Probabilistic Electricity Demand Forecasting with Transformer-Guided State Space Model

Yang Cao
China Energy Dadu River Dagangshan Power Generation Co. LTD.
Ya'an, China
12046799@ceic.com

Zhenzhen Dang
China Energy Dadu River Dagangshan Power Generation Co. LTD.
Ya'an, China
12042110@ceic.com

Feng Wu
China Energy Dadu River Dagangshan Power Generation Co. LTD.
Ya'an, China
12102448@ceic.com

Xovee Xu*
University of Electronic Science and Technology of China
Chengdu, China
*Corresponding author: xovee@ieee.org

Fan Zhou
University of Electronic Science and Technology of China
Chengdu, China
fan.zhouestc.edu.cn

***Abstract*—Forecasting the electricity demand has many critical applications for power station generation plans, energy resource allocations, power transmissions, and related decision makings. Existing electricity demand forecasting methods are mainly based on deterministic time-series prediction techniques and recurrent neural networks, ignoring the intrinsic uncertainties of electricity demands and the long-range temporal dependencies between electricity volumes and various external features. In this work, we proposed a Transformer-guided probabilistic electricity demand forecasting framework (TPEDF) that learns both the global-local electricity demand dependencies and the complex correlations of external features entangled with the electricity demands. Our proposed framework predicts electricity demand distributions with a probabilistic state space model, which can effectively eliminate the forecasting errors and enhance the expressiveness of the model by taking the electricity uncertainty into account. Extensive experiments on real-world datasets demonstrate that TPEDF significantly outperforms existing models on electricity demand forecasting.**

Keywords—electricity demand forecasting, time-series forecasting, transformer, state space model, uncertainty

I. Introduction

Electricity Demand Forecasting (EDF) is a critical time-series prediction task and has attracted tremendous research attention in recent years [1]-[5]. The goal of the EDF is to predict the future (long-term and short-term) electricity demand, given the historical observations and various external features that may have an impact on the electricity demand in the studied area. With rising concerns about global warming, environmental degradation, and natural hazards, governments and companies are increasingly turning to the use of clean energies such as wind and hydroelectric generation. Given the fact that renewable power generation approaches are less stable and uncertain than those traditional methods, and the energy consumption around the world is rapidly increasing (e.g., the electricity consumption in India increases by 7% on average each year [6]), how to accurately predict the future electricity demands becomes an essential yet challenging task for power systems on electricity planning, management, allocation, and transmission [7].

Existing electricity demand forecasting methods can be categorized into two directions: traditional and deep learning-based methods [8]. Traditional EDF methods include feature engineering and stochastic time-series pattern analysis. These methods are based on expert-specified rules and hand-crafted features, and their predictions are also highly interpretable. For example, Braun *et al*. conducted a multiple regression analysis using various climate features such as temperature and humidity for electricity consumption estimation [9]. Another line of methods uses statistical machine learning techniques such as support vector machine (SVM) [10] and Autoregressive Integrated Moving Average (ARIMA) [11] to learn data patterns from electricity time-series. More recently, deep learning-based methods achieved great success in time-series forecasting due to their capabilities of learning expressive time-series representations [8], e.g., Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [12].

However, existing prediction methods for electricity demand forecasting face several limitations. First, traditional methods are constrained by the parameter-tuning process (which is hard to be determined) and consequently less generalized for different types of electricity demands. Second, they usually rely on deterministic learning methods that cannot handle the uncertainties that exist in the electricity time-series. Moreover, recurrent-based learning frameworks cannot fully model the complex correlations between electricity observations as well as external features, especially for long time-series of electricity demands (due to error accumulations [1], [13]).

To address the aforementioned limitations, we propose a new time-series learning method termed Transformer-guided Probabilistic Electricity Demand Forecasting (TPEDF), which adopts (1) the transformer architecture [14] for long-range electricity demands modeling and (2) a probabilistic state space model [15] for time-series uncertainty handling. The multi-head self-attention mechanism in transformer structure can capture global dependencies between electricity demands and complex external factors. TPEDF goes beyond the traditional recurrent-based methods (which are prone to error accumulations) and is able to infer the electricity demand distributions at each timestamp with a non-linear state space model for data uncertainties. Extensive experiments on two large-scale real-world datasets demonstrate the effectiveness of TPEDF compared to the strong baselines. We also conduct

case studies to analyze the behaviors of the prediction model on electricity demand forecasting.

