Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ST-MRC: A Service-Oriented Probabilistic Forecasting Method for Virtual Power Plant Aggregated Coupled Loads Based on Set Attention Encoding and Multi-Scale Residual Correction

Zewei Chen

2025110861@bupt.cn

Beijing University of Posts and Telecommunications

Rongheng Lin

Beijing University of Posts and Telecommunications

Research Article

Keywords: Coupled Load Forecasting, Virtual Power Plant, Probabilistic Forecasting, Set Transformer, Multi-scale Residual Correction, Quantile Regression, Temporal Convolutional Network

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11325566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ST-MRC: A Service-Oriented Probabilistic Forecasting Method for Virtual Power Plant Aggregated Coupled Loads Based on Set Attention Encoding and Multi-Scale Residual Correction

Zewei Chen (corresponding author)
School of Computer Science (National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications
Beijing 100876, China
2025110861@bupt.cn

Rongheng Lin
School of Computer Science (National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications
Beijing 100876, China
rhlin@bupt.edu.cn

Abstract*—Virtual power plants (VPPs) represent a novel energy service paradigm that aggregates distributed resources. With the rapid development of these services, individual resources within aggregations have been increasingly integrated as flexible loads for participation in electricity market trading and demand response. However, accurate forecasting of coupled loads remains challenging due to the stochastic nature of individual loads, coupling effects among multiple entities, and dynamic membership changes within aggregations. To address these issues, this paper proposes a Set Transformer-based Multi-scale Residual Correction method ST-MRC for probabilistic load forecasting. The framework adopts a three-stage progressive architecture: individual forecasting, interaction encoding, and residual correction. Load forecasts are first generated for each entity, after which a Set Transformer encodes multi-entity interaction features via attention to capture nonlinear couplings such as coincident and complementary peak demands; subsequently, a multi-scale temporal convolutional network (TCN) with varying dilation rates extracts residual evolution patterns for adaptive error correction, and finally quantile regression produces both point forecasts and probabilistic prediction intervals to support deterministic decisions and uncertainty quantification. Experiments show that the proposed method significantly improves point forecasting accuracy. In probabilistic forecasting, it achieves high empirical coverage while producing narrower prediction intervals, effectively supporting uncertainty-aware decisions in VPP electricity market trading and demand response.

***Keywords*—Coupled Load Forecasting, Virtual Power Plant, Probabilistic Forecasting, Set Transformer, Multi-scale Residual Correction, Quantile Regression, Temporal Convolutional Network**

I. INTRODUCTION

With global climate change intensifying and dual-carbon targets driving the clean transition, virtual power plants (VPPs) have emerged as a key service solution. They aggregate distributed generation, controllable loads, and energy storage via advanced ICT and smart control to participate in power system operation and electricity markets, enhancing energy efficiency, renewable integration, and grid stability. Accurate aggregated load forecasting is critical for optimizing VPP services, affecting demand response service quality, bidding economics, and grid reliability. Considering the dynamic nature of VPPs, models must generalize across varying member sets without retraining. Current methods fail to handle set-type inputs of variable size, to model high-dimensional nonlinear interactions, or to provide uncertainty quantification for robust decisions.

The fundamental novelty of ST-MRC lies not merely in integrating existing modules, but in architecting a unified framework specifically tailored for the dynamic, heterogeneous, and uncertainty-aware nature of VPP aggregated loads. We introduce the permutation invariance and variable-length input properties of Set Transformer into VPP load forecasting, which fundamentally solves the open challenge of handling dynamically changing member sets without retraining—a limitation of conventional sequence models that assume fixed-size inputs. Furthermore, unlike generic multi-scale TCNs, we design a dedicated multi-scale TCN that explicitly extracts residual evolution patterns at short-term, medium-term, and long-term scales, directly targeting the coupling-dependent aggregation errors in traditional summation methods. By coupling set-attention-encoded interaction features with multi-scale residual corrections and quantile regression, ST-MRC simultaneously outputs point forecasts and prediction intervals in a single forward pass, offering a more efficient and robust alternative to two-stage approaches. Finally, we show that the aggregation residual variance includes covariance terms and show that our framework effectively models these nonlinear couplings, thereby providing a new theoretical perspective on VPP load forecasting beyond simple error correction.

II. RELATED WORK

From a service computing perspective, the dynamic composition and coupling of heterogeneous resources in VPPs pose unique challenges for service-level forecasting. Unlike traditional static systems, VPP services require forecasting models that are adaptable to changing service compositions and can quantify service uncertainty. Early studies combined ensemble tree models and probabilistic strategies to reduce uncertainty and enhance robustness [1]. Transfer learning leveraged knowledge from related datasets to boost forecasting across aggregation levels [2]. Multi-task learning jointly modeled correlated energy loads in integrated systems, extracting shared representations [3], and further captured couplings among wind, solar, and electrical loads [4]. Hybrid ensemble frameworks also exploited active-reactive coupling for better performance [5]. Recent research focuses on complex temporal and spatial dependencies. Spatio-temporal rolling methods update forecasting windows to improve

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

stability [6]. Hybrid architectures like DLinear with PatchTST excel in heat load forecasting by combining trend decomposition and local representations [7]. Patch-based Transformers with temporal convolution enhance cloud load prediction [8], and channel-time collaborative Transformers model both temporal evolution and cross-variable interactions for renewables [9]. Probabilistic methods combining DLinear with conformal prediction improve multi-step interval reliability while retaining point forecast accuracy [10].

Lightweight linear models have proven effective: DLinear competes with RNNs for multivariate multi-horizon forecasting at lower cost [11], and improved variants capture trends in ultra-short-term wind power load [12]. Evaluations show linear models remain competitive with Transformers across diverse scenarios [13]. Advanced architectures like Temporal Fusion Transformer model (TFT) long-term dependencies well [14], and autoregressive probabilistic RNNs handle sequential uncertainty via distributions [15]. With the rise of virtual power plants and distributed energy, forecasting now targets complex aggregation. Studies integrate load prediction with operational optimization for virtual power plants [16], use hybrid RNNs for enhanced temporal extraction [17], and employ aggregate load modeling from ambient measurements [18]. Probabilistic evaluation using CRPS emphasizes uncertainty quantification for resilience [19]. Gradient boosting with adaptive error correction improves demand forecasting under dynamic conditions [20].

Despite advances, limitations remain for dynamic aggregated coupled-load forecasting for VPPs. Existing methods assume fixed input dimensions, which hinders adaptation to changing aggregation members. Most models predict aggregated loads without modeling nonlinear interactions among heterogeneous individuals, causing systematic bias under strong coupling. Moreover, probabilistic forecasts often derive intervals directly from aggregate series, neglecting interaction-aware residuals from individual errors. To address these issues, this paper proposes a framework using Set Transformer encoding and multi-scale residual correction, combining individual forecasting, permutation-invariant interaction modeling, multi-scale residual learning, and quantile regression to capture nonlinear coupling effects and improve both deterministic accuracy and probabilistic uncertainty estimation.

In contrast to these works, ST-MRC uniquely integrates set-attention interaction modeling with multi-scale residual learning, providing a fundamentally new solution for dynamic VPP aggregation scenarios.

