POWER AND ENERGY FUTURE

Open Access
Research Article

How Shall We Evaluate Load Forecasts?

Yangze Zhou, Dalin Qin, and Yi Wang

***Abstract*—Load forecasting plays a pivotal role in energy system management, supporting critical processes including planning, operation, and control. Traditionally, the performance of load forecasting was assessed using various statistical metrics, such as mean absolute percentage error (MAPE). However, the evolution of modern power systems has resulted in increasingly diverse load forecasting scenarios, which pose new challenges and requirements for forecast evaluation. This paper provides an analysis of the current forecast evaluation framework and identifies important issues that warrant further exploration, including: (1) the critical aspects of metric calculation, such as zeros and outliers; (2) the evaluation of probabilistic forecasting; (3) the assessment of multiple forecasts with hierarchical structures; and (4) the development of cost-oriented forecast evaluation approaches. These insights underscore the need to advance load forecast evaluation from isolated accuracy metrics toward an integrated framework aligned with the practical requirements of power systems.**

***Index Terms*—Load forecasting, forecast evaluation, probabilistic forecasting, hierarchical forecasting, cost-oriented evaluation.**

ABBREVIATION

CDF	cumulative distribution function
CRPS	continuous ranked probability score
DERs	distributed energy resources
GANs	generative adversarial networks
HLF	hierarchical load forecasting
HVAC	heating, ventilation, and air conditioning
MAAPE	mean arctangent absolute percentage error
MAE	mean absolute error
MAPE	mean absolute percentage error
MedAD	median absolute deviation
RMAPE	resistant MAPE
RMSE	root mean square error
SMAPE	symmetric MAPE

Manuscript received 16 October, 2025; accepted 13 January, 2026. The work was supported by the National Natural Science Foundation of China (52477130). *(Corresponding author: Yi Wang, e-mail: yiwang@eee.hku.hk)*

Yangze Zhou, Dalin Qin, and Yi Wang are with the Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong SAR, China, and are also with The University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen, 518057, China.

Color versions of one or more of the figures in this paper are available online at https://doi.org/10.26599/PEF.2026.9650004.

Digital Object Identifier 10.26599/PEF.2026.9650004

I. INTRODUCTION

Load forecasting is a cornerstone of energy system management, underpinning essential functions such as planning, operation, control, and market-clearing processes. Recent advances in artificial intelligence, coupled with the proliferation of sensor and smart meter data, have propelled rapid progress in load forecasting methodologies. Researchers have addressed a range of challenges in this field, including probabilistic forecasting [1], multi-timescale forecasting [2], interpretable forecasting [3], and forecasting under distribution shifts [4]. In addition, a series of forecasting competitions (e.g., [5]) are instrumental in advancing the field of load forecasting. Nevertheless, the assessment of these increasingly sophisticated models still predominantly relies on conventional evaluation metrics. Traditionally, the evaluation of load forecasting models relied on metrics such as MAPE, MAE, and RMSE, which serve as standard benchmarks for model comparison and improvement. However, the evolution of modern power systems has led to increasingly diverse load forecasting scenarios, highlighting the limitations of these conventional evaluation approaches.

One major source of this challenge arises from the volatility and intricate patterns observed in modern load curves. The widespread integration of distributed renewable energy resources, such as photovoltaic installations situated behind the meter, has made accurate forecasting more challenging. Generation from these units is often invisible to retailers and system operators [6], fundamentally altering net load profiles and introducing additional forecasting uncertainty [7]. These developments not only complicate the forecasting process but also highlight the limitations of traditional evaluation metrics. Most conventional metrics focus solely on pointwise errors, overlooking the overall similarity in shape and dynamic patterns between forecasted and actual load curves.

Another important factor is the emergence of flexible loads, such as electric vehicles [8] and HVAC systems [9]. These loads can shift consumption patterns, increasing both the volatility and unpredictability of aggregate load profiles. The integration of multiple energy carriers, exemplified by the electrification of heating systems, adds further complexity to load dynamics [10]. Moreover, the increased frequency and intensity of extreme weather events have amplified the uncertainty inherent in load forecasting [11]. As a result, load distributions are increasingly complex, often exhibiting multimodal

© The author(s) 2026. Published by Tsinghua University Press.
This is an open access article under the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

[12] or heavy-tailed characteristics [13]. Traditional probabilistic forecasting metrics, such as pinball loss, which are typically designed for unimodal or central-focused distributions, are insufficient for such scenarios.

The ongoing decentralization of power systems has introduced new requirements for load forecasting as well. Forecasts are now needed for a wide range of entities, including individual users, regional grids, and microgrids, across multiple temporal [1] and spatial scales [14]. In these hierarchical contexts, traditional evaluation methods developed for single time series are often insufficient. System operators must not only evaluate the accuracy of aggregated load forecasts, but also ensure coherence between forecasts at different hierarchical levels. For example, the sum of sub-regional forecasts should align precisely with the overall city-level forecast. Consequently, focusing solely on accuracy fails to capture the critical need for consistency and structural alignment across various levels of aggregation in modern power systems.

In addition, forecasting errors in power systems can have substantial operational and economic consequences, particularly during rare but high-impact events such as extreme weather or holidays. Although these scenarios constitute only a small fraction of historical data, inaccuracies in their forecasts can lead to severe supply–demand imbalances and even catastrophic system failures [15]. Conventional evaluation metrics tend to emphasize average performance and may not adequately penalize errors in these critical scenarios. As a result, rare but consequential events can be overlooked, potentially leading to significant risks in downstream decision-making.

In summary, the changing landscape of modern power systems calls for a more comprehensive evaluation framework for load forecasting. Against this backdrop, a fundamental question arises: How can load forecasts be evaluated using the most appropriate metrics? This article seeks to address this question by providing insights from the following four perspectives:

(1) When assessing forecast accuracy, particular attention must be paid to several critical aspects. These include inherent characteristics of the time series, such as near-zero values and outliers, as well as specific requirements of power system operation, such as preserving load curve shapes.

(2) The evaluation of probabilistic forecasts requires additional considerations. In particular, it is necessary to address complex distributional characteristics such as multimodality and heavy-tailed behaviors that frequently occur in modern power systems.

(3) The evaluation of multiple forecasts with hierarchical structures introduces unique challenges. These include the proper aggregation of errors across different levels, the balance between individual accuracy and overall consistency, and the fair comparison across datasets with varying degrees of predictability.

(4) Comprehensive evaluation should extend beyond pure accuracy metrics. It should also incorporate the practical impacts of forecasts, including their influence on operational cost savings and the resources invested in producing the forecasts.

II. How accurate are forecasts

The commonly used metrics for evaluating load forecasts measure how well the forecast $\hat{l}$ aligns with the ground truth l by using statistical formulations. However, these metrics may not be unsuitable in specific scenarios, which can be divided into two primary aspects: inherent characteristics of the time series and specific requirements of power system operation.