II. Preliminaries

Electricity Demand Forecasting (EDF) is a typical time-series prediction problem associated with many external factors that pose significant impacts on the actual electricity demand. An illustration of the electricity time-series data (one week) in the Panama region is shown in Fig. 1. Now we formally define the EDF problem.

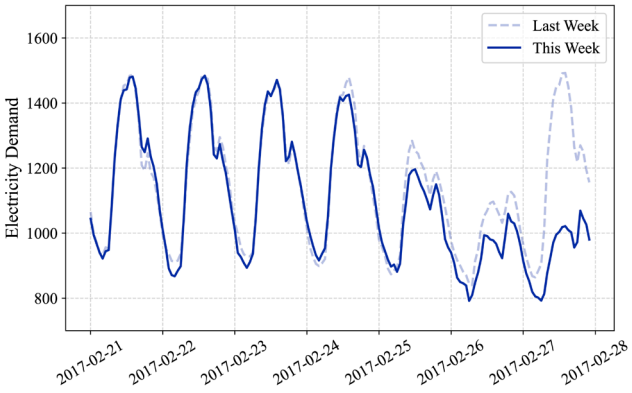

Fig. 1. An illustration of Panama electricity deman time-series.

Given a sequence of T electricity demand observations $\mathbf{X}_{1:\mathrm{T}} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_\mathrm{t}, \ldots, \mathbf{X}_\mathrm{T}\}$, each observation $\mathbf{X}_\mathrm{t} = \{\mathbf{X}_\mathrm{t}^\mathrm{e}, \mathbf{X}_\mathrm{t}^\mathrm{a}\} \in R^{d_x}$ is composed of two groups of data: (1) $\mathbf{X}_\mathrm{t}^\mathrm{e}$ is the volume of the electricity demand during the observation window; and (2) $\mathbf{X}_\mathrm{t}^\mathrm{a}$ is the auxiliary observations, e.g., *temperature* and *humidity* at time t. Here d_x is the dimension of the external factors.

In addition, we also have a set of associated external features denoted as $\mathbf{E}_{1:\mathrm{T}} = \{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_\mathrm{t}, \ldots, \mathbf{E}_\mathrm{T}\}$, where each $\mathbf{E}_\mathbf{t}$ represents certain time-dependent features (e.g., *hour of day* and *day of week*) as well as other features that may have explicit or implicit correlations with the electricity demand (e.g., price of the electricity and power plants capacities). With the above definitions, we have the EDF problem defined as:

Definition 1 (Electricity Demand Forecasting): Given historical electricity demand observations $\mathbf{X}_{1:T}$ and associated external features $\mathbf{E}_{1:T+\tau}$, we aim to predict the future electricity demands $\mathbf{X}_{T+1:T+\tau}^e$, which can be expressed as a conditional distribution:

$$p(\mathbf{X}_{T+1:T+\tau}^e | \mathbf{X}_{1:T}, \mathbf{E}_{1:T+\tau}; \Theta). \quad (1)$$

Here τ is the forecasting horizon, Θ are the learnable parameters of the forecasting model.

III. Methodology

In this section, we detail the framework of the proposed TPEDF method, which consists of a transformer-guided electricity demand learning module and a probabilistic inference module based on a state space model for estimating the electricity demand distributions. An overview of TPEDF is depicted in Fig. 2.

