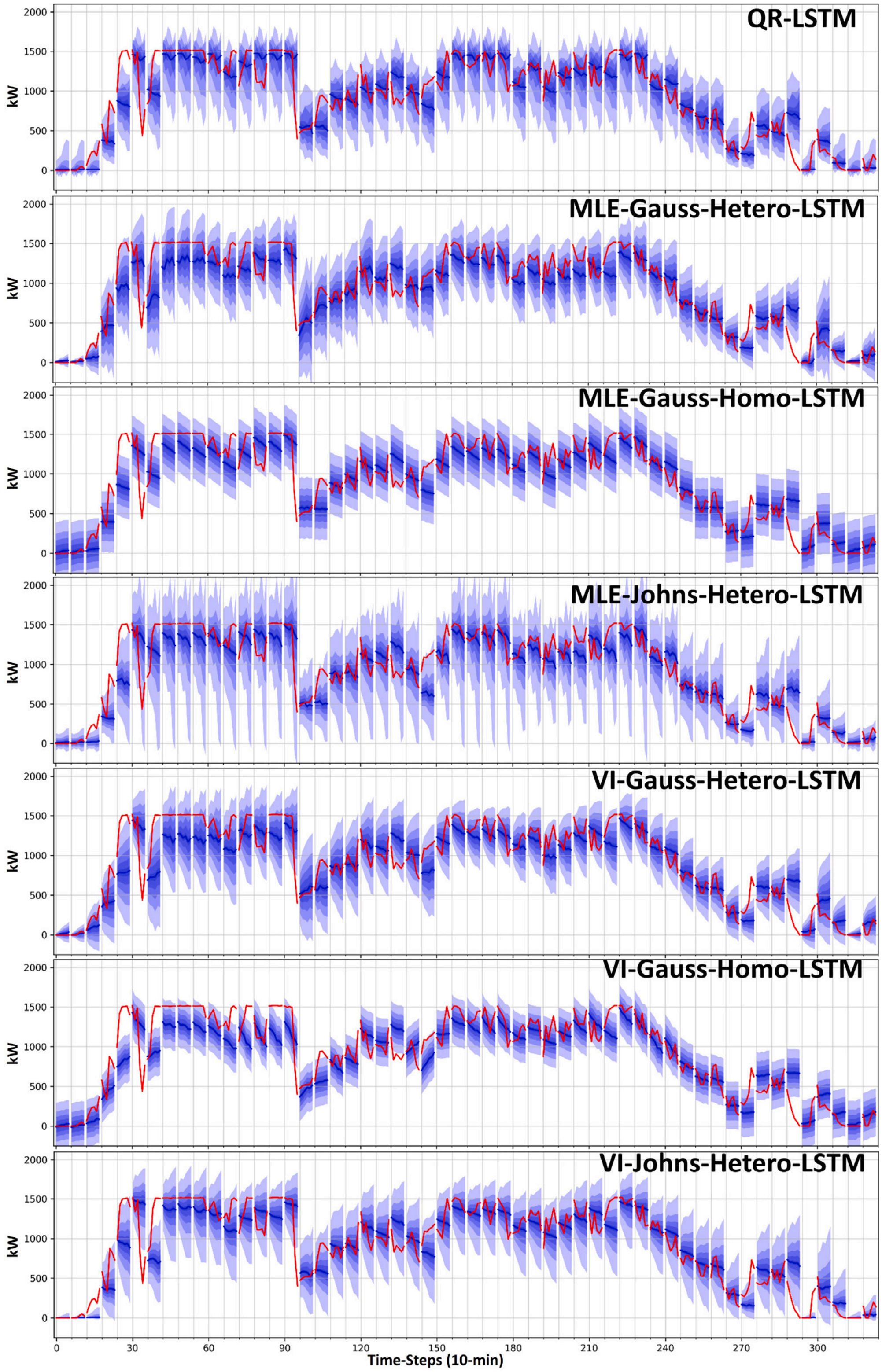

Fig. 7. Prediction examples for LSTM-based models for different methods of uncertainty modelling. Red are the observed values, and dark blue lines are the point predictions (i.e. 50% quantile levels). Shaded regions indicate the predicted 10, 30, 50, 70 and 90% confidence intervals.