Firstly, load data often involves near-zero values, outliers, and high volatility, as shown in Fig. 1. The presence of near-zero values in the time series data can compromise the numerical stability of metrics like MAPE, leading to disproportionately large error values. One solution is to replace the actual values in the denominator to truncate the excessively large MAPE, such as SMAPE and RMAPE [16]. SMAPE replaces the original denominator with the average of the forecasts and actual values, while RMAPE introduces a lower bound in the denominator. The definitions of SMAPE and RMAPE are given as follows:

$$\begin{aligned}\mathrm{MAPE} &= \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{l}_t - l_t}{l_t}\right| \\ \mathrm{SMAPE} &= \frac{1}{T}\sum_{t=1}^{T}\left|\frac{2(\hat{l}_t - l_t)}{l_t + \hat{l}_t}\right| \\ \mathrm{RMAPE} &= \frac{1}{T}\sum_{t=1}^{T}\left|\frac{l_t - \hat{l}_t}{\max(r, l_t)}\right|\end{aligned} \tag{1}$$

where T is the number of samples, l_t and $\hat{l}_t$ are the tth actual and forecast values, respectively, and r is the threshold selected for RMAPE. Another approach is to normalize the MAPE values through transformation. The MAAPE maps the MAPE into $[0, \pi/2)$ using the arctan function [17]:

$$\mathrm{MAAPE} = \frac{1}{T}\sum_{t=1}^{T}\arctan\left(\left|\frac{l_t - \hat{l}_t}{l_t}\right|\right) \tag{2}$$

This transformation mitigates sensitivity to large MAPE values. However, this approach also reduces sensitivity to extreme scenarios such as holidays or high temperatures. Consequently, it may not effectively assess the model's performance in these scenarios, which operators typically care about most.

Moreover, high volatility and outliers can significantly affect error metrics, including MAPE and MAE. Several studies suggest that the MedAD, as defined in Eq. (3), which considers the median of absolute errors, may offer improved robustness over MAE in such scenarios [18]:

$$\begin{aligned}\mathrm{MAE} &= \frac{1}{T}\sum_{t=1}^{T}\left|l_t - \hat{l}_t\right| \\ \mathrm{MedAD} &= \mathrm{Median}\left(\left|l_t - \hat{l}_t\right|\right), \quad \forall t \in [1, T]\end{aligned} \tag{3}$$

The results in Fig. 1 show that outliers have a greater impact

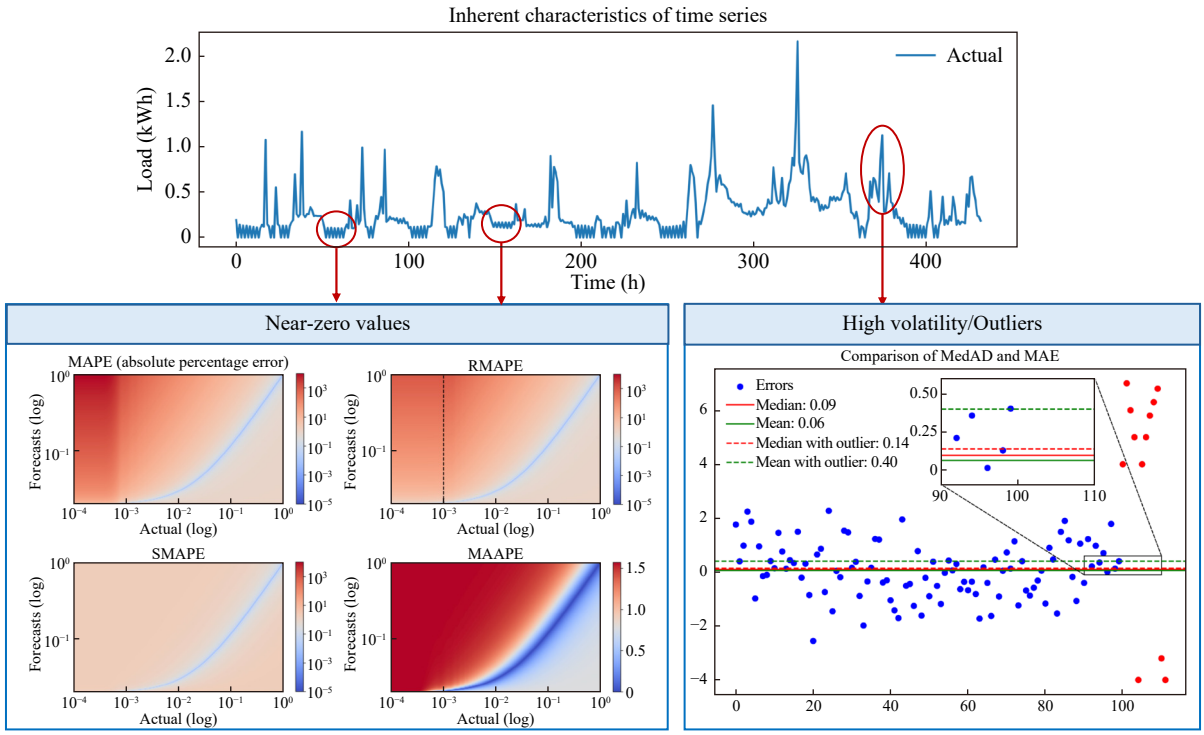

Fig. 1. The illustration of how time series inherent characteristics impact indicator selection. Near-zero values: The value of indicators with varying actual loads and forecasts, where both the *x*-axis and *y*-axis are on a logarithmic scale. High volatility or outliers: 100 data points have been generated from a Gaussian distribution (mean = 0, variance = 1) and 12 outliers have been introduced to compare the robustness of MedAD and MAE.

on the mean values (changing from 0.06 to 0.4) than on the median values (changing from 0.09 to 0.14). This suggests that MedAD is more robust to high volatility and outliers. However, this does not necessarily mean that MedAD should always be preferred. The selection of indicators often has a dual nature: Greater robustness to outliers typically comes at a cost of insufficient evaluation of extreme scenarios.

Secondly, traditional metrics may be insufficient to adequately evaluate forecasts under the specific requirements of system operation, as shown in Fig. 2. For example, system