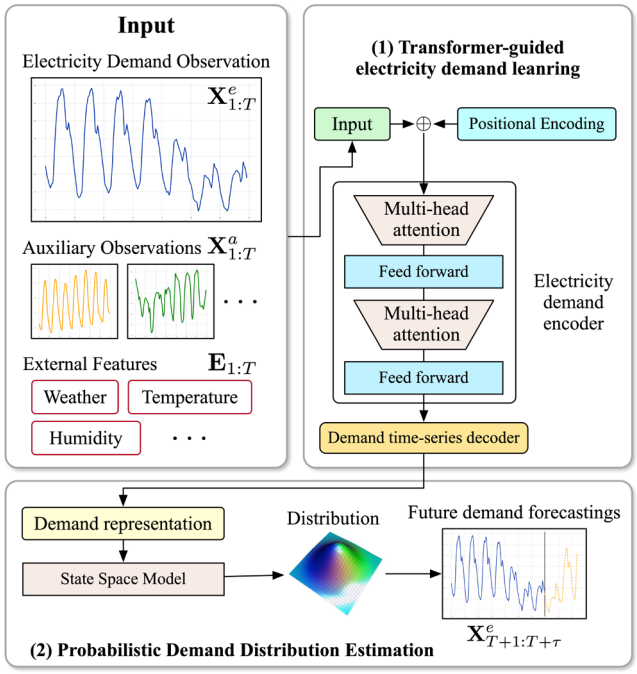

Fig. 2. An overview of the proposed TPEDF framework which composes of a transformer-guided electricity demand learning module for fusing electricity observations and features and a non-linear state space model for probabilistic demand distribution estimation and forecasting.

A. Transformer-Guided Electricity Demand Learning

To predict the future electricity demands $\mathbf{X}_{T+1:T+\tau}^e$, we need to effectively aggregate the current and historical observations of electricity demands $\mathbf{X}_{1:T}$ and related external features $\mathbf{E}_{1:T}$, such as weather and temperature. For simplicity, we denote their concatenation as $\mathcal{X}_t = [\mathbf{X}_t, \mathbf{E}_t]$. The demand learning module is followed by the encoder-decoder architecture. The encoder network comprises multiple multi-head attention layers and feed-forward layers for learning electricity demand observations while also incorporating the fusing of complex external features.

As described in [14], recurrent-based neural networks such as LSTM and GRU are inefficient in capturing the long-range dependencies between input observations, given the nature of their sequential learning manners. For complex time-series data, especially those data with long-range and periodic patterns. As we can observe from Fig. 1, the electricity demands are cyclically rising and falling, e.g., the demands are at a low level during nighttime and are rapidly growing during the daytime, and the demands are lower on weekends and higher on weekdays. To model the long-range dependencies between electricity demand observations, we use self-attention modules [16] as the building blocks of the TPEDF's encoder network. Attention modules are able to generate attention scores between the model's input variables (e.g., a large score indicates the two variables have a stronger dependency and vice versa) from a global perspective. As a consequence, both short-term and long-term dependencies of the input variables are captured, and we can learn better electricity demand observation representations for accurate demand forecasting.

Specifically, to encode the temporal information of the electricity demand sequence in a non-recurrent way, we first use the positional encoding layer for each timestamp t:

$$\mathrm{PE}_{t,2i} = \sin\left(t/10000^{2i/d_{\mathcal{X}}}\right), \quad (2)$$

$$\mathrm{PE}_{t,2i+1} = \cos\left(t/10000^{2i/d_{\mathcal{X}}}\right), \quad (3)$$

here $d_{\mathcal{X}}$ is the dimension of the input variables and $i \in \{1, \ldots, \lfloor d_{\mathcal{X}}/2 \rfloor\}$. After the positional encoding, we adopt multiple multi-head self-attention layers to learn electricity demand hidden representations. Given input $\mathcal{X}_{1:T}$, the self-attention module is defined by:

$$\mathbf{Q} = \mathcal{X}\mathbf{W}^{\mathbf{Q}}, \quad \mathbf{K} = \mathcal{X}\mathbf{W}^{\mathbf{K}}, \quad \mathbf{V} = \mathcal{X}\mathbf{W}^{\mathbf{V}}, \quad (4)$$

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_K}}\right)\mathbf{V}, \quad (5)$$

Here $\mathbf{Q}, \mathbf{K}$ and $\mathbf{V}$ are the matrices of query, key, and value obtained by multiplication of the input $\mathcal{X}$ and transformation matrices $\mathbf{W}^{\mathbf{Q}}, \mathbf{W}^{\mathbf{K}} \in R^{d_{\mathcal{X}} \times d_K}$ and $\mathbf{W}^{\mathbf{V}} \in R^{d_{\mathcal{X}} \times d_V}$. Followed by the self-attention module is the position-wise feed-forward (FF) neural networks to transform attention output $\mathbf{A}$. Then we adopt multiple self-attention networks to enhance the representation learning process, and the electricity demand hidden representation $\mathbf{H}_{1:T} = \mathrm{FF}(\mathbf{A})$ is obtained. The decoder network can be the same as the encoder network, e.g., stacking self-attention layers; other types of layers can also be used here. Next, we illustrate how to utilize the state space model for probabilistic electricity demand distribution estimation.

