Getting Uncertainty Right: Local Calibration for Wind Power Forecasting

Manush Sanchela* , Vrishin Rajpurohit*, and Pritam Anand*

Dhirubhai Ambani University, Gandhinagar, 382007 India
Email: manushsanchela796@gmail.com, vrishin_rajpurohit@dau.ac.in, pritam_anand@dau.ac.in

***Abstract*—Wind energy forecasting has increasingly shifted from point to probabilistic approaches to support risk-aware decision-making. However, most existing methods evaluate models using global calibration metrics, which fail to capture reliability in real-time operations. In this paper, we emphasize the importance of local calibration for trustworthy decision support in wind power generation. We propose a set of local calibration metrics to assess probabilistic forecasts at a finer temporal scale. Furthermore, we incorporate the Adaptive Conformal Inference (ACI) framework as a model-agnostic, post-hoc approach to improve calibration. Extensive experiments on real-world wind power datasets using deep probabilistic forecasting models show that ACI consistently enhances local calibration performance, with improvements of up to 30% in the proposed metrics. These results highlight the significance of local calibration and demonstrate the effectiveness of ACI in improving the reliability of probabilistic forecasts for real-time, risk-aware decision-making.**

I. Introduction

Wind power has emerged as a significant contributor to the global energy market in recent years, with reports indicating that it accounted for approximately 12% of global electricity generation as of early 2025 [1]. Wind power generation is different from traditional fossil fuel–based sources, as it is inherently controlled by variable atmospheric conditions. As a result, accurate forecasting plays a critical role in the domain of wind power generation. For grid operators, precise power forecasts are essential, as many day-to-day operations such as planning reserve capacity and optimizing supply to meet demand rely on reliable predictions [2].

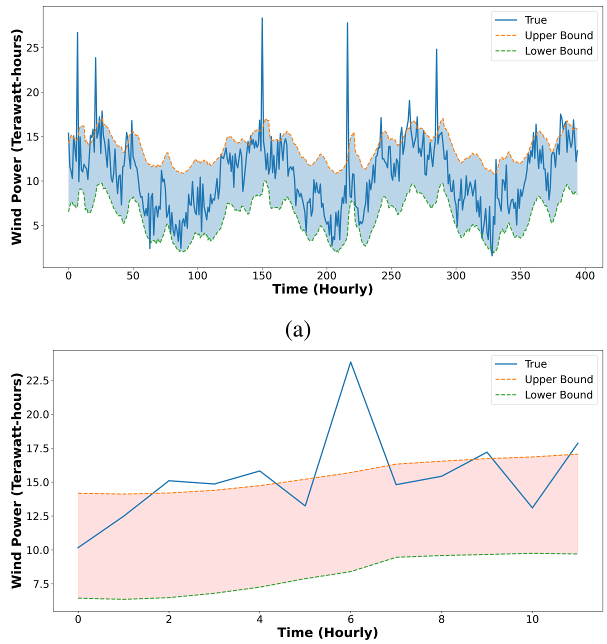
Historically, wind power forecasting has relied on point forecasting approaches, which aim to predict a single deterministic value for future generation. Early methods primarily employed statistical models such as ARMA[3], ARIMA[4], and SARIMA[5] to capture time-series dependencies, offering limited representations of uncertainty through sequential modeling. However, over the past decade, deep learning techniques have gained significant attention in wind power forecasting due to their inherent ability to extract complex temporal patterns and effectively model highly nonlinear relationships within the data. In modern energy literature, the deep learning architectures such as Long Short-Term Memory (LSTM) [6], Gated Recurrent Units (GRU) [7] , and Temporal Convolutional Networks (TCN) [8] are go-to algorithms for learning complex relationships in atmospheric data .

However, as wind power generation is a non-stationary and highly volatile process [9], achieving accurate forecasts that can be directly translated into grid actions remains a significant challenge, even with sophisticated modern deep learning architectures. This has shifted the focus of wind energy research from point forecasting to probabilistic forecasting methods, which can quantify uncertainty in predictions and support risk-aware real-time decision-making. The wind power probabilistic forecasting literature has primarily evolved along two main directions. The first approach assumes an underlying probability distribution and estimates its parameters to derive the corresponding quantile functions for probabilistic forecasting. Recent models following this paradigm include Normalizing Flows[10], Mixture Density Networks [11], and DeepAR [12], which are designed for autoregressive wind power time series.The second research direction, which has gained increasing popularity, focuses on modelling auto-regressive dependencies in a distribution-free setting. These approaches directly estimate the desired quantile functions by minimising the pinball loss [13] using various deep learning architectures, such as RNNs, LSTMs, GRUs, and TCNs. Some of the popular and recent literature that uses the distribution-free quantile approach for obtaining the probabilistic forecasting of the wind power are QR-BiLSTM [14] and QR-TCN[15] which uses a combination of Different deep learning architecture for forecasting.