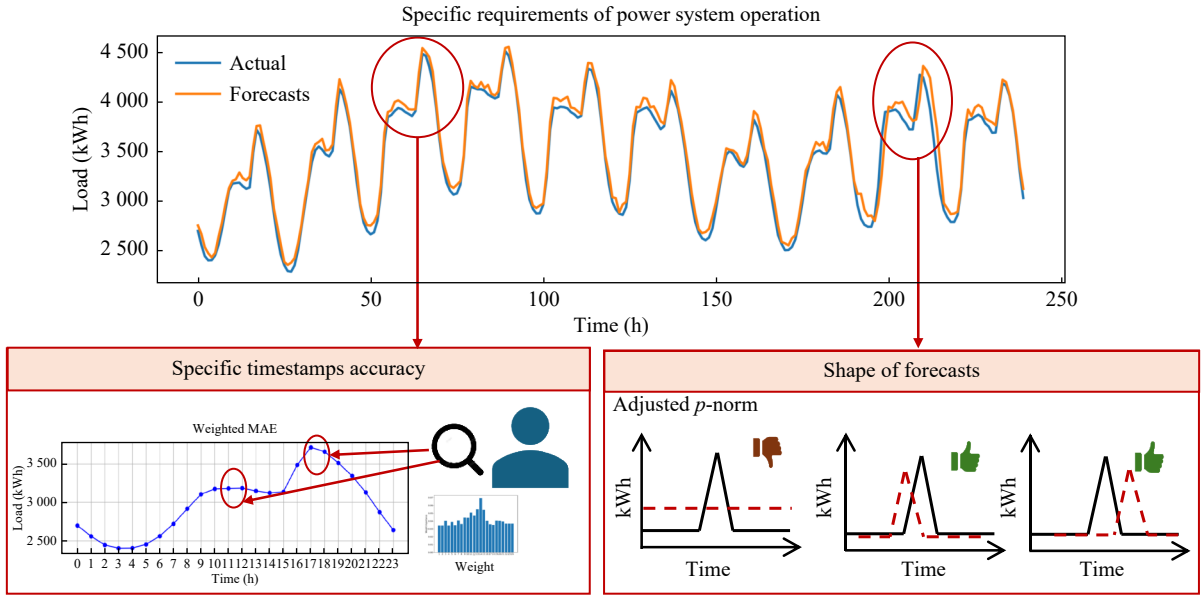

Fig. 2. The illustration of how specific requirements of power system operation impact indicator selection. The figure on the left illustrates that system operators prioritize forecasting accuracy during specific periods. The figure on the right demonstrates that mainstream accuracy metrics are insufficient for capturing the similarity between the forecasts and actual load curve shapes.

operators may prioritize accuracy during specific periods. In this case, different weights will be assigned to the accuracy values at different timestamps, resulting in a weighted accuracy metric (e.g., weighted MAE) [19]. This method is easy to compute and practical, making it popular for evaluating forecasts in power grids. The shape of load profiles is also important for system operators. When assessing the shape of the forecast curve, some methodologies employ the adjusted p-norm error, as defined in Eq. (4), which tolerates a certain degree of mistakes in forecasting the occurrence time but imposes higher penalties on errors of peak magnitude [20]:

$$P-\text{norm}(w,p)=\min_{P\in\mathcal{P}(w)}\left\|(P\hat{l})-l\right\|_p \tag{4}$$

where P is a permutation matrix that reorders the predicted values within a limited range, $\mathcal{P}$ is the set of all permutation matrices satisfying the maximum displacement constraint w, and p is the order of the norm.

The adjusted p-norm error may yield similar error values as long as the peak falls within a certain time interval, but it may cause a different impact on system operation. Specifically, forecasting peak load time early may cause grid operators to start backup generators too early, resulting in unnecessary fuel costs. Conversely, forecasting the peak load too late may result in voltage fluctuations or frequency instability. Moreover, the adjusted p-norm error can be computationally intensive, especially for large datasets. The characteristics of different metrics have been summarized in Table 1. When employing metrics in time series analysis, it is essential to incorporate practical knowledge of the actual power system operation.

TABLE 1
THE SUMMARY OF THE CHARACTERISTICS OF DIFFERENT METRICS

Metric	Characteristic
MAPE	Sensitive to near-zero values
SMAPE	Replace denominator of MAPE with the average of $\hat{l}$ and l
RMAPE	Introduces a lower bound in the denominator of MAPE
MAAPE	Normalize MAPE with arctan function
MedAD	Robust to high volatility and outliers
p-norm	Pay more attention to the shape of load profiles

III. USE OF PROBABILISTIC FORECASTS IN AN UNCERTAIN WORLD

The increasing penetration of DERs and the electrification of end-use sectors, particularly transportation, have significantly increased the complexity and uncertainty of energy demand. This paradigm shift has necessitated the adoption of probabilistic approaches in load forecasting, leveraging probability theory and stochastic modeling techniques to quantify and manage uncertainty. Probabilistic methods can be categorized into parametric and non-parametric approaches [1]. Parametric approaches quantify the uncertainty in the form of a given distribution with key parameters (e.g., mean and variance). Examples of non-parametric methods include kernel density estimation, quantile regression, and scenario generation using advanced models (e.g., GANs [21] and diffusion models [22]). In probabilistic forecasting, various indices are employed to assess the quality of probabilistic forecasts from the perspectives of reliability and sharpness. Commonly used metrics include the pinball loss, CRPS, and Winkler score. The pinball loss (also known as quantile loss) is used to evaluate the accuracy of quantile forecasts. For a quantile level $\tau\in(0,1)$, the pinball loss is defined as:

$$Q_\tau(l)=\begin{cases}\tau(l-\hat{l}), & \text{if } l\geqslant\hat{l}\\ (1-\tau)(\hat{l}-l), & \text{if } l<\hat{l}\end{cases} \tag{5}$$

CRPS is widely used for evaluating the accuracy of probabilistic forecasts, particularly for continuous variables. It measures the difference between the CDF of the forecast f^{CDF} and l:

$$\text{CRPS}(f^{\text{CDF}},l)=\int_{-\infty}^{\infty}\left(f^{\text{CDF}}(x)-\mathbb{I}\{x\geqslant l\}\right)^2dx \tag{6}$$

where f^{CDF} is the forecasted CDF, and $\mathbb{I}\{x\geqslant l\}$ is the indicator function that equals 1 if $x\geqslant l$ and 0 otherwise. Importantly, CRPS can be regarded as the average of the pinball loss over all quantile levels, which provides a comprehensive assessment of the probabilistic forecast across the entire distribution.

As for the Winkler score, it is used to evaluate forecast intervals. For a forecast interval $[L,U]$ with nominal coverage probability $1-\alpha$, the Winkler score is given as follows:

$$\text{Winkler}(L,U,l,\alpha)=\begin{cases}U-L+\dfrac{2}{\alpha}(L-l), & l<L\\ U-L, & L\leqslant l\leqslant U\\ U-L+\dfrac{2}{\alpha}(l-U), & l>U\end{cases} \tag{7}$$

where L and U are the lower and upper bounds of the forecast interval, respectively, and α is the significance level. Notably, the pinball loss can be regarded as a special case of the Winkler score when the prediction interval collapses to a single point (i.e., $L=U=\hat{l}$). In this situation, the penalty coefficients for predictions falling below or above the predicted value are $1-\alpha$ or α, respectively.