4.4.3. Analysis of forecasting examples

Some prediction examples for all LSTM-based models are provided in Fig. 7. In Fig. 7, the red lines are the observed power values for a particular turbine in the farm, while the dark blue lines are point predictions, with shaded regions representing the forecasted 10, 30, 50, 70 and 90% confidence intervals. Values in Fig. 7 are discontinuous at every sixth-time step to distinguish the individual hour-ahead predictions. It can be observed from Fig. 7 how all models with heteroscedastic likelihoods showed progressively larger confidence intervals for predictions further into the future, i.e. larger uncertainty for the sixth than for the first step ahead predictions. This makes sense, as the models are expected to be less certain about predictions made further into the future. Considering the QR and Johns-Hetero models, it was seen how the models were able to generate skewed distributions, e.g. for regions with very small power production at the start or end of the plots in Fig. 7, such as for the predictions starting at time step 270.

From the prediction examples for the QR-LSTM model in Fig. 7, intervals seemed more chaotic, with overlapping quantiles and less discernible or regular structure across intervals. Such characteristics are common for QR models since the quantiles are predicted independently and are not restricted to adhere to any particular shape. The more erratic quantiles can make the predictions more challenging to interpret for downstream users, compared to predictions with Gaussian or Johnson's SU likelihoods that are forced to produce more consistent quantile levels. Considering commercial applications, such differences could be important, as improved interpretability for downstream users might be critical for the successful integration of advanced forecasting systems. On the other hand, VI models were generally more challenging to train, requiring investigation into different KL annealing schedules and good initialisation of Johnson's SU likelihood-based models. Such challenges could mean that QR-based models might be preferred for some commercial applications, despite slightly inferior performance.

With reference to the MLE-Johns-Hetero-LSTM predictions in Fig. 7, the 90% confidence intervals for the fifth and sixth step-ahead predictions were very wide. This characteristic was commonly seen when training models with Johnson's SU likelihood, due to the ability to control tail weights and skew. The Johnson's SU distribution also posed some challenges in training, due to vanishing or exploding gradients for very wide or sharp distributions. Such challenges in training might be one of the reasons for the very wide confidence intervals for the MLE-Johns-Hetero-LSTM model and were also the motivation behind initialising the skew and shape parameters to sensible values at the start of training, as discussed in Section 4.2.

To summarise, the prediction examples confirmed some of the previous conclusions, namely that the flexibility of QR and Johnson's SU distribution improved forecasts by being able to model skew and heavy tails. However, for intermediate regions, results were similar for models with Johnson's SU and Gaussian distributions. Finally, from the prediction results in Fig. 7, it was evident how parametric methods are easier to interpret, due to consistent interval levels, with the QR-LSTM model producing more erratic and inconsistent quantile levels.

4.4.4. *Power- and horizon-specific analysis*

As an additional analysis, we investigate the results for the different 1–6 step-ahead predictions (10–60 min), as well as for specific observed wind speed bins. This was done as an additional analysis to further strengthen our conclusions on the relative characteristics of the different probabilistic methods. Fig. 8 shows the computed evaluation metrics in Tables 2 and 3 for the different prediction steps and observed wind speeds for LSTM-based models. From the left-hand column in Fig. 8, it was found that MSEs and MAEs increased for predictions further into the future, which was expected. Pinball losses also increased for predictions further into the future, which indicates that the predicted quantile levels had lower precision. Recall that for the prediction examples in Fig. 7, interval widths generally increased further into the future. When quantile levels are further apart, it is challenging for the models to have very small pinball losses since the metric takes into account the distance of observations away from the predicted quantiles. In terms of calibration, most models did not seem to perform worse for predictions further into the future, as seen by the non-increasing ACE values in the bottom-left plot of Fig. 8. This was interesting as it meant that, since interval predictions generally had similar reliability for the later prediction steps, the increase in pinball loss might largely be a result of wider intervals, rather than the models performing much worse for the later prediction steps.