B. Probabilistic State Space Model

The probabilistic state space model is able to account for electricity demand uncertainties by providing a traceable multi-step forecast distribution [15]. In particular, the non-linear state space model can be described as:

$$\mathbf{L}'_t = \mathbf{T}_t\mathbf{L}_{t-1} + \mathbf{B}_t + \mathbf{O}_t\epsilon, \quad (6)$$

$$\mathcal{X}_t = \mathbf{C}_t\mathbf{L}_t + \mathbf{D}_t + \mathbf{R}_t\epsilon, \quad (7)$$

$$\mathbf{L}_t = \delta(\mathbf{L}'_t), \quad (8)$$

$$\mathcal{X}_t = \delta(\mathcal{X}'_t), \quad (9)$$

$$\epsilon \sim \mathcal{N}(0,1), \quad (10)$$

Where δ is the tan activation function, $\mathbf{L}_t \in R^{d_s}$ is the latent states, d_s is the dimension of the latent states, $\mathbf{T}_t, \mathbf{C}_t \in R^{d_s \times d_s}$ are the transition matrices from the previous timestamp to the current timestamp of the electricity demand, $\mathbf{O}_t$ and $\mathbf{R}_t$ are the variances of the latent states and observations, respectively. Then the probabilistic distributions at each timestamp t can be described as:

$$p(\mathbf{L}_0) = \mathcal{N}(\mathbf{L}_0|\mathbf{0}, \mathbf{I}), \quad (11)$$

$$p(\mathbf{L}_t) = \mathcal{N}\left(\mathbf{L}_t\middle|\delta(\mathbf{T}_t\mathbf{L}_{t-1} + \mathbf{B}_t), \mathrm{diag}(\mathbf{O}_t^2)\right), \quad (12)$$

$$p(\mathcal{X}_t) = \mathcal{N}\left(\mathcal{X}_t\middle|\delta(\mathbf{C}_t\mathbf{L}_t + \mathbf{D}_t), \mathrm{diag}(\mathbf{R}_t^2)\right), \quad (13)$$

where $\mathbf{L}_0$ is the initial latent states and diag() is the diagonal function.

Next, we show how to optimize the network and sample electricity demand forecastings from the distributions. Let $\mathbf{X}_{\mathbf{t}}^{(\mathbf{i})}$ be the $i-th$ value of the input observation $\mathbf{X}_t$ and Θ be the learnable parameters of the model, the optimization objective of TPEDF is defined as a combination of two losses:

$$\mathcal{L}_\Theta = \log p\left(\mathbf{X}_{1:T+\tau}|\mathbf{X}_{1:T}, \mathbf{E}_{1:T+\tau}\right) \quad (14)$$

$$= \sum_{t=1}^{T+\tau} \log p\left(\mathbf{X}_{\mathbf{t}}|\mathbf{X}_{1:t-1}, \mathbf{E}_{1:t}\right) \quad (15)$$

$$= \sum_{t=1}^{T+\tau} \log c_t^{\mathrm{cross}} + \sum_{i=1}^{d_s} \log p\left(\mathbf{X}_t^{(i)}\middle|\Sigma_{1:t}^*, \Phi_{1:t}\right), \quad (16)$$

where the first term of the objective optimizes a covariance $\log c_t^{\mathrm{cross}}$ defined in [1] and the second term is the log-likelihood function calculated with the Kalman filter [17]. The model parameters Φ can be updated by standard gradient descent algorithms such as Adam [18]. The electricity demand forecasting process is then iteratively performed at each timestamp $t \in [T+1, T+\tau]$.

IV. Emipirical evaluation

In this section, we evaluate the prediction performance of the proposed TPEDF model compared with state-of-the-art baselines on two real-world datasets. We first introduce the details of datasets, baselines, metrics, and experimental settings. Then we analyze the performance comparison results and conduct case studies.