However, these metrics may have limitations when the predictive distribution is multimodal or heavy-tailed, as illustrated in Fig. 3. The pinball loss and the Winkler score can be problematic because they implicitly favor unimodal predictive behavior. In practice, the distribution of load tends to be multimodal, exhibiting multiple peaks. When traditional metrics are applied to such multimodal distributions, they might fail to capture the true performance of the forecast. For instance, during a typical day in a metropolitan area, the load might have distinct morning and evening peaks due to commuting patterns and household activities. Alternatively, when multiple types of renewable energy are present at a single node, the net

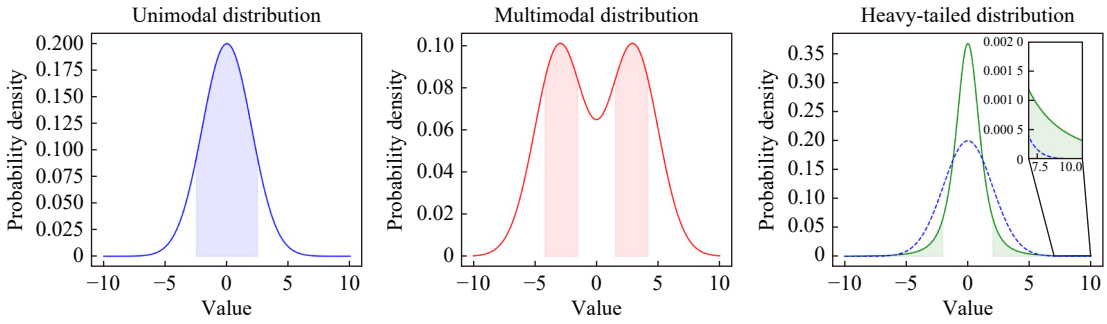

Fig. 3. The illustration of multimodal distribution and heavy-tailed distribution. When loads follow multimodal or heavy-tailed distributions, existing probabilistic metrics often fail to adequately reflect multiple peaks or the tail behavior.

load can also exhibit a multimodal distribution [12]. Furthermore, the occurrence of extreme scenarios, such as holidays or extreme weather, will impact the distribution of load, leading to a heavy-tailed distribution [13]. Traditional probabilistic metrics tend to focus more on the central part of the distribution and do not adequately penalize errors in the tails. As a result, rare but high-impact events may not be well considered, potentially causing a significant impact on the downstream decision-making process. In contrast to deterministic forecasting, where various solutions have already been proposed to address different scenarios, probabilistic forecasting currently lacks well-established metrics that effectively address these two issues.

IV. Forecasting in Future Decentralized Power Systems

The decentralization of modern power systems has led to a multitude of load profiles organized in deep spatio-temporal hierarchies, fundamentally redefining the requirements for load forecast evaluation. Although conventional metrics like MAE or RMSE are sufficient for assessing individual time series, they are conceptually inadequate for hierarchical contexts. They fail to capture the critical need for coordination in multi-level forecasts and cannot offer a comprehensive evaluation of forecast performance across different parts of the hierarchy or among various datasets. This section critically examines the challenges of transitioning from traditional single-forecast evaluation to the more complex domains of multi-hierarchy and multi-dataset forecast evaluation.

Multi-hierarchies: When performing multi-time series forecasting, different time series often exhibit hierarchical relationships from both temporal and spatial perspectives. Hierarchical forecasting aims to exploit these structural relationships to generate forecasts that are not only accurate at each level but also coherent, meaning that the forecasts for lower-level series sum up precisely to the forecasts of the series above them [23]. For instance, the sum of hourly forecasts for a day must equal the total daily forecast (temporal coherence), and the sum of regional load forecasts must match the system-wide forecast (spatial coherence).

Let the actual value of the aggregated (top-level) load at time t be $l_{0,t}$, and the sub-layer (bottom-level) loads be $l_{i,t}$, for $i \in [1,n]$. The hierarchical structure imposes the aggregation constraint $l_{0,t} = \sum_{i=1}^{n} l_{i,t}$. We first generate base forecasts, $\hat{l}_{0,t}$ and $\hat{l}_{i,t}$, which are generally incoherent. A reconciliation process then adjusts them to produce coherent forecasts, $\widetilde{l}_{0,t}$ and $\widetilde{l}_{i,t}$, which must satisfy the structural constraint:

$$\widetilde{l}_{0,t} = \sum_{i=1}^{n} \widetilde{l}_{i,t}, \quad \forall t \in [1,T] \tag{8}$$

A truly effective evaluation framework must therefore assess not only the accuracy of the base forecasts ($\hat{l}$) but also the quality of the final coherent forecasts ($\widetilde{l}$). However, evaluating HLF still faces unresolved challenges, especially concerning the spatio-temporal coupling of error aggregation, the assessment of multi-level probabilistic forecast errors, and the inherent trade-off between forecast accuracy and coherence.

Current error metrics for HLF predominantly treat the temporal and spatial dimensions as separable, which is a critical oversimplification that ignores their intrinsic coupling. This limitation is critical because forecast errors at lower levels are often correlated, not independent [24], highlighting that ignoring such cross-sectional correlations can lead to suboptimal reconciliation. For example, consider the error of an aggregate forecast formed via the simple bottom-up method, where $\widetilde{l}^{\mathrm{BU}}_{0,t} = \sum_{i=1}^{n} \hat{l}_{i,t}$. The aggregate forecast error, $e^{\mathrm{BU}}_{0,t} = l_{0,t} - \widetilde{l}^{\mathrm{BU}}_{0,t}$, is the sum of the bottom-level base forecast errors, $e_{i,t} = l_{i,t} - \hat{l}_{i,t}$:

$$e^{\mathrm{BU}}_{0,t} = \sum_{i=1}^{n} l_{i,t} - \sum_{i=1}^{n} \hat{l}_{i,t} = \sum_{i=1}^{n} (l_{i,t} - \hat{l}_{i,t}) = \sum_{i=1}^{n} e_{i,t} \tag{9}$$

This mathematical relationship demonstrates how a large positive error in one sub-region (e.g., $e_{1,t} = +100$ MW) can be canceled out by a large negative error in another (e.g., $e_{2,t} = -95$ MW), resulting in a deceptively small aggregate error ($e^{\mathrm{BU}}_{0,t} = 5$ MW). This spatio-temporal error compensation creates an illusion of accuracy while masking critical imbalances, rendering existing evaluation frameworks mathematically incomplete and operationally misleading.

Probabilistic HLF further compounds the challenges [25], [26], where ensuring coherence is far more complex. In this setting, the goal is to produce coherent predictive distributions.

Let the random variable representing the coherent forecast for the bottom-level series i be $\tilde{L}_{i,t}$, and for the top-level series be $\tilde{L}_{0,t}$. Probabilistic coherence requires that the distribution of the aggregate forecast is mathematically derived from the joint distribution of its constituent forecasts. Formally, this means their CDF must satisfy:

$$f^{\mathrm{CDF}}(\tilde{L}_{0,t}) = f^{\mathrm{CDF}}\left(\sum_{i=1}^{n} \tilde{L}_{i,t}\right) \tag{10}$$

This formal requirement directly implies that simply aggregating properties of the marginal distributions, such as quantiles, is statistically invalid. If $Q_\tau(l)$ denotes the τ-quantile of the distribution of a random variable l, then in general:

$$Q_\tau\left(\sum_{i=1}^{n} l_{i,t}\right) \neq \sum_{i=1}^{n} Q_\tau(l_{i,t}) \tag{11}$$

This inequality holds because the distribution of a sum of random variables is the convolution of their individual distributions, which depends heavily on their correlation. As discussed in [27], conventional methods often ignore this. Furthermore, conventional probabilistic metrics like CRPS or Winkler scores, applied to marginal distributions, cannot distinguish between errors from poor individual forecasts versus those from incoherent aggregation. Existing studies still use "standard" probabilistic metrics designed for flat hierarchies to evaluate inherently hierarchical systems while lacking an appropriate measure theory to account for coupled uncertainty spaces. To this end, probabilistic HLF evaluation needs to be further investigated, possibly through methodologies like copula theory, which can explicitly model the joint distribution and complex dependencies of forecast errors across the hierarchy [28].