From the right-hand column in Fig. 8, it was found that MSEs, MAEs and CRPS values increased for larger observed wind speeds. Again, this was expected, since power values are likely to vary more for higher wind speeds and absolute point prediction accuracy will therefore be lower. For very high wind speeds, greater than 15 m/s, MSE, MAE and CRPS values flattened out, due to powers being closer to rating and therefore being potentially easier to accurately predict. Evaluating the quantile prediction performance, it was found that Johnson's SU likelihood and QR models performed much better than Gaussian likelihoods for lower wind speeds, smaller than 7 m/s. Models with Gaussian likelihoods achieved very precise quantile predictions, as seen by small pinball losses, for intermediate wind speeds around

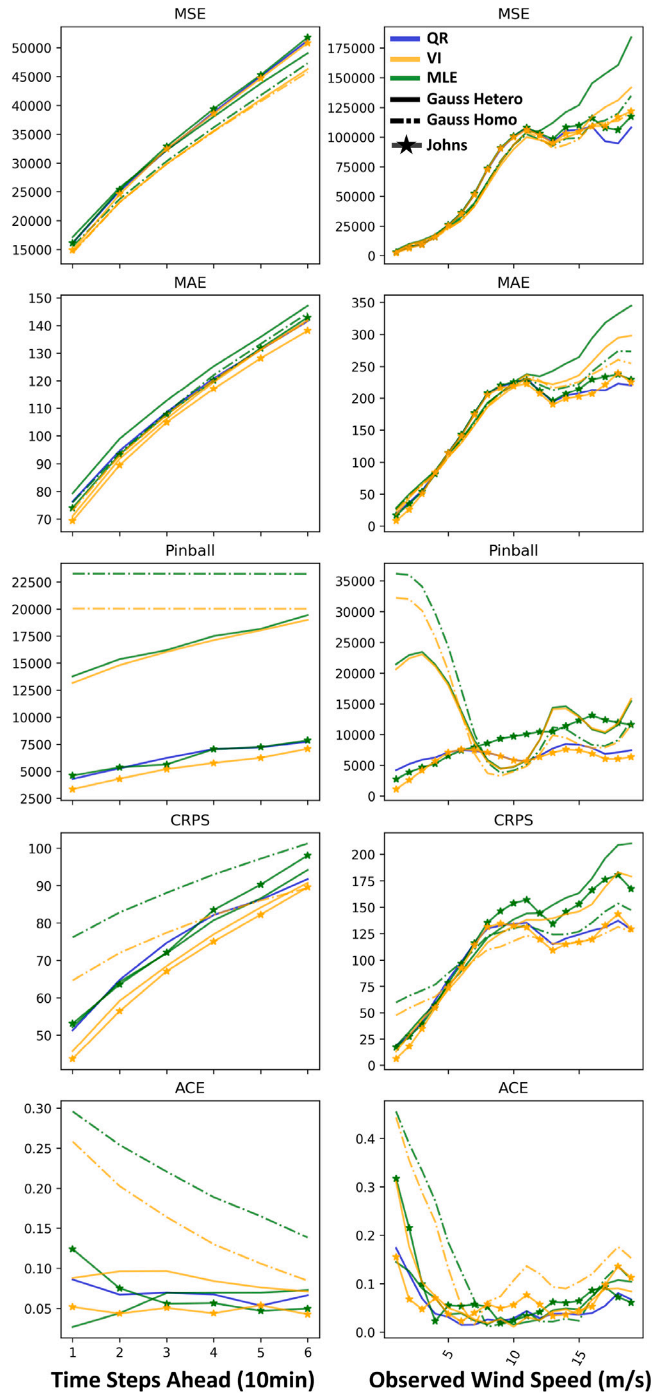

Fig. 8. Visualisation of point- and probabilistic prediction metrics for the different step-ahead predictions (left column) and for observed wind speeds (right column). Only results for the LSTM-based models are shown here for brevity.