III. PROBLEM DESCRIPTION

In this paper, a coupled unit is the smallest measurable entity within an aggregate that possesses independent load characteristics. Let $\mathcal{U} = \{u_1, u_2, \dots, u_i, \dots\}$ denote the set of units. Each unit u_i has independent observability: its load $y_{i,t}$ can be independently measured or predicted at time t. The aggregation has a variable scale, so its membership size may change dynamically:

$$N \in \mathbb{N}^+, N \geq 1 \tag{1}$$

The aggregation is order invariant. For any permutation π, reordering its units leaves the set unchanged:

$$\{u_1, u_2, \dots, u_N\} = \{u_{\pi(1)}, u_{\pi(2)}, \dots, u_{\pi(N)}\} \tag{2}$$

Heterogeneity tolerance allows the units to have different physical attributes, load characteristics, and temporal patterns. Consider a temporary VPP aggregate containing N heterogeneous units. The actual load of unit i at time t is $y_{i,t}$, and its prediction from a unit-level model is $\hat{y}_{i,t}$. Direct summation yields the aggregated forecast:

$$\overline{Y}_t = \sum_{i=1}^{N} \hat{y}_{i,t} \tag{3}$$

The true aggregated load is the sum of the actual unit loads:

$$Y_t = \sum_{i=1}^{N} y_{i,t} = \overline{Y}_t + \sum_{i=1}^{N} \epsilon_{i,t} \tag{4}$$

Here, $\epsilon_{i,t} = y_{i,t} - \hat{y}_{i,t}$ is the unit prediction error. The aggregation residual is defined as:

$$e_t = Y_t - \overline{Y}_t = \sum_{i=1}^{N} \epsilon_{i,t} \tag{5}$$

The variance of this residual contains both individual error variances and cross-unit covariance terms:

$$\begin{aligned} Var(e_t) = {} & \sum_{i=1}^{N} Var(\epsilon_{i,t}) \\ & +2 \sum_{1 \leq i < j \leq N} Cov(\epsilon_{i,t}, \epsilon_{j,t}) \end{aligned} \tag{6}$$

The covariance terms characterize coupling effects such as coincident peaks and peak staggering. When these terms dominate the residual variance, independent unit models cannot fully capture aggregation uncertainty. Covariance affects variance rather than the residual mean directly; systematic bias arises when unit prediction errors have nonzero means. ST-MRC learns the interaction-dependent residual characteristics to obtain $\hat{e}_t$, yielding the corrected forecast:

$$\tilde{Y}_t = \overline{Y}_t + \hat{e}_t \tag{7}$$

The framework also outputs the quantile interval $[\tilde{Y}_t^{0.1}, \tilde{Y}_t^{0.9}]$ to support uncertainty-aware decisions. These endpoints define a central interval with nominal coverage of 80%; empirical coverage is evaluated separately using PICP.

IV. ALGORITHMIC FLOW DESIGN

A. Overall Framework

Although ST-MRC incorporates Set Transformer, multi-scale TCN, residual correction, and quantile regression, its core innovation resides in the novel combination and problem-specific adaptation of these modules. Specifically, the Set Transformer is tailored to encode multi-entity load interactions under dynamic VPP membership rather than performing generic set processing, whereas the multi-scale TCN captures residual evolution patterns instead of raw load sequences, directly addressing coupling-dependent residual variation. Quantile regression is applied to residuals rather than original loads, enabling sharper interval forecasts. This tight coupling creates a synergistic effect that cannot be achieved by simply stacking them, as validated by our ablation studies in Section V.D.

To address the problem defined in Section III for VPP services, we propose ST-MRC, a probabilistic forecasting method for aggregated coupled loads based on set-attention encoding and multi-scale residual correction. It adopts a three-stage architecture comprising individual prediction, interaction encoding, and residual correction to accommodate dynamic VPP membership.

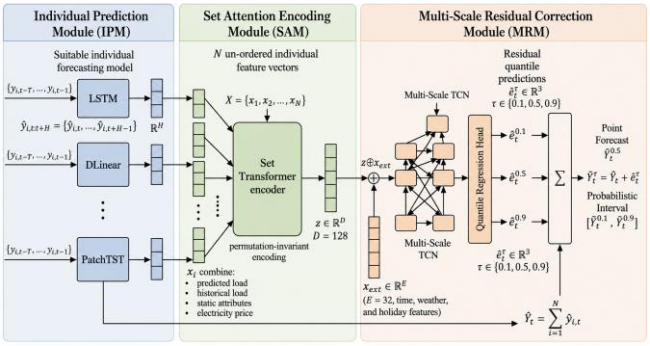


Fig. 1. Overall framework of ST-MRC

In the individual prediction module, the historical load sequence $\{y_{i,t-T}, \dots, y_{i,t-1}\}$ of each unit i is input into a suitable model (PatchTST, DLinear, or LSTM) to obtain future H-step predictions $\hat{y}_{i,t:t+H-1}$. Direct summation gives $\overline{Y}_t = \sum_{i=1}^{N} \hat{y}_{i,t}$. In the interaction encoding module, the feature set $X = \{x_1, x_2, \dots, x_N\}$ is encoded by a Set Transformer into a global interaction vector $z \in \mathbb{R}^{128}$. This module supports permutation invariance and variable membership sizes. The vector z is concatenated with external features $x_{\text{ext}} \in \mathbb{R}^{32}$, and a multi-scale TCN extracts residual evolution patterns. A quantile regression head outputs $\hat{e}_t^{0.1}$, $\hat{e}_t^{0.5}$, and $\hat{e}_t^{0.9}$. The corrected point forecast is computed as:

$$\tilde{Y}_t^{0.5} = \overline{Y}_t + \hat{e}_t^{0.5} \tag{8}$$

The corresponding central 80% prediction interval is:

$$[\tilde{Y}_t^{0.1}, \tilde{Y}_t^{0.9}] = [\overline{Y}_t + \hat{e}_t^{0.1}, \overline{Y}_t + \hat{e}_t^{0.9}] \tag{9}$$

The selection of PatchTST and DLinear as individual load forecasters is motivated by their complementary strengths in modeling heterogeneous unit loads. PatchTST, a Transformer-based model with a patching strategy, excels at capturing local temporal dependencies and seasonal patterns in high-frequency load data, making it suitable for units with strong intraday periodicity such as commercial buildings. DLinear, on the other hand, decomposes time series into trend and seasonal components and applies linear layers, which is particularly effective for loads dominated by linear trends or gradual drifts, such as slow-varying industrial processes. Both models are lightweight, support multi-step forecasting, and have demonstrated superior performance in recent energy forecasting benchmarks. Using them as representatives of nonlinear and linear paradigms ensures that the individual forecasts $\hat{y}_{i,t}$ are reasonably accurate, which is crucial because the residual correction module in ST-MRC relies on high-quality base predictions to isolate coupling effects. LSTM is included as a traditional recurrent baseline for comparison, but PatchTST and DLinear form the primary choices due to their robustness and computational efficiency.