However, regardless of the approach, two key limitations persist in existing probabilistic forecasting models, restricting their effectiveness for real-time decision support in wind farm operations. First, existing wind probabilistic forecasting models focus on producing *globally well-calibrated* prediction intervals. In contrast, for daily grid operations and risk-aware decision-making, *locally calibrated prediction intervals* are more actionable and essential. This is because grid operators procure reserves over limited time horizons (e.g., 6 to 24 hours). The existing probabilistic forecasting models available in the wind power literature may claim to attain the target coverage level (such as 0.95) globally; however, they may be operationally unreliable if they tend to obtain poor coverage during a key 24-hour ramp event. Figure 1 shows an example on a toy wind dataset of how a modern probabilistic forecast-

* Equal contribution.

This project is funded by Smart Energy Learning Center, Dhirubhai Ambani University, Gandhinagar, under the CSR grant no. CSR-25/BSES/A6-PA/SELC.



(a)

(b)

Fig. 1: (a) Rolling forecasts of the LSTM-Quantile model over 400 future time steps. The model achieves a global PICP of **89.6%**, which suggests overall reliability. However, it fails to maintain the required level of local calibration, which is essential for risk-aware daily decision-making in grid operations. (b) Local performance of the LSTM-Quantile model over selected 24-hour windows. Only **66.7%** of the observations fall within the estimated prediction intervals in these windows, highlighting poor local calibration and reducing the practical usefulness of the forecasts.

ing deep learning model may look globally well-calibrated, yet fail to attain the desired coverage locally, which limits its ability to support risk-aware decisions. Second, in real-time grid operations, wind power data arrives in an online, streaming manner, requiring deep learning models to be periodically updated at fixed intervals. However, most wind power forecasting models are typically trained in an offline, batch-data setting, limiting their adaptability to evolving real-time conditions.

In view of the above research gap, we outline the key contribution of this paper as follows.

1) This paper highlights and emphasizes the importance of local calibration in wind probabilistic forecasting models for real-world, risk-aware decision-making. Through illustrative examples, we demonstrate how a well-calibrated probabilistic forecasting model can enable reliable and trustworthy day-to-day decisions in the wind power generation process.
2) We develop a set of local evaluation metrics that effectively quantify the quality of probabilistic wind power forecast at the local level and advocate their use in future studies for a robust assessment of probabilistic forecast performance from a decision-support perspective.
3) We conduct a series of experiments with deep probabilistic forecasting models that are periodically fine-tuned every six hours using streaming time-series wind power data to provide real-time decision support. More importantly, we show that the use of the Adaptive Conformal Inference (ACI) framework [16] in the deep probabilistic forecasting models leads to better local calibration ability in wind power forecasting. In particular, our numerical results conclude that the use of the ACI in wind power probabilistic forecasting leads to around 30% improvement in our developed local metrics that effectively quantify the local calibration ability of the probabilistic forecasting models.

The rest of the paper is organized as follows. Section II reviews global evaluation metrics, motivates the need for local calibration, and introduces localized metrics for real-time decision support. Section III presents numerical experiments demonstrating the effectiveness of Adaptive Conformal Inference (ACI) in improving local calibration under an online streaming setting.

II. Local calibration in wind probabilistic forecasting models for effective real-time decision support

In this section, we strongly emphasize the need for local calibration for real-time decision-making in the wind power industry. In the literature, probabilistic wind power forecasting models are primarily evaluated using global metrics that remain fixed over a given test set. However, real-time applications involve continuously streamed data and require strong local calibration over specific time windows to enable adaptive and timely decision-making.

In literature, the probabilistic forecasting models are commonly trained upon the time-series wind power data $\{x_1, x_2, \dots x_T\}$ and evaluated upon the test data $\{x_{T+1}, x_{T+2}, \dots x_{T+k}\}$ with the following metrics.