8–11 m/s, before the VI-Johns and QR models again outperformed Gaussian models for very high wind speeds. This indicates that wind power might be well-modelled by symmetric Gaussian distributions for specific wind conditions, although skew and heavy tails are important for low and high wind speeds. In theory, Johnson's SU likelihood and QR models should be able to achieve competitive results even for the intermediate regions since they can also produce symmetric distributions similar to a Gaussian. However, most of the training data contained examples of lower wind speeds, which might in part, explain why the more complex distributions did not perform as well for the less represented wind conditions. More data would potentially be required for these complex models to produce a more versatile range of different distribution shapes. The limited training data might also explain some of the variability observed for MSE, MAE, pinball and CRPS values for higher wind speeds.

Finally, looking at all the computed metrics, for both different time steps and wind speeds in Fig. 8, it was generally found that the VI-Johns and QR models had less variability within the different subplots. This was desirable as it meant that these models were more consistent across prediction steps and for different wind conditions, making them potentially more reliable for downstream users.

To summarise, it was found that Gaussian likelihoods seemed well-equipped at modelling uncertainty for intermediate wind speeds, where wind power distributions are more symmetrical. QR and Johnson's SU distribution showed more consistent performance across wind speeds, which is a desirable property of a forecasting system to improve reliability. Considering the VI-Johns-LSTM model, it showed very small variability in calibration across different step-ahead predictions, indicating that the model was able to produce accurate predictions across the entire forecast horizon.

5. Conclusions

This study has researched three different methods for uncertainty modelling in wind power forecasting using DL, namely VI, MLE and QR, with either Gaussian or Johnson's SU likelihoods for the parametric methods. Non-parametric models based on QR, which can model any underlying distribution, were found to produce fairly precise and well-calibrated forecasts that generally outperformed VI and MLE methods with simpler Gaussian likelihoods. The main novelty of this paper was the investigation into the use of Johnson's SU distribution for probabilistic spatio-temporal wind forecasting, a distribution which has not been extensively studied in the literature. Parametric MLE and VI methods with Johnson's SU distribution yielded highly calibrated and precise forecasts, compared to Gaussian likelihoods. Calibration and precision were evaluated through the pinball loss, ACE, PICP and CRPS metrics, while general conclusions were supported by investigating results for different wind conditions, step-ahead times and qualitative assessment of produced forecasts. VI models with Johnson's SU likelihoods were found to outperform all other methods, with better results across all the aforementioned evaluation metrics. Since Johnson's SU distribution has limited research being applied to wind modelling and is very flexible, unbounded and easy to integrate into modern DL frameworks, this paper should enable a new direction of research that might significantly advance the field of probabilistic wind forecasting. However, despite the superior performance of VI architectures with Johnson's SU distribution, there are a few trade-offs that should be considered. Since VI is a true probabilistic model that requires sampling to produce probabilistic forecasts and can have challenges with posterior collapse, QR-based models are easier to implement out of the box and are faster. Furthermore, for the parametric VI and MLE methods with Johnson's SU likelihoods, it was found that it was important to initialise the distribution prior to training. This could make it more difficult to effectively implement Johnson's SU distribution and to obtain models that achieve the best results. For future work, it would be particularly interesting to further investigate the performance of Johnson's SU distribution for modelling uncertainty in wind for additional datasets and applications. Furthermore, Johnson's SU distribution should be further compared against more complicated distributions than a Gaussian, such as Beta or Gamma distributions.

CRediT authorship contribution statement

Lars Ødegaard Bentsen: Project administration, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft. **Narada Dilp Warakagoda:** Supervision, Writing – review & editing. **Roy Stenbro:** Supervision, Writing – review & editing. **Paal Engelstad:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lars Odegaard Bentsen reports financial support was provided by Research Council of Norway.

Data availability

Links to dataset is provided and openly available at https://aistudio.baidu.com/aistudio/competition/detail/152/.