A. Data

TABLE I. Panama Data Statistics

Dataset	***Panama 17-18***	***Panama 19-20***
Time	2017-2018	2019-2020
Interval	1 hour	1 hour
Electricity range (MW·h)	[380.6, 1635.9]	[85.2, 1754.9]
Avg electricity demand	1194.0	1217.5
Temperature	[19.8, 27.8]	[20.22, 39.1]
Humidity	[17.5, 90.6]	[16, 100]
Percipitation (mm)	[0, 45.8]	[0, 52.0]

We use two large-scale electricity demand forecasting datasets collected from https://www.cnd.com.pa/index.php/informes-disponibles in four years: *Panama 17-18* and *Panama 19-20*. The two datasets contain hourly electricity consumption in Panama country from 2017 to 2020. The detailed statistics are shown in Table I. External features are collected from two major cities (Panama and San Miguelito), including *hour of day*, *week day* and *holiday*.

B. Baseline

We compare TPEDF with traditional time-series methods, recurrent-based methods, as well as state-of-the-art time-series methods.

- Historical Average (HA) predicts the future electricity demands by the average of previous demands.
- AutoregRessive Integrated Moving Average ARIMA) [19] is a popular time-series forecasting model and has been used in many applications.
- Support Vector Regression (SVR) is a variant of SVM for regression task [20].

TABLE II. PERFORMANCE COMPARISON OF ELECTRICITY DEMAND FORECASTING WITH TPEDF AND BASELINES ON TWO PANAMA DATASETS IN TERMS OF RMSE, MAE, LOG p AND CRPS. THE BEST RESULTS ARE BOLDED AND THE SECOND BEST RESULTS ARE UNDERLINED. SINCE HA AND ARIMA CANNOT MAKE PROBABILISTIC PREDICTIONS, WE USE N/A TO DENOTE THEIR PERFORMANCES OF LOG p AND CRPS.

Dataset	***Panama 17-18***				***Panama 19-20***			
Metric	**RMSE**	**MAE**	**Log p**	**CRPS**	**RMSE**	**MAE**	**Log p**	**CRPS**
HA	132.2	96.79	n/a	n/a	123.7	86.10	n/a	n/a
ARIMA	101.5	86.05	n/a	n/a	100.6	80.43	n/a	n/a
SVR	92.76	70.77	2.453	0.017	78.30	58.48	2.667	0.016
LSTM	79.24	56.68	2.649	0.015	73.60	52.04	2.694	0.015
GRU	80.24	57.49	2.677	0.015	74.34	51.84	2.680	0.016
CNN-RNN	72.43	51.60	2.734	0.012	80.85	51.74	2.784	0.014
GRU-VAE	76.04	60.24	2.727	0.013	73.74	52.18	2.675	0.015
DSSM	64.14	43.66	2.801	0.010	71.21	51.87	2.793	0.013
DeepAR	66.85	45.49	2.764	0.011	67.87	46.03	2.753	0.015
PrEF	66.00	40.84	3.118	0.008	63.94	42.01	2.851	0.012
TPEDF (ours)	**51.10**	**39.56**	**3.170**	**0.007**	**54.66**	**38.76**	**3.055**	**0.008**

- Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are two dominant types of Recurrent Neural Networks (RNN) for learning sequential data such as language and time-series [12].
- CNN-RNN [21] is a combination of RNN and Convolutional Neural Network (CNN) for multi-scale electricity load forecasting with rich features and electricity price.
- GRU-VAE [22] employs a variational autoencoder (VAE) framework and takes the GRU as the encoding network.
- DSSM [15] integrates the state space model with deep learning for time-series forecasting.
- DeepAR [23] is a deep learning time-series prediction model based on autoregressive RNNs.
- PrEF [1] is the state-of-the-art electricity demand forecasting model incorporating Copula-augmented SSM and RNNs for probabilistic prediction.