B. Set Attention Feature Encoding Module

To accommodate dynamic aggregation membership, ST-MRC uses a Set Transformer with permutation invariance and variable input length. For an input set $X = \{x_1, x_2, \dots, x_N\}$, N denotes the current number of units and $x_i \in \mathbb{R}^d$ is the feature vector of unit i. Each feature vector combines future predictions, historical load observations, static attributes, and historical electricity prices.

The forecasting features comprise the future H-step predictions $\hat{y}_{i,t}, \hat{y}_{i,t+1}, \dots, \hat{y}_{i,t+H-1}$, while the historical features contain the past T load observations $y_{i,t-T}, \dots, y_{i,t-1}$. Static attributes include unit type, rated capacity, historical mean and standard deviation, and the peak-to-valley load difference. The past T time-of-use electricity prices $p_{t-T}, \dots, p_{t-1}$ provide additional information about shared economic drivers.

All the above features undergo Z-score standardization before input to eliminate dimensional differences.

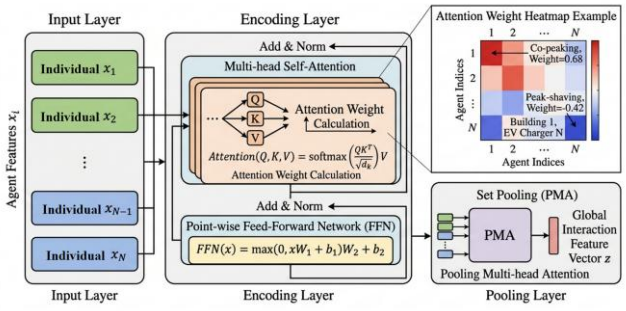


Fig. 2. Process of set attention feature encoding in ST-MRC

As shown in Figure 2, the set attention feature encoding process mainly consists of three stages. First, in the Input Projection and Multi-head Self-Attention Encoding stage, the input features $X \in \mathbb{R}^{N \times d}$ are projected into a hidden space:

$$X_{proj} = XW^{proj} + b^{proj} \tag{10}$$

Where $b^{proj} \in \mathbb{R}^{d_h}$ are learnable parameters, and d_h is the hidden layer dimension. A multi-head self-attention mechanism calculates the interaction weights between any two units within the set, automatically capturing coupling patterns such as peak coincidence and peak staggering. The scaled dot-product attention is computed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K, V are the query, key, and value matrices, $d_k = d_v = d_h/h$ is the key-value vector dimension, and $\sqrt{d_k}$ is a scaling factor to prevent vanishing gradients. Multi-head attention concatenates the results of multiple independent attention heads:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head_h)W^0 \tag{12}$$

$$head_i = Attention(X_{proj}W_i^Q, X_{proj}W_i^K, X_{proj}W_i^V) \tag{13}$$

$$W^0 \in \mathbb{R}^{hd_v \times d_h} \tag{14}$$

where h is the number of attention heads, and W^0 is the output projection matrix.

Next, in the Feed-Forward Network Transformation stage, the output of the self-attention layer undergoes residual connection and layer normalization:

$$X_{attn} = LayerNorm(X_{proj} + MultiHead(Q,K,V)) \tag{15}$$

This is then fed into a point-wise feed-forward network for non-linear transformation:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{16}$$

where $W_1 \in \mathbb{R}^{d_h \times 4d_h}, b_1 \in \mathbb{R}^{4d_h}, W_2 \in \mathbb{R}^{4d_h \times d_h}, b_2 \in \mathbb{R}^{d_h}$ are learnable parameters shared across all positions. This is followed again by residual connection and layer normalization:

$$X_{ffn} = LayerNorm(X_{attn} + FFN(X_{attn})) \tag{17}$$

The above self-attention and feed-forward network layers are stacked for a total of L layers.

Finally, in the set pooling and global feature output stage, after L layers of encoding, the encoded feature matrix $X_{enc} \in \mathbb{R}^{N \times d_h}$ is obtained. Pooling multi-head attention maps the variable-length set encoding into a fixed-length global feature vector:

$$z = MultiHead(S, X_{enc}, X_{enc}) \tag{18}$$

where $S \in \mathbb{R}^{1 \times d_h}$ is a learnable query vector. Through the attention mechanism, it adaptively aggregates the encoded features of all individual units, ultimately obtaining a global interaction feature $z \in \mathbb{R}^{D}$, and we set $D = 128$ in this paper. The global feature z contains interaction information among all units within the aggregate, effectively capturing coupling effects like peak coincidence and peak staggering between buildings and charging stations, which is crucial for accurate VPP aggregated load forecasting.

C. Multi-scale Residual Correction and Probabilistic Prediction Module

Considering multi-time scale evolutionary characteristics of aggregate residuals, we design a multi-scale TCN for residual prediction, combined with quantile regression to achieve probabilistic output. The ST-MRC method adopts three types of time scales: short-term (15 minutes) residuals are mainly affected by random fluctuations in individual loads; medium-term (1 hour) residuals are primarily influenced by time-period coupling effects; long-term (4 hours) residuals are mainly driven by external factors like electricity prices and weather within a VPP context. The specific structure is shown in Figure 3.

The input received by this module is the concatenated vector of the global interaction feature $z \in \mathbb{R}^{128}$ and the external features $x_{ext} \in \mathbb{R}^{32}$, which can be expressed as:

$$x_{input} = z \oplus x_{ext} \in \mathbb{R}^{160} \tag{19}$$

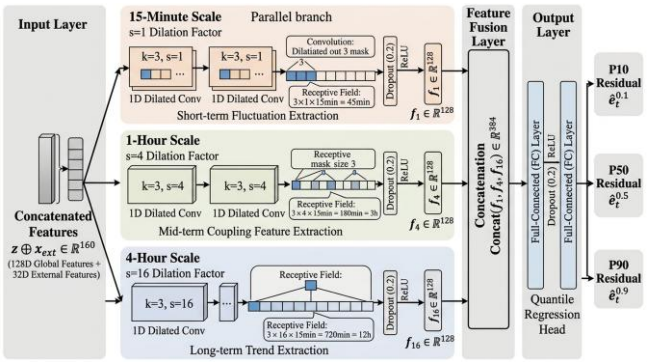


Fig. 3. The structure of multi-scale residual correction and probabilistic prediction module in ST-MRC

The aggregation residual $e_t = \sum_{i=1}^{N} \epsilon_{i,t}$ has a variance that can be dominated by covariance terms reflecting physical coupling among units. Positive residual values, which indicate that the true aggregated load exceeds the sum of individual predictions, can occur under specific load profiles. For instance, when multiple heterogeneous units such as commercial buildings and EV charging stations simultaneously experience their daily peaks in the early evening, their individual loads become positively correlated; individual models ignore these cross-unit interactions and tend to underpredict each unit's peak, leading to a large positive residual. Another common scenario is extreme weather events like heatwaves, which cause a widespread increase in cooling or heating loads across all buildings and charging infrastructure, which can induce positively correlated forecast errors and underprediction. Additionally, during demand response events, distributed resources may shift their consumption to the same off-peak periods, creating an unintended synchronized load pattern that individual forecasters fail to capture. Conversely, negative residuals arise when loads are complementary or when peak-staggering strategies are effectively employed. By modeling these residual patterns through multi-scale TCN, ST-MRC learns statistical signatures of these physical coupling mechanisms, thereby enhancing both interpretability and accuracy.