Acknowledgements

This work was financed by the research project ELOGOW (Electrification of Oil and Gas Installation by offshore Wind). ELOGOW (project nr: 308838) is funded by the Research Council of Norway under the PETROMAKS2 program with financial support of Equinor Energy AS, ConocoPhillips Skandinavia AS and Aibel AS.

References

Afrasiabi, M., Mohammadi, M., Rastegar, M., Afrasiabi, S., 2020. Advanced deep learning approach for probabilistic wind speed forecasting. IEEE Trans. Ind. Inform. 17 (1), 720–727.

Aslam, S., Herodotou, H., Mohsin, S.M., Javaid, N., Ashraf, N., Aslam, S., 2021. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. Renew. Sustain. Energy Rev. 144, 110992.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bazionis, I.K., Georgilakis, P.S., 2021. Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research. Electricity 2 (1), 13–47.

Bazionis, I.K., Karafotis, P.A., Georgilakis, P.S., 2022. A review of short-term wind power probabilistic forecasting and a taxonomy focused on input data. IET Renew. Power Gener. 16 (1), 77–91.

Billinton, R., Huang, D., 2008. Aleatory and epistemic uncertainty considerations in power system reliability evaluation. In: Proceedings of the 10th International Conference on Probabilistic Methods Applied to Power Systems. IEEE, pp. 1–8.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. J. Am. Statist. Assoc. 112 (518), 859–877.

Bludszuweit, H., Domínguez-Navarro, J.A., Llombart, A., 2008. Statistical analysis of wind power forecast error. IEEE Trans. Power Syst. 23 (3), 983–991.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S., 2015. Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349.

Brody, S., Alon, U., Yahav, E., 2021. How attentive are graph attention networks? arXiv preprint arXiv:2105.14491.

Cadenas, E., Rivera, W., 2010. Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model. Renew. Energy 35 (12), 2732–2738.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

GWEC, 2022. Global wind report. URL: https://gwec.net/global-wind-report-2022/.

He, Y., Li, H., 2018. Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. Energy Convers. Manage. 164, 374–384.

He, Y., Zheng, Y., 2018. Short-term power load probability density forecasting based on Yeo–Johnson transformation quantile regression and Gaussian kernel function. Energy 154, 143–156.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36 (1/2), 149–176.

Kavasseri, R.G., Seetharaman, K., 2009. Day-ahead wind speed forecasting using f-ARIMA models. Renew. Energy 34 (5), 1388–1393.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, Vol. 30.

Khodayar, M., Wang, J., 2018. Spatio-temporal graph deep neural network for short-term wind speed forecasting. IEEE Trans. Sustain. Energy 10 (2), 670–681.

Koenker, R., Hallock, K.F., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156.

Li, Y., Wang, R., Li, Y., Zhang, M., Long, C., 2023. Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach. Appl. Energy 329, 120291.

Li, Y., Wang, R., Yang, Z., 2021. Optimal scheduling of isolated microgrids using automated reinforcement learning-based multi-period forecasting. IEEE Trans. Sustain. Energy 13 (1), 159–169.

Li, H., Zhang, Z., Zhang, B., 2020. Johnson system for short-term wind power forecast error modeling. In: 2020 IEEE 14th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG), Vol. 1. IEEE, pp. 377–381.

Liao, W., Bak-Jensen, B., Pillai, J.R., Wang, Y., Wang, Y., 2021. A review of graph neural networks and their applications in power systems. J. Mod. Power Syst. Clean Energy 10 (2), 345–360.

Liu, Y., Qin, H., Zhang, Z., Pei, S., Jiang, Z., Feng, Z., Zhou, J., 2020. Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model. Appl. Energy 260, 114259.

Liu, Y., Qin, H., Zhang, Z., Pei, S., Wang, C., Yu, X., Jiang, Z., Zhou, J., 2019. Ensemble spatiotemporal forecasting of solar irradiation using variational Bayesian convolutional gate recurrent unit network. Appl. Energy 253, 113596.