In both deterministic and probabilistic HLF, it is critical to ensure that the forecasts are consistent across hierarchies, which is mainly achieved by post-hoc reconciliation. Methodologies range from simple scaling approaches like top-down (which disaggregates a top-level forecast) and bottom-up (which sums lower-level forecasts) to more advanced optimal combination methods like minimum trace (MinT) reconciliation (which finds a linear combination of the base forecasts that minimizes the trace of the reconciled forecast error covariance matrix) [27]. Each approach embodies a different position on the accuracy-coherence spectrum. There is a fundamental trade-off between accuracy and consistency to be considered during the evaluation of HLF. The bottom-up method preserves local accuracy but is often incoherent and can produce highly biased aggregate forecasts by accumulating errors from the lower levels. Optimal reconciliation methods attempt to find a statistically superior balance, but they still represent a compromise [29]. A comprehensive HLF evaluation framework must therefore move beyond a simple accuracy score. It should be a multi-objective system that can simultaneously quantify local and global inaccuracies (ideally using scale-free metrics to allow fair comparison across levels) while also measuring the degree of incoherence. Without such a holistic framework that explicitly evaluates this trade-off, it is impossible to determine whether a given hierarchical forecasting model is genuinely effective or simply hiding poor performance through error cancellation and biased reconciliation. While research has proposed scale-independent error metrics to compare accuracy across levels [30], a unified framework that holistically evaluates this trade-off is still lacking.

Multi-datasets: Currently, the majority of forecast evaluation metrics are limited to single datasets, and the evaluation strategy on multiple datasets is seldom considered [31]. Practically, a regional power system operator usually receives forecasts from various subordinate regions and needs a way to evaluate the forecasting performance of each subregion. This evaluation is crucial for incentivizing subregions to submit more accurate forecasts. Different time series contain varying proportions of unpredictability and exhibit different levels of volatility, which means that the forecasting difficulty across different datasets also varies significantly. To this end, using the traditional metrics to evaluate the forecasts across different datasets is not reasonable, where the predictability of the dataset should be considered.

As shown in Fig. 4, one possible solution is to consider a forecasting skill score that normalizes accuracy metrics S by using a lower bound S_l and an upper bound S_u to balance the forecasting difficulty across different datasets:

$$\widetilde{S} = \frac{S - S_l}{S_u - S_l} \tag{12}$$

where S_l represents the accuracy of a benchmark model, while S_u represents the accuracy of a "perfect" model with access to future information. Due to the presence of unpredictable uncertainty, even a "perfect" model cannot achieve an accuracy of 100%. For instance, a "perfect" model has been developed using multiple-seasonal time series decomposition, where only the residual is forecasted while the other components are assumed to be perfectly accurate [32].

Even though the above skill score demonstrates the consideration of predictability, it still has intrinsic shortcomings. The reference upper and lower performance bounds are calculated based on a series of assumptions and specific models. The irrationality of assumptions and the inevitable model uncertainty may generate inappropriate performance references that lead to unreasonable evaluation scaling. Further investigation is still needed to fairly quantify dataset predictability and propose a robust evaluation framework to rank load forecasts across multiple datasets.

V. How much do forecasts cost?

The goal of load forecasting is to help system operators achieve optimal downstream decision-making. Therefore, the evaluation of load forecasts should not only measure the accuracy but also quantify the cost-saving capability, which requires the consideration of decision-making costs and

清華大學出版社 Tsinghua University Press | SciOpen

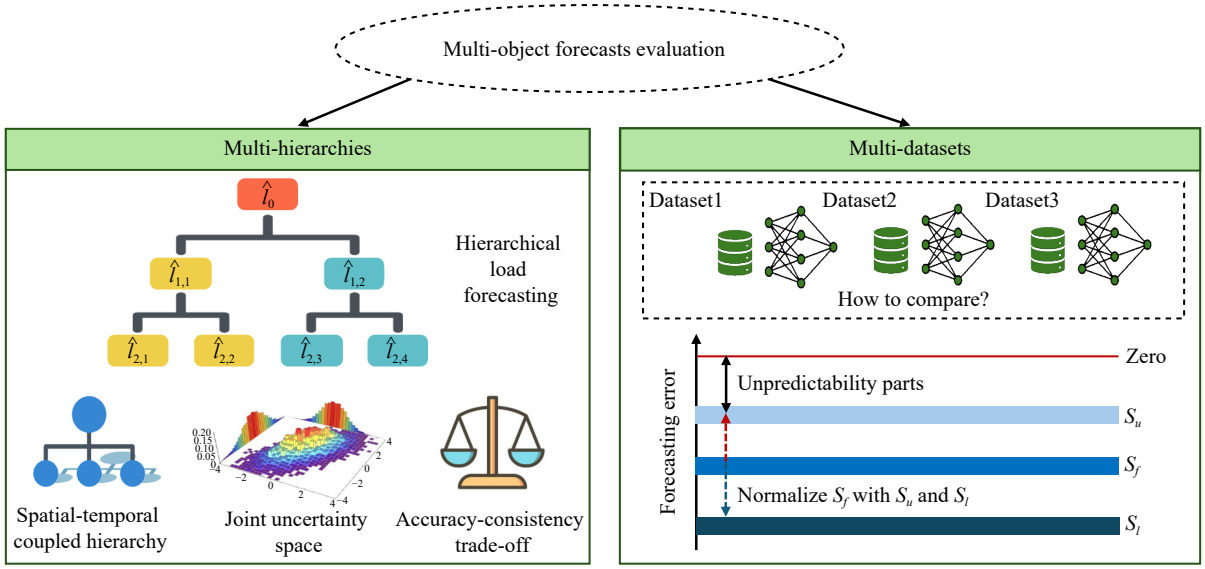

Fig. 4. Multi-object forecast evaluation from the perspectives of multi-level hierarchies and multi-dataset settings. The left panel highlights key issues that must be considered when forecasting loads within hierarchical structures. The right panel illustrates a potential solution for comparing forecast performance across datasets with varying degrees of predictability.

return-on-investment for estimating forecasts.