The external features x_{ext} specifically include 12-dimensional time features and 20-dimensional weather features. Specifically, the time features include hour encoding (1-24 normalized), day-of-week encoding (1-7 normalized), month encoding (1-12 normalized), and holiday indicator (0-1 encoding). The weather features specifically include temperature, humidity, wind speed at the prediction time t, and one-hot encoding of weather conditions (sunny, cloudy, overcast, rainy). Weather features change dynamically with the forecast period, reflecting the temporal heterogeneity of meteorological conditions. The multi-scale TCN uses dilated convolutions to extract meteorological evolution patterns at different time scales from historical weather time series, thereby modeling the impact of weather on aggregate residuals.

The multi-scale TCN in the ST-MRC method adopts a parallel branch structure, where each branch uses dilated convolution with different dilation rates to capture residual patterns at different time scales. This can be formally expressed as:

$$f_s = TCN_s(x_{input}), \quad s \in \{1,4,16\} \tag{20}$$

In our ST-MRC method, each TCN branch has the same structure. The feature maps from the three TCN branches are concatenated along the feature dimension to obtain the multi-scale fusion feature:

$$f_{fusion} = Concat(f_1, f_4, f_{16}) \in \mathbb{R}^{384} \tag{21}$$

And the fused feature is input into the quantile regression head to output residual predictions at different quantile levels:

$$\hat{e}_t^{\tau} = FC_{\tau}(f_{fusion}), \quad \tau \in \{0.1,0.5,0.9\} \tag{22}$$

The structure of the above quantile regression head consists of a first fully connected layer:

$$g_1 = max\left(0, W_{f1} f_{fusion} + b_{f1}\right) \tag{23}$$

where $W_{f1} \in \mathbb{R}^{128 \times 384}$ is the weight matrix, and $b_{f1} \in \mathbb{R}^{128}$ is the bias. This is followed by a dropout layer, and then subsequently, there is a second fully connected layer:

$$g_2 = W_{f2} g_{1,drop} + b_{f2} \tag{24}$$

The output is:

$$\begin{bmatrix} \hat{e}_t^{0.1} \\ \hat{e}_t^{0.5} \\ \hat{e}_t^{0.9} \end{bmatrix} = g_2 \tag{25}$$

The quantile regression uses the pinball loss as the training objective. For a given quantile level τ, the loss function is defined as:

$$L_\tau = \frac{1}{M} \sum_{t=1}^{M} \rho_\tau \left(e_t - \hat{e}_t^\tau\right) \tag{26}$$

Here, e_t is the true residual defined in Section III, M is the number of training samples, and $\rho_\tau(u)$ is the pinball loss function:

$$\rho_\tau(u) = \begin{cases} \tau u u \geq 0 \\ (1-\tau)(-u) u < 0 \end{cases} \tag{27}$$

The ST-MRC method can be naturally extended to multi-step prediction scenarios. For a prediction horizon k > 1 , the load at future times is predicted recursively:

$$\tilde{Y}_{t+k}^{0.5} = \sum_{i=1}^{N} \hat{y}_{i,t+k} + \hat{e}_{t+k}^{0.5} \tag{28}$$

$$[\tilde{Y}_{t+k}^{0.1}, \tilde{Y}_{t+k}^{0.9}] = [\overline{Y}_{t+k} + \hat{e}_{t+k}^{0.1}, \overline{Y}_{t+k} + \hat{e}_{t+k}^{0.9}] \tag{29}$$

where $\hat{y}_{i,t+k}$ is the k-step ahead prediction from the individual model, and $\hat{e}_{t+k}^{0.1}, \hat{e}_{t+k}^{0.5}, \hat{e}_{t+k}^{0.9}$ are the corresponding residual predictions.

V. EXPERIMENTS

A. Experimental Datasets

To comprehensively validate the proposed method, this paper uses coupled load datasets from three independent sources: two domestic commercial park datasets (Dataset 1, Dataset 2) and one European large-scale heterogeneous aggregator dataset (Dataset 3).

For domestic commercial parks, experiments use coupled load data from commercial buildings and charging stations. Dataset 1 contains 3 months of 15-minute granularity data (Oct-Dec 2024), including aggregated building and charging station loads and time-of-use electricity prices. Average building load is 544.8 kW (range: 186.4-1078.8 kW); average charging station load is 233.8 kW (range: 2.6-1206.6 kW). Dataset 2 covers the same period with similar structure; average building load is 931.7 kW (181.5-2407.2 kW), average charging station load is 217.5 kW (2.1-1142.4 kW).

For the European dataset, based on the PanTaGruEl pan-European transmission grid and heterogeneous aggregation data source [21], we take the European regional transmission network as the backbone and combine it with OPSD (Open Power System Data) load time series [22], which are disaggregated to consumer nodes via the admittance matrix, to construct virtual power plant aggregator units. Ten nodes are selected as aggregation members, covering six heterogeneous resource clusters: commercial building groups (2 nodes), distributed PV (2 nodes), EV charging stations (2 nodes), energy storage (2 nodes), adjustable industrial loads (1 node), and cold-chain flexible loads (1 node). Nodes can dynamically join or leave under market demand response scenarios, forming Dataset 3. The dataset includes day-ahead electricity prices, renewable outputs (solar and wind, MW), temporal features, temperature, humidity, and wind speed. The data span from January 2015 to September 2020 (approximately 5.75 years, hourly granularity, 50377 records), covering multiple annual cycles to test generalization and stability.

B. Baseline Models

For point forecasting, we compare ST-MRC with seven baseline methods. The direct-summation baseline aggregates unit-level predictions obtained from LSTM models. PatchTST applies Transformer-based patching to extract temporal patterns, whereas DLinear provides a computationally efficient linear forecasting model. LightGBM-Res corrects aggregation residuals using gradient boosting, and DeepSets provides a permutation-invariant set-based comparator.

For probabilistic forecasting, six baseline methods use quantile regression to produce prediction intervals. PatchTST-QR combines patch-based temporal encoding with three quantile output branches, and DLinear-QR predicts trend and seasonal components using branches optimized with pinball loss.

TFT-QR combines the Temporal Fusion Transformer with quantile regression. Its attention mechanisms and gated residual networks capture temporal dependencies and interactions with external features, while parallel output branches predict the selected quantiles. DeepAR-QR uses a recurrent encoder followed by quantile output layers to model sequential uncertainty.

To further demonstrate the advantage of ST-MRC, we additionally include two recent VPP-specific methods for point prediction: VPP-STGAT, a spatio-temporal graph attention network designed for VPP aggregated load forecasting that models inter-unit dependencies using graph structures (2025), and DeepRes-VPP, a deep residual network with channel-wise attention for coupled load forecasting in virtual power plants (2024). For probabilistic interval prediction, we also compare against VPP-STGAT-QR and DeepRes-VPP-QR, which combine these new architectures with quantile regression. These additions allow for a more comprehensive evaluation against state-of-the-art methods tailored to dynamic aggregation scenarios.