Liu, H., Tian, H.-q., Pan, D.-f., Li, Y.-f., 2013. Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. Appl. Energy 107, 191–208.

Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Pobočíková, I., Sedliačková, Z., Michalková, M., 2017. Application of four probability distributions for wind speed modeling. Procedia Eng. 192, 713–718.

Qu, K., Si, G., Shan, Z., Kong, X., Yang, X., 2022. Short-term forecasting for multiple wind farms based on transformer model. Energy Rep. 8, 483–490.

Quan, H., Khosravi, A., Yang, D., Srinivasan, D., 2019. A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids. IEEE Trans. Neural Netw. Learn. Syst. 31 (11), 4582–4599.

Ren, Y., Suganthan, P., 2014. Empirical mode decomposition-k nearest neighbor models for wind speed forecasting. J. Power Energy Eng. 2 (04), 176–185.

Ritter, H., Karaletsos, T., 2022. Tyxe: Pyro-based Bayesian neural nets for pytorch. In: Proceedings of Machine Learning and Systems.

Rogers, T., Gardner, P., Dervilis, N., Worden, K., Maguire, A., Papatheou, E., Cross, E., 2020. Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression. Renew. Energy 148, 1124–1136.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecast. 36 (3), 1181–1191.

Santamaría-Bonfil, G., Reyes-Ballesteros, A., Gershenson, C., 2016. Wind speed forecasting for wind farms: A method based on support vector regression. Renew. Energy 85, 790–809.

Sfetsos, A., 2002. A novel approach for the forecasting of mean hourly wind speed time series. Renew. Energy 27 (2), 163–174.

Shang, Z., Wen, Q., Chen, Y., Zhou, B., Xu, M., 2022. Wind speed forecasting using attention-based causal convolutional network and wind energy conversion. Energies 15 (8), 2881.

Sloughter, J.M., Gneiting, T., Raftery, A.E., 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. J. Am. Statist. Assoc. 105 (489), 25–35.

Valdenegro-Toro, M., Mori, D.S., 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1508–1516.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, Vol. 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

Wang, L., He, Y., Li, L., Liu, X., Zhao, Y., 2022a. A novel approach to ultra-short-term multi-step wind power predictions based on encoder–decoder architecture in natural language processing. J. Clean. Prod. 354, 131723.

Wang, Y., Hu, Q., Meng, D., Zhu, P., 2017. Deterministic and probabilistic wind power forecasting using a variational Bayesian-based adaptive robust multi-kernel regression model. Appl. Energy 208, 1097–1112.

Wang, Y., Xu, H., Zou, R., Zhang, L., Zhang, F., 2022b. A deep asymmetric Laplace neural network for deterministic and probabilistic wind power forecasting. Renew. Energy 196, 497–517.

Wang, Y., Zou, R., Liu, F., Zhang, L., Liu, Q., 2021. A review of wind speed and wind power forecasting with deep neural networks. Appl. Energy 304, 117766.

Zeng, A., Chen, M., Zhang, L., Xu, Q., 2023. Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37. pp. 11121–11128.

Zhang, L., Cheng, H., Zhang, S., Zeng, P., Yao, L., 2016. Probabilistic power flow calculation using the Johnson system and Sobol's quasi-random numbers. IET Gener. Transm. Distrib. 10 (12), 3050–3059.

Zhang, Y., Wang, J., Wang, X., 2014. Review on probabilistic forecasting of wind power generation. Renew. Sustain. Energy Rev. 32, 255–270.

Zhou, J., Lu, X., Xiao, Y., Su, J., Lyu, J., Ma, Y., Dou, D., 2022. Sdwpf: A dataset for spatial dynamic wind power forecasting challenge at kdd cup 2022. arXiv preprint arXiv:2208.04360.

Zhu, Q., Chen, J., Shi, D., Zhu, L., Bai, X., Duan, X., Liu, Y., 2019. Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. IEEE Trans. Sustain. Energy 11 (1), 509–523.