Decision-making costs: The assessment of decision-making costs in power system operations requires careful consideration of how load forecasting impacts downstream decisions. Load forecasting serves as the foundation for economic dispatch, reliability management, and other operational tasks. Although traditional statistical metrics impose symmetric penalties on positive and negative forecast errors, these errors induce asymmetrical operational consequences as quantified in Fig. 5. Consequently, minimizing conventional error metrics (e.g., MAE and RMSE) does not guarantee minimized decision costs. This necessitates augmenting forecast evaluation with decision cost quantification. Within load forecasting, such cost-driven approaches are gaining significant traction and are termed value-oriented learning [33], cost-oriented learning [34], [35], decision-focused learning [36], [37], or end-to-end modeling [38], yet unified by their core principle: embedding downstream decision processes directly into forecast assessment and model training. Mathematically, cost-oriented load forecasting can be expressed as Eq. (13). The goal is to identify forecasting parameters and decision variables that simultaneously minimize decision costs.

$$\min_{\theta,z} \sum_{t=1}^{T} C(z, \eta, f(X|\theta), l) \tag{13}$$

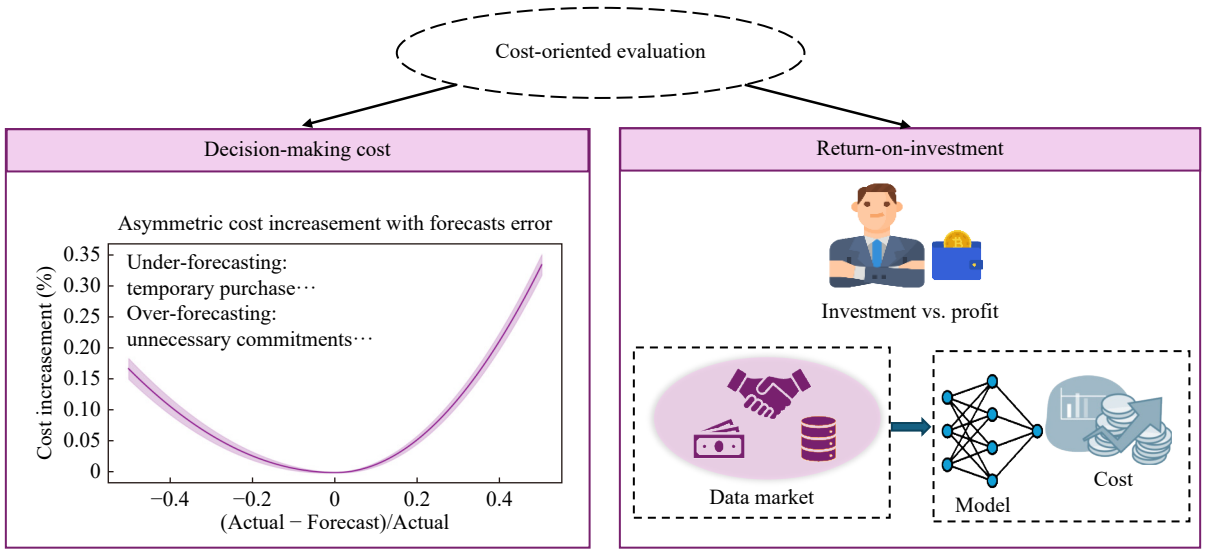

Fig. 5. Cost-oriented forecasts evaluation. The figure on the left indicates that errors of equal magnitude, but opposite sign, can have asymmetric impacts on operational costs. The figure on the right illustrates the necessity of balancing the investment required to improve forecasting accuracy against the profits.

where θ is the parameters of the load forecasting model f, η denotes the parameters in decision-making problems, C is the objective function, and z is the decision variable.

Although there are some attempts to bridge various load forecasting with downstream decision-making, there are many interesting topics that deserve to be discussed about cost-oriented evaluation. To begin with, load forecasting encompasses a wide range of methodologies, implying that there are various forms of f in Eq. (13). Deterministic forecasting provides single-point estimates, while probabilistic forecasting generates forecasts in the form of probability distributions. As previously mentioned, there are also various approaches to conducting probabilistic forecasting. On the decision-making side, optimization frameworks are equally diverse (C in Eq. (13)), including stochastic optimization, robust optimization, and distributionally robust optimization. Despite this variety, most existing research tends to focus on a single forecasting approach and primarily considers stochastic optimization for downstream decision-making. *This highlights the need for a unified framework that can integrate multiple forecasting and decision-making modules.* It is noted that cost-driven approaches are effectively suitable for probabilistic forecasting, as the full predictive distribution (rather than a single point forecast) allows for a more direct and nuanced calculation of expected operational costs under uncertainty. In addition, it can also address the heavy-tailed and multimodal characteristics of probabilistic forecasts discussed above.

However, simply integrating various forecasting and decision-making methods is not sufficient to address the complexities of real-world applications. In practice, the relationship between forecasting and decision-making is dynamic and interactive, rather than one-way or static. When forecasts are used to support decisions, they can actively influence the outcomes they are intended to predict, a phenomenon known as performative prediction [39]. This implies that, in the context of Eq. (13), the actual values of the loads l may change after the forecasts $\hat{l} = f(X|\theta)$. For example, system-level day-ahead load forecasts affect electricity prices in both day-ahead and real-time markets. These price signals (η in Eq. (13)), in turn, prompt price-sensitive users to adjust their energy consumption, thus reshaping the system's load profile [36]. As a result, the realized outcomes may deviate from the original forecast targets, complicating the evaluation of forecasting models. Further complicating the matter is the collective behavior of market participants. When a large number of agents in the market adopt cost-oriented forecasting objectives, their aggregated actions can alter the market-clearing equilibrium [40], demonstrating that in two-stage electricity markets, a competitive regression equilibrium can emerge under such circumstances, highlighting the systemic impact of widespread cost-oriented forecasting. Hence, it is essential to fully consider the impact of forecasts on downstream decisions, as well as the potential feedback effects that these decisions may have on the forecasts themselves.

Cost-oriented evaluation also presents unique challenges in the context of multi-object forecasting and diverse downstream applications. Different loads contribute unequally to the system's total cost, causing cost-oriented evaluation metrics to focus more heavily on larger or more volatile loads and making it difficult to fairly assess forecasting performance for each object. Moreover, load forecasts are often used for a variety of downstream tasks (various C), such as economic dispatch or demand response, each requiring different evaluation criteria. Additionally, when downstream tasks shift over time (e.g., η dynamically changes), relying solely on historical data to train cost-oriented forecasting models can lead to severe overfitting. Even for a fixed task, the accuracy of cost-oriented evaluation depends on how well the downstream decision-making process is modeled. Any inaccuracies can introduce bias into the evaluation [41]. Therefore, while cost-oriented evaluation can help bridge forecasting and decision-making, it also introduces additional uncertainty and bias, making the already complex task of evaluating load forecasts even more challenging.