C. Evaluation Metrics

Point forecasts are evaluated using mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE), measuring absolute accuracy, sensitivity to large errors, and relative error. Prediction interval coverage probability (PICP) measures empirical coverage, while the reported PINAW measures interval width. Coverage and width are assessed jointly against the nominal coverage implied by the interval endpoints.

D. Experimental Results and Analysis

Point forecasting experiments were conducted on Datasets 1 and 2 to evaluate aggregated load accuracy under different heterogeneous load profiles. Tables I and II report the resulting MAE, RMSE, and MAPE values.

Tables I and II demonstrate the effectiveness of ST-MRC for aggregated load point forecasting. ST-MRC achieves the lowest MAE and MAPE on both datasets, with MAE values of 24.54 and 40.34 and MAPE values of 3.94% and 4.43%, respectively. Relative to direct summation, MAE decreases by

48.9% on Dataset 1 and 40.7% on Dataset 2, while RMSE decreases by 48.6% and 34.3%. On Dataset 1, the RMSE of 32.09 also represents reductions of 25.5% relative to DLinear and 5.5% relative to DeepRes-VPP. On Dataset 2, the RMSE of 58.08 improves upon DLinear and DeepSets by 9.0% and 1.4%, respectively. These results show that set-attention encoding captures interaction information that complements individual forecasts, while multi-scale residual learning corrects aggregation errors. The consistent MAE and MAPE improvements across the two datasets demonstrate the effectiveness of this joint design for heterogeneous load profiles. Figures 4 and 5 illustrate its behavior over typical daily trajectories.

TABLE I. COMPARISON OF POINT FORECASTING RESULTS FOR DATA SOURCE 1

Model	MAE	RMSE	MAPE(%)
Baseline	48.07	62.49	6.86
PatchTST	36.30	47.19	5.18
DLinear	33.14	43.08	5.32
LightGBM-Res	29.84	38.79	4.79
DeepSets	27.97	36.36	4.49
VPP-STGAT	26.52	34.81	4.28
DeepRes-VPP	25.88	33.95	4.15
ST-MRC	**24.54**	**32.09**	**3.94**

TABLE II. COMPARISON OF POINT FORECASTING RESULTS FOR DATA SOURCE 2

Model	MAE	RMSE	MAPE(%)
Baseline	68.03	88.44	7.40
PatchTST	53.60	75.04	5.83
DLinear	49.10	63.82	5.34
LightGBM-Res	46.15	60.07	5.02
DeepSets	45.33	58.92	4.93
VPP-STGAT	43.21	57.45	4.71
DeepRes-VPP	42.56	56.78	4.65
ST-MRC	**40.34**	**58.08**	**4.43**

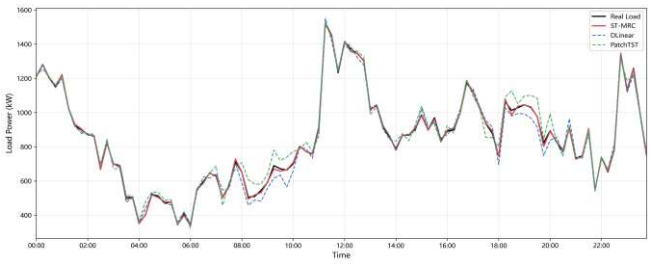


Fig. 4. Comparison of typical daily coupled load forecasting results (Data source 1)

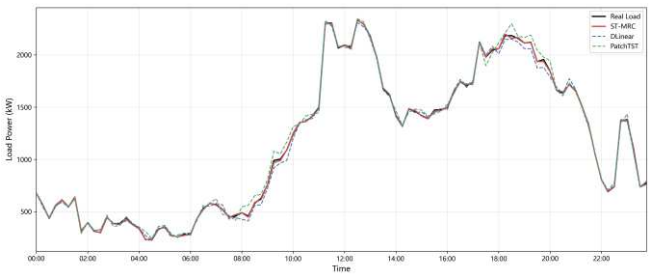


Fig. 5. Comparison of typical daily coupled load forecasting results (Data source 2)

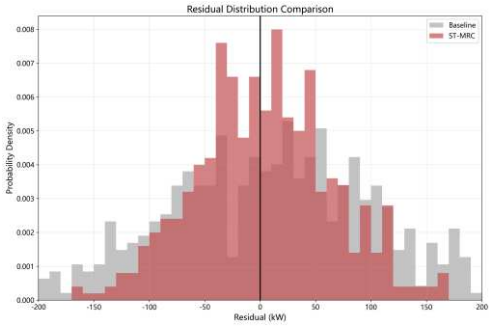


Fig. 6. Histogram of residual distribution (Dataset 1, Baseline vs ST-MRC)

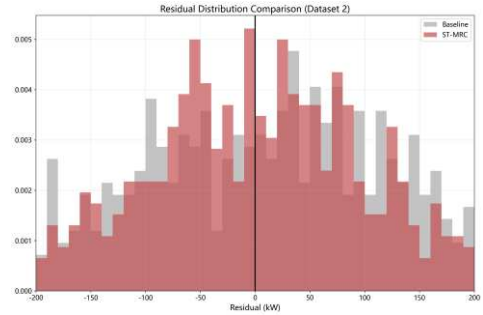


Fig. 7. Histogram of residual distribution (Dataset 2, Baseline vs ST-MRC)

Figures 6 and 7 illustrate the residual correction achieved by ST-MRC on Datasets 1 and 2. The horizontal axis represents the difference between actual and predicted aggregated load in kW, and the vertical axis shows normalized probability density. Compared with the direct-summation residuals, the ST-MRC residuals are more concentrated around zero, with a tighter distribution and a higher central peak. This concentration indicates improved agreement between the predicted and observed loads and a reduction in systematic aggregation error. The consistent pattern across both datasets supports the effectiveness of interaction-aware residual learning under different load scales and coupling conditions.

For Dataset 3, multi-step forecasting was evaluated at horizons of 1, 6, 12, 24, and 48 steps. Because this dataset has hourly granularity, these horizons correspond to 1, 6, 12, 24, and 48 hours, respectively. Figure 8 and Table III summarize the results.

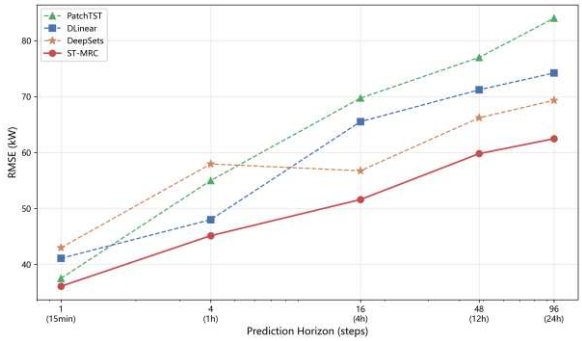


Fig. 8. RMSE comparison under different forecast horizons (Data source 3)

Table III shows that ST-MRC maintains the lowest RMSE at every evaluated horizon, from 1 to 48 steps. Its RMSE is 10,443 at one step and 17,184 at 48 steps, corresponding to reductions of 22.6% and 10.8% relative to DLinear at the

respective horizons. At 6, 12, and 24 steps, ST-MRC likewise achieves RMSE values of 14,038, 14,462, and 15,527, demonstrating consistent accuracy across short and extended forecast windows. For the hourly Dataset 3, the 48-step horizon spans two days. The sustained accuracy over this range supports the use of ST-MRC for VPP scheduling and market decisions that require forecasts at multiple lead times.