In the current literature, time-series forecasting models are commonly evaluated using the following metrics:

(a) **PICP (Prediction Interval Coverage Probability)**:- It measures the fraction of true observations that fall within the PI over the test horizon. It is defined as:

$$\textbf{PICP} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\left(x_{T+i} \in [\hat{L}(x_{T+i}), \hat{U}(x_{T+i})]\right) \quad (1)$$

where $[\hat{L}(x), \hat{U}(x)]$ is the model estimated PI. For a well-calibrated model, the PICP should be close to or greater than the nominal confidence level.

(b) **MPIW (Mean Prediction Interval Width)** :- MPIW quantifies the average width of the estimated PI and is given by:

$$\textbf{MPIW} = \frac{1}{k} \sum_{i=1}^{k} \left(\hat{U}(x_{T+i}) - \hat{L}(x_{T+i})\right) \quad (2)$$

Lower MPIW values correspond to sharper PIs.

(c) **CRPS (Continuous Ranked Probability Score)** :- CRPS evaluates the quality of the full predictive distribution. However, it is primarily relevant for models that explicitly estimate the conditional autoregressive density and subsequently derive the quantile function from it, such as DeepAR and mixture density network (MDN)-based models . Also, it rarely participates in the final decision function for risk-aware decision support.

A. Limitation of Global Metrics in Real-Time Risk-Aware Decision Making

In real-world electricity markets, forecasting errors translate directly into economic cost through time-varying imbalance prices. The total cost over a horizon is given by

$$C = \sum_{t=1}^{N} e_t \cdot \mathrm{AEP}_t \tag{3}$$

where $e_t = y_t - \hat{y}_t$ is the forecast error (actual generation minus scheduled position) and AEP_t (Ausgleichsenergiepreis) is the imbalance energy price (€/MWh) at time t, determined under the reBAP mechanism [17] based on real-time system imbalance and activated balancing reserve costs. Because AEP_t varies over time, errors during high-price periods have a disproportionately large impact on total cost.

Global metrics such as PICP and MPIW summarize performance over the entire dataset and do not capture time-specific behavior. A model may achieve a high global PICP yet exhibit poor coverage over a critical 24-hour interval. In day-ahead settings, poor calibration during a few key hours can lead to significant economic loss, even if the global PICP appears satisfactory. Local calibration evaluated over the 24-hour decision window provides time-specific reliability, making it more relevant for cost-sensitive, real-time decision-making.

B. Local Calibration Metrics for Local Evaluation

Motivated by the limitations of global calibration metrics, we propose a local calibration framework that enables real-time, risk-aware decision support assessment of model reliability over localized temporal regions.

Let $\alpha \in (0,1)$ denote the miscoverage rate, such that $1-\alpha$ represents the target confidence level of the prediction interval. To enable risk-aware decision-making, it is essential to capture the local calibration ability of the probabilistic forecasting model within a rolling window of size w. This leads to the formulation of the following evaluation metrics that quantify the true quality of probabilistic forecasts at a local level, thereby making them more relevant and actionable for day-to-day operational decision-making.

1) **Local Prediction Interval Coverage Probability (LPICP):** At the i^{th} time-stamp, it quantifies the local calibration ability of the model within a *centered* sliding window. The window is defined as

$$W_i = \{t \mid t = i - h_l, \ldots, i + h_r\}$$

where

$$h_l = \begin{cases} \frac{w}{2} - 1, & \text{if } w \text{ is even}, \\ \left\lfloor \frac{w}{2} \right\rfloor, & \text{if } w \text{ is odd}, \end{cases} \qquad h_r = \left\lfloor \frac{w}{2} \right\rfloor.$$

The local coverage at index i is given by

$$\mathbf{LPICP}_i = \frac{1}{w} \sum_{t \in W_i} \mathbf{1}\big[\hat{L}(x_t) \le x_t \le \hat{U}(x_t)\big], \tag{4}$$

which is defined only for indices where the full window lies within the data range.

The overall local coverage is computed as

$$\mathbf{LPICP} = \frac{1}{K} \sum_{i=T+1}^{T+k} \mathbf{LPICP}_i, \tag{5}$$

that is different from the commonly PICP detailed in (1). For taking the risk-aware decision with confidence, say $1-\alpha$, all of the observed $\mathbf{LPICP}_i$ should be greater than or equal to $1-\alpha$ for $i = T+1, \ldots T+k$, which seems to be a much stricter condition than obtaining the PICP greater than or equal to $1-\alpha$ in real-time experiments, particularly in the presence of highly volatile and chaotic components in time-series wind power data.