Return-on-investment: Nowadays, there is a consensus that obtaining more features related to energy consumption behaviors (such as meteorological information) often helps improve forecast accuracy. However, in the era of big data, where concerns surrounding data privacy and ownership are becoming more prevalent, acquiring such data comes at a cost [42]. For instance, system operators often need to purchase high-quality meteorological forecasts or other external data services to improve forecast accuracy. Recent research has explored various data-sharing mechanisms to further enhance load forecasting performance. For example, [43] designed a regression market in which participants are incentivized to share their proprietary data in exchange for monetary compensation. Reference [44] proposed a framework that allows distribution system operators to acquire traffic data from data holders, such as traffic authorities and navigation application companies, to improve the forecast of EV charging demand. Similarly, [38] introduced a cross-sector data sharing scheme for multi-energy systems. Therefore, while developing load forecasting models, we also need to consider the cost of obtaining data from data markets and the downstream decision-making benefits that such data can provide [45]. By evaluating the return-on-investment ratio, which means the cost of data acquisition against the potential decision-making benefits, we can more effectively optimize resource allocation and enhance overall economic efficiency.

VI. CONCLUSIONS

The evolution of load forecasting in modern power systems has revealed significant gaps in current evaluation frameworks. While forecasting methodologies have advanced considerably, their evaluation paradigms remain inadequate in addressing the multifaceted challenges and the evolving needs of grid operations. Our analysis demonstrates that a compre-

hensive evaluation framework must simultaneously address four critical dimensions: (1) the development of specialized metrics for deterministic forecasts that account for inherent characteristics of load curves and specific operational requirements; (2) the establishment of advanced probabilistic evaluation techniques capable of handling multimodal and heavy-tailed distributions; (3) the creation of hierarchical evaluation frameworks and fair cross-dataset comparisons considering predictability; and (4) the integration of decision-impact analysis to quantify cost savings in the short and long term, as well as returns on forecasting investments. These insights collectively highlight that load forecast evaluation must transition from isolated accuracy assessment to a holistic framework that bridges the gap between theoretical forecasting performance and practical grid operation requirements. In the future, we also call for a "next-generation" open evaluation platform or forecasting competition to benchmark forecasts on hierarchical coherence, cost-based decision impact, probabilistic calibration, and reproducibility/operational realism, boosting the development of the load forecasting field.

DECLARATION OF COMPETING INTEREST

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

FUNDING

This work is carried out without the support of any funding agency.

AI USAGE STATEMENT

During the preparation of this manuscript, the authors used AI-assisted tools for language polishing only. The authors have reviewed and edited the content and take full responsibility for the final version of the manuscript.

REFERENCES

[1] T. Hong and S. Fan, "Probabilistic electric load forecasting: a tutorial review," *Int. J. Forecast*., vol. 32, no. 3, pp. 914−938, Jul./Sep. 2016.

[2] C. X. Wang, P. Pinson, and Y. Wang, "Seamless and multi-resolution energy forecasting," *IEEE Trans. Smart Grid*, vol. 16, no. 1, pp. 383–395, Jan. 2025.

[3] L. Baur, K. Ditschuneit, M. Schambach, C. Kaymakci, T. Wollmann, and A. Sauer, "Explainability and interpretability in electric load forecasting using machine learning techniques–a review," *Energy AI*, vol. 16, pp. 100358, May 2024.

[4] D. L. Qin, X. Wu, D. Y. Sun, Z. F. Liang, and N. Zhang, "Load forecasting under distribution shift: an online quantile ensembling approach," *Appl. Energy*, vol. 401, pp. 126812, Dec. 2025.

[5] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecast*., vol. 30, no. 2, pp. 357−363, Apr./Jun. 2014.

[6] S. E. Razavi, A. Arefi, G. Ledwich, G. Nourbakhsh, D. B. Smith, and M. Minakshi, "From load to net energy forecasting: short-term residential forecasting for the blend of load and PV behind the meter," *IEEE Access*, vol. 8, pp. 224343–224353, Dec. 2020.

[7] A. Stratman, T. Q. Hong, M. Yi, and D. B. Zhao, "Net load forecasting with disaggregated behind-the-meter PV generation," *IEEE Trans. Ind. Appl*., vol. 59, no. 5, pp. 5341−5351, Sep./Oct. 2023.

[8] J. L. Zhao, J. H. Qu, H. R. Ji, J. Xu, H. M. Hasanien, R. A. Turky, and P. Li, "Robust operation of flexible distribution network with large-scale EV charging loads," *IEEE Trans. Transp. Electr.*, vol. 10, no. 1, pp. 2207–2219, Mar. 2024.

[9] D. Wang, W. F. Zheng, Z. Wang, Z. F. Wu, B. Q. Shen, and S. M. Tian, "Quantifying the potential of load flexibility for building HVAC system using model predictive control strategy," *Energy Build.*, vol. 323, pp. 114819, Nov. 2024.

[10] C. Wang, Y. Wang, Z. T. Ding, and K. F. Zhang, "Probabilistic multi-energy load forecasting for integrated energy system based on Bayesian transformer network," *IEEE Trans. Smart Grid*, vol. 15, no. 2, pp. 1495–1508, Mar. 2024.

[11] R. Z. Yang, Y. J. Li, C. X. Hu, and F. Y. Hou, "Exdiffusion: classifier-guidance diffusion model for extreme load scenario generation with extreme value theory," *IEEE Trans. Smart Grid*, vol. 16, no. 5, pp. 3887–3903, Sep. 2025.

[12] M. L. Kloubert and C. Rehtanz, "Probabilistic load flow method for correlated multimodally distributed input variables," in *Proceedings of 2018 Power Systems Computation Conference*, Dublin, Ireland, 2018, pp. 1–6.

[13] Z. X. Wang, Q. S. Wen, C. L. Zhang, L. Sun, and Y. Wang, "DiffLoad: Uncertainty quantification in electrical load forecasting with the diffusion model," *IEEE Trans. Power Syst.*, vol. 40, no. 2, pp. 1777–1789, Mar. 2025.

[14] D. A. G. Vieira, B. E. Silva, T. V. Menezes, and A. C. Lisboa, "Large scale spatial electric load forecasting framework based on spatial convolution," *Int. J. Electr. Power Energy Syst.*, vol. 117, pp. 105582, May 2020.

[15] G. Zhou, J. X. Shi, B. J. Chen, Z. Y. Qi, and L. C. Wang, "Risk assessment of power supply security considering optimal load shedding in extreme precipitation scenarios," *Energies*, vol. 16, no. 18, pp. 6660, Sep. 2023.

[16] J. J. M. Moreno, A. P. Pol, A. S. Abad, and B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, May. 2013.

[17] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecast*., vol. 32, no. 3, pp. 669−679, Jul./Sep. 2016.

[18] R. H. Gerlach, C. W. S. Chen, and N. Y. C. Chan, "Bayesian time-varying quantile forecasting for value-at-risk in financial markets," *J. Bus. Econ. Stat.*, vol. 29, no. 4, pp. 481–492, Jan. 2011.

[19] Y. Z. Zhou, G. X. Lin, D. Y. Sun, Z. F. Liang, and N. Zhang, "Interpretable load forecasting via representation learning of geo-distributed meteorological factors," *Int. J. Electr. Power Energy Syst.*, vol. 172, pp. 111283, Nov. 2025.