C. Evaluation Metric and Experimental Settings

We employ four evaluation metrics to verify the electricity demand forecasting performance of our method and baselines: rooted mean squared error (RMSE) and mean absolute error (MAE) are used to measure prediction residuals; logarithmic density $\log p$ and continuous ranked probability score (CRPS) are used to evaluate the quality of the inferred electricity demand distributions. The four metrics are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}\left(\hat{\mathbf{X}}_{T+t}^{e} - \mathbf{X}_{T+t}^{e}\right)^2},$$

$$\text{MAE} = \frac{1}{\tau}\sum_{t=1}^{\tau}\left|\hat{\mathbf{X}}_{T+t}^{e} - \mathbf{X}_{T+t}^{e}\right|,$$

$$\log p = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{\sqrt{2\pi\hat{\sigma}_t^e}}\exp\left(-\frac{(\mathbf{X}_{T+t}^{e} - \hat{\mu}_t^e)^2}{2(\hat{\sigma}_t^e)^2}\right),$$

$$\text{CRPS} = \frac{1}{\tau}\sum_{t=1}^{\tau}\int_{-\infty}^{+\infty}\left[\left(F(\hat{\mathbf{X}}_{T+t}^{e}) - \mathbf{1}(\hat{\mathbf{X}}_{T+t}^{e} - \mathbf{X}_{T+t}^{e})\right)\right]^2$$

where $\hat{\mu}_t^e$ and $\hat{\sigma}_t^e$ are the mean and variance of the electricity distribution, respectively. $F(\cdot)$ is the cumulative distribution function of the forecasting and $\mathbf{1}(\cdot)$ is the Heaviside step function.

We split each dataset as training (50%), validation (25%), and test (25%) sets, and report the best test performance when validation loss is not declined for 50 consecutive epochs. The length of the electricity demands T is set to 48 (hours) and the forecasting horizon τ is set to 24 (hours), i.e., we utilize two days of observations to predict the hourly electricity demands of the next day. The latent states dimension in the state space model and the transformer learning module is set to 128 and 256 for all methods, respectively.

D. Experimental Results

In this section, to demonstrate the superiority of TPEDF, we conduct an experimental performance comparison on electricity demand forecasting for the proposed method and ten baselines. The results are shown in Table II and we have the following observations. Specifically, traditional time-series methods [19], [20] perform poorly and have the highest prediction errors compared to other baselines. Recurrent-based neural networks such as LSTM and CNN-RNN [21]

have better performance than traditional methods, demonstrating the superior learning capability of neural networks. In addition, state-of-the-art time-series methods generally have comparable or better prediction performance than recurrent-based methods. This might be because they have more advanced learning modules on top of recurrent/sequential neural networks such as autoregressive RNNs and state space models. However, their learning capabilities are constrained by the recurrent architecture and are inefficient in capturing long-range dependencies between electricity demand observations and external features which may have a great influence on future electricity demands. In contrast, with the transformer-guided electricity demand learning module and probabilistic demand distribution estimation incorporated in TPEDF, we achieve significant performance improvements on electricity demand forecasting by up to 14.8% and 14.5% on Panama 17-18 and Panama 19-20, respectively, in terms of RMSE. These promising results demonstrate the importance of modeling long-range time-series dependencies and electricity demand uncertainties.

We provide intuitive visualizations of the electricity demands predicted by TPEDF compared to the ground truth in Fig. 3. We can observe that TPEDF's prediction fits the ground truth well and captures the periodic variations of electricity demands. It is worth mentioning that the predicted demands are not as smooth as the ground truth lines, which contain small vibrations on local demand extremums.

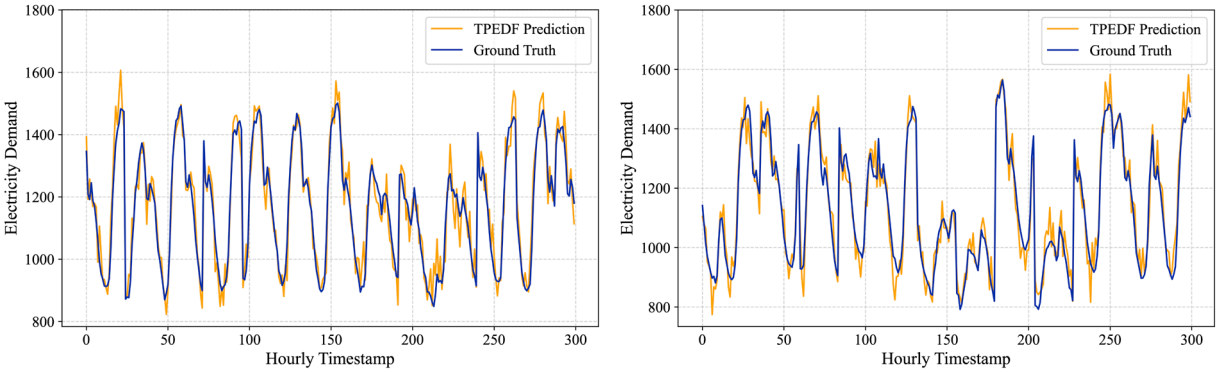

Fig. 3. Case studies on the ground-truth and prediction results of TPEDF on Panama 17-18 electricity demands.