2) **Local Prediction Interval Coverage Error (LPICE):-** To quantify the local calibration error, we further define the LPICE using

$$\mathbf{LPICE} = \sqrt{\frac{1}{k} \sum_{i=T+1}^{T+k} \Big(\max\big((1-\alpha) - \mathbf{LPICP}_i,\, 0\big)\Big)^2}. \tag{6}$$

The LPICE quantifies the average local calibration error and should be as small as possible. Apart from the average, the variance of the local error should be expected to be as small as possible for a locally well-calibrated PI.

3) **Local Prediction Interval Accuracy (LPIA):-** It quantifies the fraction of windows the model satisfies the local calibration target. It is given by

$$\mathbf{LPIA} = \frac{1}{k} \sum_{i=T+1}^{T+k} \mathbf{1}\big[\mathbf{LPICP}_i \ge 1-\alpha\big] \tag{7}$$

It reflects the actual reliability and trustworthiness of the underlying decisions.

Choice of window size w:- The window size w defines the temporal scale of local calibration and should align with the operational decision horizon (e.g., 24–48 hours in wind grid operations). Reliable decision-making requires consistent calibration within this window. Larger w relaxes the calibration requirement, and in the extreme case $w = k$, it reduces to global PICP. In contrast, smaller w imposes stricter local calibration, making it more challenging due to the inherent volatility of wind power data.

C. Adaptive Conformal Inference (ACI) Formulation

Standard Conformal Regression (CR) calibrates prediction intervals using past residuals. Given predicted lower and upper quantile bounds $\hat{L}(x_t)$ and $\hat{U}(x_t)$, the non-conformity score is

$$s_t = \max\Big(\hat{L}(x_t) - x_t,\ x_t - \hat{U}(x_t)\Big). \tag{8}$$

The empirical quantile $q_t = \text{Quantile}_{1-\alpha}(\{s_1, \ldots, s_{t-1}\})$ yields the calibrated prediction interval

$$C_t = \left[\hat{L}(x_t) - q_t,\ \hat{U}(x_t) + q_t\right]. \tag{9}$$

However, CR assumes exchangeability, which is violated in time-series data due to temporal dependencies and distributional shift.

The Adaptive Conformal Inference (ACI) framework [16] addresses this by replacing the fixed miscoverage level α with an adaptive level α_t, updated online as

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - 1\{x_t \notin C_t\}), \tag{10}$$

where $\alpha \in (0, 1)$ is the target miscoverage level and γ controls the adaptation rate. When x_t falls outside C_t, α_t decreases, widening subsequent intervals; when covered, α_t increases, sharpening them. This enables the PI to adapt to distributional shifts without retraining the base model.

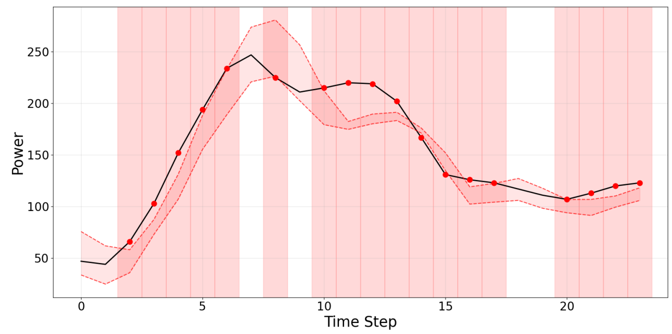
In our pipeline, ACI serves as a model-agnostic, post-hoc recalibration layer on top of the quantile-based deep learning models. Non-conformity scores are computed over a rolling buffer, ensuring intervals track evolving wind conditions. Fig. 2(a)–(b) illustrates the effect: without ACI, interval violations are frequent during shifts; with ACI, intervals adapt, improving short-term calibration.

III. Empirical Analysis of ACI-Driven Calibration

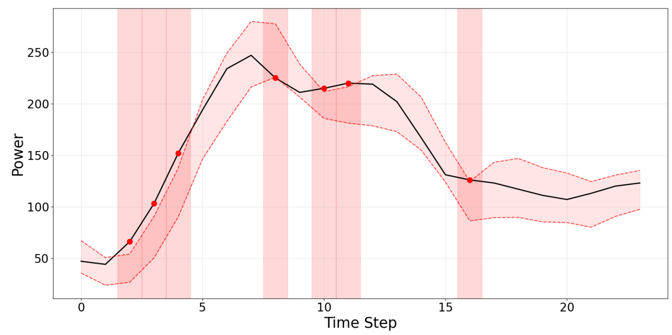
In this section, we present the empirical results and their analysis to demonstrate that the Adaptive Conformal Inference (ACI) framework improves our local calibration metrics in probabilistic wind power forecasting models to help us make risk aware decisions. We have targeted to obtain the interval for wind power with the target nominal coverage 0.90