The multi-step results further demonstrate the value of combining cross-unit interaction encoding with residual evolution modeling. Set attention preserves information about the dependencies among individual unit forecasts, and the multi-scale TCN uses interaction features, external drivers, and historical information to learn residual patterns over different temporal ranges. This complementary design delivers the lowest absolute RMSE throughout Table III. From one to 48 steps, ST-MRC increases from 10,443 to 17,184 while retaining its accuracy advantage at each intermediate horizon. These results confirm the effectiveness of residual correction for producing useful long-horizon aggregated forecasts and support forecasting-based VPP operational planning.

TABLE III. RMSE COMPARISON ACROSS DIFFERENT FORECAST HORIZONS (UNIT: MW, DATA SOURCE 3)

Model	1 step	6 steps	12 steps	24 steps	48 steps
PatchTST	15289	15791	17404	18247	20888
DLinear	13495	14732	16487	18276	19267
DeepSets	12085	15126	16784	17251	18785
ST-MRC	**10443**	**14038**	**14462**	**15527**	**17184**

For probabilistic forecasting, ST-MRC is compared with PatchTST-QR, DLinear-QR, TFT-QR, DeepAR-QR, VPP-STGAT-QR, and DeepRes-VPP-QR. Each method outputs quantiles at 0.1, 0.5, and 0.9. Tables IV, V, and VI report interval coverage and width for Datasets 1, 2, and 3, respectively.

TABLE IV. COMPARISON OF PROBABILISTIC PREDICTION INTERVAL RESULTS (DATA SOURCE 1)

Model	PICP(%)	PINAW
PatchTST-QR	77.4	134.0
DLinear-QR	79.4	126.7
TFT-QR	76.7	157.1
DeepAR-QR	85.2	206.0
VPP-STGAT-QR	86.5	118.2
DeepRes-VPP-QR	87.2	115.6
ST-MRC	**90.2**	**105.3**

TABLE V. COMPARISON OF PROBABILISTIC PREDICTION INTERVAL RESULTS (DATA SOURCE 2)

Model	PICP(%)	PINAW
PatchTST-QR	76.4	141.7
DLinear-QR	78.7	134.6
TFT-QR	81.4	208.1
DeepAR-QR	85.8	158.8
VPP-STGAT-QR	85.9	127.5
DeepRes-VPP-QR	86.3	124.1
ST-MRC	**88.9**	**117.3**

Tables IV and V demonstrate that ST-MRC combines high empirical coverage with the narrowest reported intervals on Datasets 1 and 2. ST-MRC achieves PICP values of 90.2% and 88.9%, together with PINAW values of 105.3 and 117.3, respectively. Relative to DeepRes-VPP-QR, the interval width is reduced by 8.9% on Dataset 1 and 5.5% on Dataset 2, while empirical coverage increases by 3.0 and 2.6 percentage points. Relative to PatchTST-QR, the width decreases by 21.4% and 17.2%, respectively. The joint improvement in coverage and sharpness indicates that learning the conditional residual distribution captures aggregation uncertainty efficiently. The specified 0.1 and 0.9 quantiles define nominal 80% central intervals; their measured PICP values describe the empirical inclusion of observed loads. Figures 9 and 10 illustrate how the intervals follow changing load trajectories, providing useful uncertainty information for electricity market bidding and demand response decisions.

TABLE VI. COMPARISON OF PROBABILISTIC PREDICTION INTERVAL RESULTS (DATA SOURCE 3)

Model	PICP(%)	PINAW
PatchTST-QR	74.5	29579
DLinear-QR	79.6	27903
TFT-QR	80.6	26768
DeepAR-QR	83.5	24256
VPP-STGAT-QR	85.1	23120
DeepRes-VPP-QR	85.8	22540
ST-MRC	**90.7**	**21859**

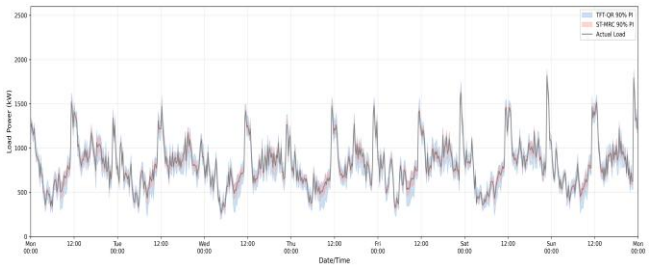


Fig. 9. Visualization of probabilistic prediction intervals (Data source 1)

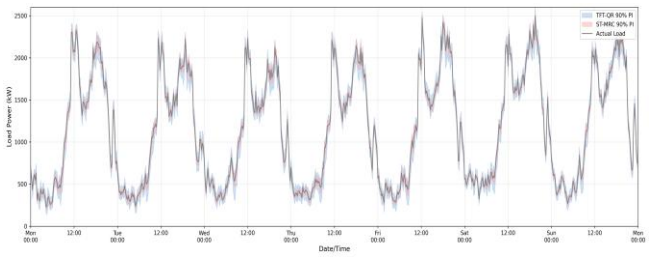


Fig. 10. Visualization of probabilistic prediction intervals (Data source 2)

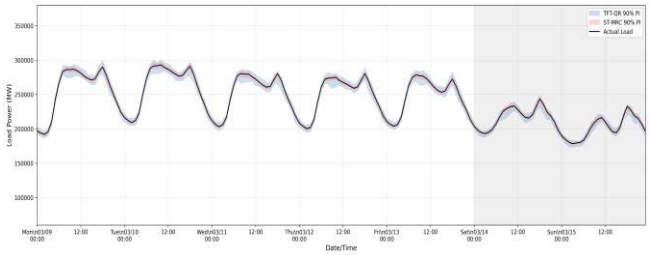


Fig. 11. Visualization of probabilistic prediction intervals (Data source 3)

Table VI extends the probabilistic evaluation to the large heterogeneous aggregation in Dataset 3. ST-MRC achieves a PICP of 90.7% and the smallest reported PINAW of 21,859.

Relative to DeepRes-VPP-QR, the reported interval width decreases by 3.0% and coverage increases by 4.9 percentage points. The width reductions relative to DLinear-QR and DeepAR-QR are 21.7% and 9.9%, respectively. These results demonstrate that interaction-aware residual forecasting retains its effectiveness at a larger aggregation scale. Figure 11 shows that the predicted intervals track the aggregate load trajectory, complementing the quantitative evidence of efficient uncertainty representation.