V. Conclusion

This work presents TPEDF, a Transformer-guided probabilistic electricity demand forecasting framework for modeling the long-range dependencies of demand time-series and complex external features. TPEDF incorporates probabilistic distribution estimation for electricity demands and is able to handle the uncertainty of demands. We conducted extensive experiments on two large-scale electricity demand datasets, and the results show that our method significantly improved the performance in comparison to strong baselines. Future work including extend the model with disentangled learning [24].

References

[1] Z. Wang, X. Xu, G. Trajcevski, K. Zhang, T. Zhong, and F. Zhou, "PrEF: Probabilistic electricity forecasting via Copula-augmented state space model," in *AAAI*, 2022.

[2] A. K. Singh, S. K. Ibraheem, M. Muazzam, and D. Chaturvedi, "An overview of electricity demand forecasting techniques," *Network and Complex Systems*, vol. 3, no. 3, pp. 38-48, 2013.

[3] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing," *Journal of the Operational Research Society*, vol. 54, no. 8, pp. 799-805, 2003.

[4] A. A. Mir, M. Alghassab, K. Ullah, Z. A. Khan, Y. Lu, and M. Imran, "A review of electricity demand forecasting in low and middle income countries: The demand determinants and horizons," *Sustainability*, vol. 12, no. 15, p. 5931, 2020.

[5] A. T. Eseye, M. Lehtonen, T. Tukia, S. Uimonen, and R. J. Millar, "Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems," *IEEE Access*, vol. 7, pp. 91463-91475, 2019.

[6] "Transition in indian electricity sector the energy & resource institute," 2017. [Online]. Available: http://www.teriin.org/files/transitionreport/files/downloads/Transitions-in-Indian-Electricity-Sector Report.pdf

[7] A. Q. Huang, M. L. Crow, G. T. Heydt, J. P. Zheng, and S. J. Dale, "The feature renewable electric energy delivery and management (freedm) system: the energy internet," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 133-148, 2010.

[8] J. Bedi and D. Toshniwal, "Deep learning framework to forecast electricity deman," *Applied Energy*, vol. 238, pp. 1312-1326, 2019.

[9] M. Braun, H. Altan, and S. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," *Applied Energy*, vol. 130, pp. 305-313, 2014.

[10] P. Pełka, "Pattern-based forecasting of monthly electricity demand using support vector machine," in *IJCNN*, 2021, pp. 1-8.

[11] R. Jamil, "Hydroelectricity consumption forecast for pakistan using ARIMA modeling and supply-demand analysis for the year 2030," *Renewable Energy*, vol. 154, pp. 1-10, 2020.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

[13] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *AAAI*, vol. 35, no. 12, 2021, pp. 11106-11115.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[15] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski, "Deep state space models for time series forecasting," in *NeurIPS*, 2018.

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

[17] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, pp. 35-45, 1960.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[19] C.-M. Lee and C.-N. Ko, "Short-term load forecasting using lifting scheme and ARIMA models," *Expert Systems With Applications*, vol. 38, no. 5, pp. 5902-5911, 2011.

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

[21] X. Guo, Q. Zhao, D. Zheng, Y. Ning, and Y. Gao, "A short-term load forecasting model of multi-scale CNN-LSTM hybrid neural network

considering the real-time electricity price," *Energy Reports*, vol. 6, pp. 1046-1053, 2020.

[22] Y. Qiu, Y. Sun, C. Liu, B. Li, S. Wang, and T. Peng, "Aggregate model for power load forecasting based on conditional autoencoder," in *International Conference on Intelligent Computing*, 2021.

[23] 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.

[24] Z. Wang, X. Xu, G. Trajcevski, W. Zhang, T. Zhong, and F. Zhou, "Learning latent seasonal-trend representations for time-series forecasting," in NeurIPS, 2022.