Datasets:- The experiments are conducted on three publicly available wind power time series datasets, namely **50Hertz** [18], **Amprion** [18], and **TransnetBW** [18]. These datasets represent three major wind power transmission system operators in Germany. The original measurements are recorded at a 15-minute resolution. To align with standard grid dispatch practices, the data are downsampled to an hourly frequency by retaining every fourth observation. Each dataset is thereby standardized to 9,480 hourly observations, expressed in normalized wind power units.

Experimental setup:- To evaluate these datasets, we adopt a sliding window approach to simulate real-time forecasting scenarios. The initial 5,000 observations are utilized for the offline training of our models over 50 epochs. After preliminary tuning, we formulate the task as an autoregressive prediction problem with a sequence length of 48; that is, the model uses the previous 48 time steps to predict the subsequent target value (x_{t+1}). Following this initial phase, predictions are generated sequentially for the remaining observations (indices 5,001 to 9,450). To maintain an online setting and adapt to evolving patterns and temporal dependencies, the model is fine-tuned every 6 hours using newly arriving observations to update its parameters. We assume that grid operators make



(a)



(b)

Fig. 2: (a) Rolling forecasts of the QR-LSTM model on the Amprion dataset without adaptive calibration. Although the model produces reasonable overall trends, it exhibits frequent violations of the prediction intervals (highlighted in red), indicating poor adaptation to local distributional shifts and unreliable uncertainty estimates. (b) Forecasts enhanced with Adaptive Conformal Inference (ACI) using a rolling window of size 24. The prediction intervals dynamically adjust to changing local conditions, significantly reducing the number of out-of-interval observations. This demonstrates improved local calibration, leading to more reliable and risk-aware forecasts in the presence of temporal variability..

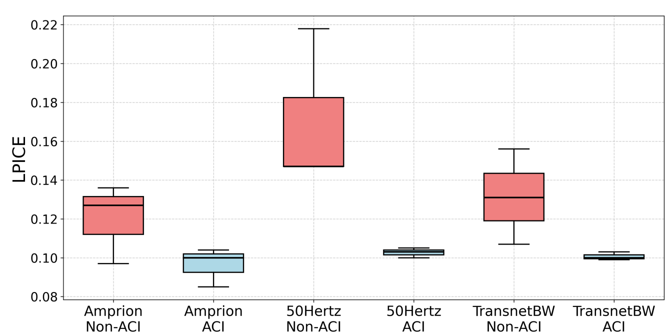


Fig. 3: Distribution of LPICE across datasets under Non-ACI and ACI settings. ACI consistently reduces both the median and variance of local calibration error, indicating improved reliability.

risk-aware decisions for procurement and dispatch at a 24-hour frequency. Accordingly, the local window size is set to 24 in our experiments.

Probabilistic Forecasting Method:- In this work, we employ quantile-based deep learning models to generate probabilistic forecasts under both ACI and non-ACI settings. These

TABLE I: Performance of the quantile based deep forecasting models in ACI and Non-ACI setting with target coverage ($1 - \alpha = 0.90$ and $w = 24h$)

Model	Amprion								50Hertz								TransnetBW							
	Non-ACI				ACI				Non-ACI				ACI				Non-ACI				ACI			
	PICP	LPICP	LPICE	LPIA	PICP	LPICP	LPICE	LPIA	PICP	LPICP	LPICE	LPIA	PICP	LPICP	LPICE	LPIA	PICP	LPICP	LPICE	LPIA	PICP	LPICP	LPICE	LPIA
LSTM	84.6	81.0	0.136	0.1990	**90.0**	**86.2**	**0.100**	**0.4422**	83.2	79.7	0.147	0.1734	**89.7**	**86.0**	**0.105**	**0.4534**	82.3	78.8	0.156	0.1144	**90.0**	**86.1**	**0.099**	**0.4617**
GRU	84.8	81.2	0.127	0.1990	**89.7**	**86.0**	**0.104**	**0.4411**	83.1	79.7	0.147	0.1672	**89.8**	**86.1**	**0.103**	**0.4508**	84.8	81.2	0.131	0.1838	**89.8**	**86.0**	**0.103**	**0.4247**
TCN	88.2	84.4	0.097	0.3013	**89.8**	**86.1**	**0.085**	**0.3899**	86.7	83.1	0.2179	0.2170	**89.7**	**86.0**	**0.100**	**0.4232**	86.8	83.2	0.107	0.2255	**89.9**	**86.1**	**0.100**	**0.4234**

models provide a distribution-free framework for uncertainty estimation and have been shown to perform consistently across a variety of datasets. To obtain probabilistic forecasts of wind power, we train a pair of deep learning models by minimizing the pinball loss at quantile levels $\tau = 0.05$ and $\tau = 0.950$, corresponding to the lower and upper bounds of the prediction interval, respectively.