Concerning the coupled load forecasting performance under different aggregation scales, experiments were designed for Data Source 3 involving different numbers of units in a multi-coupled-unit scenario. Comparison charts of weekly coupled load forecasting results under different aggregation scales are shown in Figure 12 below, and the variation of prediction errors under different aggregation scales is shown in Figure 13 below.

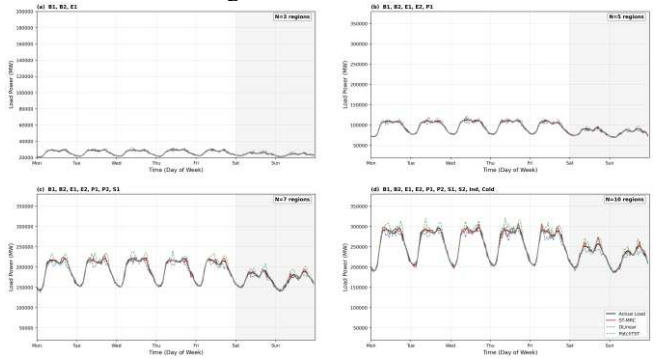


Fig. 12. Comparison of weekly coupled load forecasts under different aggregation scales

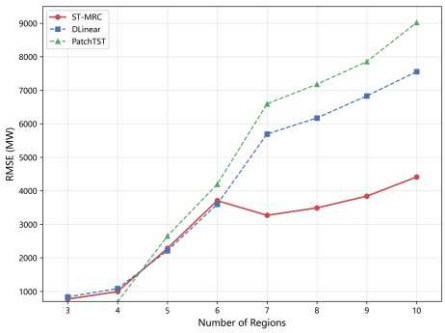


Fig. 13. Variation of forecast error under different aggregation scales

Figures 12 and 13 evaluate ST-MRC across aggregation sizes ranging from three to ten units. The weekly trajectories in Figure 12 show that ST-MRC follows the principal changes in aggregate load as membership size increases. Figure 13 highlights its favorable error behavior at larger aggregation scales, where the information captured by set attention and multi-scale residual learning becomes particularly useful. The relatively flat RMSE growth from three to ten units indicates that the framework effectively represents cross-unit dependencies as more heterogeneous resources are aggregated. These results demonstrate the value of permutation-invariant encoding for VPP forecasting under variable membership and support its applicability to larger commercial aggregations.

To evaluate each core component of ST-MRC, ablation experiments on Dataset 1 separately remove or replace Set Transformer encoding, residual correction, multi-scale TCN processing, quantile regression, and external features. Table VII reports point forecast errors and the relative RMSE increase against the full model. Table VIII evaluates the corresponding changes in probabilistic interval quality.

Table VII demonstrates the complementary contributions of the core ST-MRC components. The full model achieves the lowest RMSE of 32.09. Compared with the mean-pooling variant, Set Transformer encoding reduces RMSE by 12.8%, confirming the value of attention-based cross-unit interaction modeling. Relative to direct summation, the full residual correction framework reduces RMSE by 16.5%, demonstrating its effectiveness in learning aggregation errors. Multi-scale processing improves RMSE by 6.2% relative to the single-scale variant with dilation 1 and by 4.0% relative to the variant with dilation 4, supporting the joint use of residual patterns at short, medium, and long temporal ranges. Quantile regression also contributes to point forecasting, with a 2.4% RMSE improvement relative to MSE-only training. Including external features yields a further 5.0% improvement relative to the variant without weather and electricity-price information. These results show that interaction encoding, residual correction, temporal scale fusion, distributional learning, and external drivers contribute distinct predictive information, enabling the full framework to achieve the strongest point forecasting performance in the ablation study.

TABLE VII. POINT FORECASTING COMPONENT ABLATION RESULTS ON DATASET 1

Model Variant	RMSE	Relative Increase in RMSE (%)
ST-MRC Full Model	**32.09**	–
w/o Set Transformer (replaced by mean pooling)	36.82	14.7%
w/o Residual Correction (direct summation only)	38.45	19.8%
w/o Multi-scale TCN (single-scale, dilation=1)	34.21	6.6%
w/o Multi-scale TCN (single-scale, dilation=4)	33.41	4.1%
w/o Quantile Regression (point forecast via MSE loss)	32.87	2.4%
w/o External Features (no weather/electricity price)	33.78	5.3%

Table VIII demonstrates the effectiveness of quantile regression and its integration with interaction encoding and residual correction. The full ST-MRC model achieves a PICP of 90.2% and a PINAW of 105.3. Relative to the Gaussian-assumption variant, quantile regression increases empirical coverage by 7.7 percentage points and reduces interval width by 24.1%, demonstrating more efficient uncertainty representation. With quantile regression retained, integrating Set Transformer encoding increases PICP from 86.4% to 90.2% and narrows PINAW from 121.9 to 105.3. Similarly, the complete residual correction and multi-scale TCN components improve coverage and sharpness relative to their respective ablation variants. The full model combines the highest empirical coverage with the narrowest intervals among the tested configurations, confirming the

complementary roles of its components in probabilistic aggregated load forecasting.

TABLE VIII. PROBABILISTIC COMPONENT ABLATION RESULTS ON DATASET 1

Model Variant	PICP (%)	PINAW
ST-MRC Full Model (with QR)	**90.2**	**105.3**
w/o Quantile Regression (Gaussian assumption + MSE)	82.5	138.7
w/o Set Transformer + QR	86.4	121.9
w/o Residual Correction + QR	85.1	128.4
w/o Multi-scale TCN (single-scale) + QR	88.3	112.6

VI. CONCLUSIONS

A probabilistic load forecasting method for aggregated coupled loads, termed ST-MRC, is proposed by integrating set attention-based feature encoding and multi-scale residual correction. Experimental results demonstrate that the set attention-based feature encoding effectively captures the interactive features among heterogeneous resources within an aggregation, adapting well to the dynamic membership of temporary aggregations. The multi-scale residual correction module further captures combined effect deviations across different temporal scales. Compared with baseline models, the proposed method achieves improved accuracy while providing reliable probabilistic prediction intervals for temporary aggregated coupled load forecasting, which is fundamental for delivering dependable and efficient VPP services.

Future work will proceed in several directions. First, a dynamic attention mechanism can be introduced to further enhance adaptability to changes in aggregation members. Second, reinforcement learning methods may be incorporated to realize joint optimization of forecasting and scheduling. Moreover, the approach can be extended to multi-energy coupled system forecasting scenarios to broaden its applicability.

DECLARATIONS

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Funding Declaration

This work was supported by the State Grid Corporation of China (52060026006W).

Ethics approval

Not applicable.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

REFERENCES

[1] H. Shahinzadeh, H. Nafisi, A. Zamani, S. Mehrabani-Najafabadi, A. Mahmoudi and F. Ebrahimi, "Probabilistic Short-Term Load Forecasting Using GBDT-Based Sister Forecasts and Ensemble Methods," 2025 15th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, Islamic Republic of, 2025, pp. 1-6, doi: 10.1109/ICCKE68588.2025.11273339.