[20] S. Haben, J. Ward, D. V. Greetham, C. Singleton, and P. Grindrod, "A new error measure for forecasts of household-level, high resolution electrical energy consumption," *Int. J. Forecast*., vol. 30, no. 2, pp. 246−256, Apr./Jun. 2014.

[21] Z. Wang and T. Z. Hong, "Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN)," *Energy Build.*, vol. 224, pp. 110299, Oct. 2020.

[22] S. Y. Li, H. Xiong, and Y. Z. Chen, "DiffPLF: a conditional diffusion model for probabilistic forecasting of EV charging load," *Electr. Power Syst. Res.*, vol. 235, pp. 110723, Oct. 2024.

[23] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and A. Panagiotelis, "Forecast reconciliation: a review," *Int. J. Forecast*., vol. 40, no. 2, pp. 430−456, Apr./Jun. 2024.

[24] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," *Eur. J. Oper. Res.*, vol. 262, no. 1, pp. 60–74, Oct. 2017.

[25] S. S. Rangapuram, S. Kapoor, R. S. Nirwan, P. Mercado, T.

Januschowski, Y. Y. Wang, and M. Bohlke-Schneider, "Coherent probabilistic forecasting of temporal hierarchies," in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023, pp. 9362–9376.

[26] S. S. Rangapuram, L. D. Werner, K. Benidis, P. Mercado, J. Gasthaus, and T. Januschowski, "End-to-end learning of coherent probabilistic forecasts for hierarchical time series," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8832–8843.

[27] S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman, "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization," *J. Am. Stat. Assoc.*, vol. 114, no. 526, pp. 804–819, Oct. 2018.

[28] S. B. Taieb, J. W. Taylor, and R. J. Hyndman, "Coherent probabilistic forecasts for hierarchical time series," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3348–3357.

[29] R. J. Hyndman and G. Athanasopoulos, "Forecasting hierarchical or grouped time series," in *Forecasting: Principles and Practice*, 2nd ed. , R. J. Hyndman and G. Athanasopoulos, Eds. Melbourne: OTexts, 2018.

[30] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679−688, Oct./Dec. 2006.

[31] H. Li, M. Heleno, W. N. Zhang, K. Y. Sun, L. R. Garcia, and T. Z. Hong, "A cross-dimensional analysis of data-driven short-term load forecasting methods with large-scale smart meter data," *Energy Build.*, vol. 344, pp. 115909, Oct. 2025.

[32] S. Liu, D. L. Qin, and Y. Wang, "Rank load forecasting performances of multiple datasets," in *Proceedings of 2023 IEEE Sustainable Power and Energy Conference*, 2023, pp. 1–6.

[33] Y. F. Zhang, M. S. Jia, H. L. Wen, Y. X. Bian, and Y. Y. Shi, "Toward value-oriented renewable energy forecasting: an iterative learning approach," *IEEE Trans. Smart Grid*, vol. 16, no. 2, pp. 1962–1974, Mar. 2025.

[34] C. F. Zhao, C. Wan, and Y. H. Song, "Cost-oriented prediction intervals: on bridging the gap between forecasting and decision," *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 3048–3062, Jul. 2022.

[35] S. Barja-Martinez, F. Teng, A. Junyent-Ferré, and M. Aragüés-Peñalba, "Personalized federated learning with cost-oriented load forecasting for home energy management systems," *IEEE Trans. Ind. Appl.*, vol. 61, no. 1, pp. 1410−1419, Jan./Feb. 2025.

[36] P. Ellinas, V. Kekatos, and G. Tsaousoglou, "Decision-focused learning under decision dependent uncertainty for power systems with price-responsive demand," *Electr. Power Syst. Res.*, vol. 235, pp. 110665, Oct. 2024.

[37] Y. Z. Zhou, R. Y. Yao, D. L. Qin, Y. X. Jia, and Y. Wang, "Privacy-preserving decision-focused learning for multi-energy systems," arXiv: 2510.21858, 2025.

[38] Y. Z. Zhou, Q. S. Wen, J. Song, X. Y. Cui, and Y. Wang, "Load data valuation in multi-energy systems: An end-to-end approach," *IEEE Trans. Smart Grid*, vol. 15, no. 5, pp. 4564–4575, Sep. 2024.

[39] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 7599–7609.

[40] V. Dvorkin, "Regression equilibrium in electricity markets," *IEEE Trans. Energy Mark. Policy Regul.*, vol. 3, no. 1, pp. 121–132, Mar. 2025.

[41] H. P. Zhang, R. Li, Q. T. Du, J. Y. Tao, S. Pineda, G. Kariniotakis, S. Camal, C. B. Monroc, M. Y. Sun, C. Wan, W. K. Xu, and F. Teng, "Decision-focused learning for power system decision-making under uncertainty," *IEEE Trans. Power Syst.*, vol. 41, no. 1, pp. 307–323, Jan. 2026.

[42] J. X. Wang, F. Gao, Y. Z. Zhou, Q. L. Guo, C. W. Tan, J. Song, and Y. Wang, "Data sharing in energy systems," *Adv. Appl. Energy*, vol. 10, pp. 100132, Jun. 2023.

[43] P. Pinson, L. Y. Han, and J. Kazempour, "Regression markets and application to energy forecasting," *TOP*, vol. 30, no. 3, pp. 533–573, Oct. 2022.

[44] S. Lv, S. Chen, T. F. Zhang, C. Chen, J. J. Xu, and Z. N. Wei, "Promote data sharing in integrated power-traffic networks: a coalition game approach," *IEEE Trans. Sustain. Energy*, vol. 16, no. 3, pp. 2171–2181, Jul. 2025.

[45] D. L. Qin, P. Pinson, and Y. Wang, "Load forecasting model trading: a cost-oriented and auction-based approach," *IEEE Trans. Smart Grid*, vol. 17, no. 1, pp. 766–779, Jan. 2026.



Yangze Zhou (Graduate Student Member, IEEE) received the B.S. degree from the College of Energy Engineering at Zhejiang University in 2023. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Hong Kong. His research includes data analytics and optimization in energy systems.



Dalin Qin (Graduate Student Member, IEEE) received the B.S. degree in Electrical Engineering and its Automation from South China University of Technology, Guangzhou, China, in 2022. He is now pursuing a Ph.D. degree in Electrical and Computer Engineering at the University of Hong Kong. His current research interests include energy forecasting and privacy-preserving data analytics in smart grids.



Yi Wang (Senior Member, IEEE) received the B.S. degree from Huazhong University of Science and Technology in June 2014, and the Ph.D. degree from Tsinghua University in January 2019. He was a visiting student with the University of Washington from March 2017 to April 2018. He served as a Postdoctoral Researcher in the Power Systems Laboratory, at ETH Zurich from February 2019 to August 2021. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of Hong Kong. His research interests include data analytics in smart grids, energy forecasting, multi-energy systems, Internet-of-things, and cyber-physical-social energy systems.