Adaptive Conformal Inference (ACI) Configuration. After obtaining the quantile bounds functions through a quantile-based deep learning model, we apply the adaptive conformal procedure detailed in Section (II) (C) for obtaining the locally calibrated PI. To compute the non-conformity score, we maintain a rolling calibration buffer consisting of the most recent 200 data points. The update parameter is set to $\gamma = 0.001$, which adaptively adjusts the miscoverage level to preserve conformal guarantees under distributional shift.

Impact of ACI upon local calibration:- Table I compares quantile-based deep learning models, namely LSTM+Quantile, GRU+Quantile, and TCN+Quantile, trained under both ACI and Non-ACI settings. The models are evaluated using a local window size of $w = 24\mathrm{h}$ and a target calibration level of $1 - \alpha = 0.90$, based on the proposed evaluation metrics.

The numerical results presented in Table I clearly demonstrate that incorporating ACI into quantile-based deep forecasting models leads to improved local calibration performance, as effectively quantified by the proposed evaluation metrics.

Fig. 3 shows that ACI reduces both the median and variance of LPICE across all models and datasets, with an average reduction of **20–35%**, indicating improved local calibration accuracy. Additionally, ACI improves LPICP by **5–7%**, enhancing local coverage reliability. LPIA also increases significantly by **80–120%**, reflecting more reliable and better-calibrated prediction intervals. Overall, these results confirm the effectiveness of ACI in improving local calibration.

IV. Future Work

The ACI configuration is a post-hoc conformal procedure that is model-agnostic and independent of the underlying probabilistic forecasting method. However, in this work, we have limited our analysis to quantile-based deep learning models to demonstrate that the ACI framework leads to significant improvements in local calibration performance. As part of future work, we plan to extend this investigation to other popular deep forecasting models available in the literature, in order to further evaluate and validate the effectiveness of ACI in enhancing local calibration across diverse model architectures.

References

[1] World Wind Energy Association (WWEA), "Half year report 2025," 2025.

[2] S. Diop, P. S. Traore, and M. L. Ndiaye, "Wind power forecasting using machine learning algorithms," in *2021 9th International Renewable and Sustainable Energy Conference (IRSEC)*. IEEE, 2021, pp. 1–6.

[3] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1970.

[4] G. M. Jenkins, "Autoregressive–integrated moving average (arima) models," in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, 2014.

[5] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2008.

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

[7] K. Cho, B. Van Merriënboer, Ç. Gulçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1724–1734.

[8] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[9] L. Shi, S. Sun, L. Yao, Y. Ni, and M. Bazargan, "Effects of wind generation intermittency and volatility on power system transient stability," *IET Renewable Power Generation*, vol. 8, no. 5, pp. 509–521, 2014.

[10] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[11] C. Bishop, "Mixture density networks," Tech. Rep., January 1994.

[12] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International journal of forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

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

[14] L. Chen, M. Yang, X. Guan, and L. Ma, "Research on probabilistic wind power prediction based on qrcnn-bilstm model," in *2025 4th International Conference on Energy, Power and Electrical Technology (ICEPET)*. IEEE, 2025, pp. 468–472.

[15] J. Hu, Q. Luo, J. Tang, J. Heng, and Y. Deng, "Conformalized temporal convolutional quantile regression networks for wind power interval forecasting," *Energy*, vol. 248, p. 123497, 2022.

[16] I. Gibbs and E. Candes, "Adaptive conformal inference under distribution shift," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1660–1672, 2021.

[17] German TSOs (Grid Control Cooperation), "Model description report of the rebap calculation since 01.11.2023," German Transmission System Operators (TSOs), Tech. Rep., 2023.

[18] J. G. S. Simão and Netztransparenz, "Wind power generation data," 2021. [Online]. Available: https://www.kaggle.com/dsv/1825731