[2] K. Jiang and D. L. Donaldson, "Transfer Learning Based Short-Term Load Forecasting at Different Aggregation Levels," 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), Muscat, Oman, 2024, pp. 1-5, doi: 10.1109/ICIESTR60916.2024.10798263.

[3] Z. Li, G. Li, H. Liu and Y. Shi, "Short-Term Prediction of Multi-Load in Integrated Energy Systems Based on Coupling Characteristic Analysis and Multi-Task Learning," 2024 Second International Conference on Cyber-Energy Systems and Intelligent Energy (ICCSIE), Shenyang, China, 2024, pp. 1-6, doi: 10.1109/ICCSIE61360.2024.10698123.

[4] F. Yang et al., "A Short Term Wind-Photovoltaic-Load Forecasting Model Based on Coupling Relationship and Multi-Layer Extraction Multi-Task Learning," 2024 7th International Conference on Power and Energy Applications (ICPEA), Taiyuan, China, 2024, pp. 621-625, doi: 10.1109/ICPEA63589.2024.10785028.

[5] C. Li et al., "Active load forecasting based on hybrid ensemble learning framework considering active-reactive coupling," 2025 IEEE Industry Applications Society Annual Meeting (IAS), Taipei, Taiwan, 2025, pp. 1-6, doi: 10.1109/IAS62731.2025.11061413.

[6] J. Chen, B. Fan, Z. Xu and C. Hu, "Spatio-Temporal Rolling Forecasting Method for Power Load Process Characteristics," 2025 5th International Conference on New Energy and Power Engineering (ICNEPE), Guangzhou, China, 2025, pp. 152-155, doi: 10.1109/ICNEPE67923.2025.11384310.

[7] J. Song, Q. Dong, K. Wang and X. Gao, "A novel heat load prediction algorithm based on DLinear-PatchTST model," 2025 IEEE 7th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2025, pp. 829-833, doi: 10.1109/CISCE65916.2025.11065816.

[8] X. Xue, Z. Xia, Y. Wang and W. Huo, "Cloud Computing Resource Load Prediction Model Based on TCN-PatchTST," 2025 9th International Conference on Cloud and Big Data Computing (ICCBDC), Manchester, United Kingdom, 2025, pp. 45-49, doi: 10.1109/ICCBDC67784.2025.00014.

[9] K. Lu, M. Huo, Y. Li, Q. Zhu and Z. Chen, "CT-PatchTST: Channel-Time Patch Time-Series Transformer for Long-Term Renewable Energy Forecasting," 2025 10th International Conference on Computer and Information Processing Technology (ISCIPT), Fushun, China, 2025, pp. 86-95, doi: 10.1109/ISCIPT67144.2025.11265471.

[10] Y. Mei, M. He, L. Zhou and J. Che, "Reliable Multi-Step Probabilistic Forecasting for Electricity Load: A DLinear and Split Conformal Prediction Approach," 2025 9th International Conference on Smart Grid and Smart Cities (ICSGSC), Chengdu, China, 2025, pp. 451-455, doi: 10.1109/ICSGSC68313.2025.11288275.

[11] M. Kamal, F. Rahman, Q. A. Fiddina, S. Rafif and R. A. Setyo Prayoga, "A Comparative Study of LSTM and DLinear for Multi Horizon Multivariate Traffic Forecasting," 2025 5th International Conference of Science and Information Technology in Smart Administration (ICSINTESA), Yogyakarta, Indonesia, 2025, pp. 684-689, doi: 10.1109/ICSINTESA68165.2025.11413717.

[12] P. Cai, C. Peng, K. Zou and P. Huang, "Ultra-Short-Term Wind Power Forecasting Based on the Improved Dlinear Model," 2024 14th International Conference on Power and Energy Systems (ICPES), Chengdu, China, 2024, pp. 414-419, doi: 10.1109/ICPES63746.2024.10856666.

[13] S. Albelali and M. Ahmed, "Noise and Sequence Length in Time-Series Forecasting: a Unified Evaluation of Transformers, DLinear, LSTM, and GRU," 2026 22nd International Conference on Intelligent Environments (IE), Lisbon, Portugal, 2026, pp. 1-4, doi: 10.1109/IE69249.2026.11539036.

[14] S. Rath and M. P, "Air Pollution Forecasting Using Machine Learning with Temporal Fusion Transformer and Graph Neural Networks," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2025, pp. 1798-1803, doi: 10.1109/IDCIOT64235.2025.10914714.

[15] S. Shi, X. Qiu, Y. Ru and X. Tan, "A DeepAR-Based Neural Network for Time Series Forecasting," 2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS), Tianjin, China, 2023, pp. 1-7, doi: 10.1109/DOCS60977.2023.10294820.

[16] S. Wang, Z. Guo, X. Li, J. Chen and Y. Guo, "Research on Load Forecasting and Dispatching Strategy of Virtual Power Plant Based on Big Data," 2025 2nd International Conference on Telecommunications and Power Electronics (TELEPE), Frankfurt, Germany, 2025, pp. 67-71, doi: 10.1109/TELEPE66277.2025.00018.

[17] A. Han, X. Chen, S. Chen, X. Wang, X. Zhang and C. Li, "Research on Virtual Power Plant Load Forecasting Based on RNN-LSTM Hybrid Model," 2025 International Conference on Electrical Engineering, Automation and Information Science (EEAIS), Lanzhou, China, 2025, pp. 91-94, doi: 10.1109/EEAIS66172.2025.11170672.

[18] M. Mukherjee and A. M. Kulkarni, "Low-Frequency and Small-Signal Aggregate Load Model Estimation using Ambient Measurements," 2024 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/ISGTAsia61245.2024.10876325.

[19] V. P. Singh, "Probabilistic Forecasting Accuracy Assessment for GW-Scale Solar PV Plants: Leveraging CRPS and Ensemble Methods for Grid Resilience," 2025 IEEE PES Conference on Innovative Smart Grid Technologies - Middle East (ISGT Middle East), Dubai, United Arab Emirates, 2025, pp. 1-6, doi: 10.1109/ISGTMiddleEast65737.2025.11314298.

[20] M. Irfan Khan, P. Swati Vaidya, R. Pathak, D. Kumar Yaduwanshi, C. Nayak and I. Khan, "Adaptive Demand Forecasting Using LightGBM with Agentic AI Error Correction," 2025 International Conference on Emerging Engineering Technologies and Applications (IC-EETA), Indore, India, 2025, pp. 1095-1099, doi: 10.1109/IC-EETA66496.2025.11548026.

[21] M. Gillioz, G. Dubuis, and P. Jacquod, "A large synthetic dataset for machine learning applications in power transmission grids," Sci. Data, vol. 12, Art. no. 168, 2025, doi: 10.1038/s41597-025-04479-x.

[22] F. Wiese, I. Schlecht, W.-D. Bunke, C. Gerbaulet, L. Hirth, M. Jahn, F. Kunz, C. Lorenz, J. Mühlenpfordt, J. Reimann, and W.-P. Schill, "Open Power System Data – Frictionless data for electricity system modelling," Appl. Energy, vol. 236, pp. 401–409, 2019, doi: 10.1016/j.apenergy.2018.11.097.