

Power Indicator Benchmarking Evaluation System Based on Multi-Source Data Fusion

Xiaomeng Liu^{1,a*}, Jianming Zhang^{1,b}, Xiaoxia Kou^{1,c}, Zhenguo Peng^{2,d} and Tingting Liu^{2,e}

¹State Grid Gansu Electric Power Company, 730000, Gansu, China

²Gansu Tongxing Intelligent Technology Development Co., LTD., 730050, Gansu, China

Traditional power indicator benchmarking systems are unable to meet the multi-dimensional dynamic evaluation needs of new power systems. In this paper, a dynamic benchmarking evaluation system is constructed based on multi-source data fusion. By establishing a multi-dimensional indicator library of “source-grid-load-storage-carbon”, integrating data from several domains such as meteorology, geography, and policy, an integrated methodological framework comprising dynamic weighting, comprehensive evaluation, and trend prediction is designed to achieve adaptive assessment and the analysis of future system status. Empirical results show that this system can improve the coverage of traditional evaluation methods for dimensions such as clean and low-carbon energy from 5 points to over 75 points, and effectively support risk warning and proactive dispatch by means of ongoing prediction of key indicators such as the rate of utilization of renewable energy (RMSE=0.18%).

Keywords: Data Fusion; Power System Evaluation; LSTM Prediction; Dynamic Benchmarking Framework; Renewable Energy Utilization

1. INTRODUCTION

With the increasing development and construction of new power systems, multiple factors have made the structure and operation of power systems increasingly complex. These factors include: the high proportion of renewable energy grid connection, multi-dimensional interaction between power generation, grid, load, and storage, and carbon emission reduction targets[1–3]. Traditional power indicator benchmarking systems rely mainly on historical operating data and static weights. However, their evaluation is often limited to issues of safety and economy, making it difficult to comprehensively depict the system’s overall performance in terms of multiple objectives such as clean and low-carbon development, flexible regulation, and high-quality service. Furthermore, they cannot dynamically respond to the impact of time-varying factors such as external policies and climate.

This results in untimely and one-sided evaluation results that cannot effectively support precise decision-making and proactive optimization.

This paper focuses on the benchmarking evaluation of the comprehensive operation level of provincial and above-scale power grids. Power system benchmarking evaluation is an important tool for measuring the quality of power grid development, guiding optimal resource allocation, and promoting lean management. Given the characteristics of multi-objective collaboration, strong uncertainty, and high dynamic interaction in new power systems, the construction of an intelligent evaluation system that can integrate multi-source information, dynamically adjust weights, and combine current status diagnosis and trend prediction capabilities has become the key to improving the level of power grid modernization governance. It has important theoretical research value and addresses the urgency of practical application [4–6]. Recently, advanced artificial intelligence technologies, particularly generative AI, have also been introduced to construct intelligent decision-making systems for power

*Corresponding Author. ^aliuxm3030@163.com, ^bzhangjianm030@163.com, ^ckouxianx090@163.com, ^dpengzg1284@163.com, ^eliutingt4119@163.com

grid dispatching, further driving the modernization of grid operations [7].

For the problem of multidimensional dynamic evaluation, existing research on power index evaluation focuses mainly on index construction, weight determination, and comprehensive evaluation. Regarding index construction, some studies have introduced external data to expand the evaluation dimensions [8]. In terms of weight determination, objective weighting methods such as entropy weighting and the Critic method are widely used to reduce subjectivity [9–10]. For comprehensive evaluation, models such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) and fuzzy comprehensive evaluation are commonly-used tools [11–12]. Different approaches for the development of index prediction are being researched [13–14], including time series models and machine learning techniques. However, current methodologies divide these tasks into piecemeal steps, whereas the optimal solution is a single, continuously flowing methodology combining multiple data sources using dynamic monitoring, benchmarking, and trend prediction. Furthermore, the majority of the methodologies employ weights for each of the indicators based on a static assumption rather than using adaptive methods to manipulate weights based on the environment. In addition, when performing a comprehensive evaluation, the vast majority of methodologies ignore the correlation among all of the indicators, thereby compromising the soundness of their rankings. In addition, when using the methodology for prediction, the various combinations of time-series characteristics do not include sufficient multi-source characteristics for predictive accuracy or predictive foresight.

To address the aforementioned limitations, a dynamic benchmarking and evaluation system for power indicators is constructed; it integrates multi-source data fusion, a dynamic indicator system, adaptive weighting, precise comprehensive benchmarking, and forward-looking trend prediction. To achieve this goal, firstly, a dynamic indicator library covering five dimensions—source, grid, load, storage, and carbon—is created by integrating internal operational data, external environmental data, and unstructured data. Secondly, at the model level, an improved Critic-Entropy Weight Combination Method [15] is proposed, which uses external environmental factors as dynamic adjustment coefficients to achieve adaptive adjustment of indicator weights; an improved TOPSIS model based on Mahalanobis distance is used for comprehensive ranking to overcome the influence of indicator correlation [16–17]; an LSTM (Long Short-Term Memory) network is used for rolling prediction of key indicators. By systematically combining multi-source data governance, dynamic indicator management, intelligent model library and decision application layer, a closed-loop logical chain of “data-driven evaluation, evaluation-guided decision-making, and decision feedback optimization” is formed.

Based on the aforementioned methodology, in regard to theoretical methods, this study achieves the deep integration and value mining of multi-source heterogeneous data in power benchmarking evaluation. In terms of technical application, it proposes an intelligent benchmarking tool

that can dynamically respond to environmental changes, accurately identify shortcomings, and effectively warn of risks. This new system enhances not only the scientific rigor, timeliness and guidance of evaluation results, but also provides complete, real-time data-driven decision making from “insight” on the current state to “forecast” on likely future results for the development, operation, scheduling, trading/marketing and performance management of all aspects of new electrical power systems. Therefore, this system has both significant theoretical innovation and value in engineering.

2. CONSTRUCTION OF MULTI-SOURCE DATA FUSION EVALUATION SYSTEM

2.1 Overall System Architecture and Multi-Source Data Governance

In response to the varied, evolving, and precise assessment needs associated with the development of new-age power generation systems, a logical architecture based upon a construction of “data-based layered decoupled intelligent decision making” is developed. This architecture includes three basic components: Data Layer, Model Layer, and Application Layer. The effective co-operation of the inter-layers is attained by creating a standard method for defining the control and data flow from input raw data to decision-making (the closed-loop).

The data layer is the underlying structure for the entire system. It is the point at which a system’s data from many separate sources, or “heterogeneous” sources, can be collected and managed. All raw inputs to the system come through the data layer as either one of three main source groups (the types of sources are discussed below). The first group of source data that comes through the data layer includes all internal operational data, most of which comes from Energy Management Systems (EMS), Production Management Systems (PMS), and Advanced Metering Infrastructure (AMI)—time series operational data and event data related to all aspects of energy generation, transmission, distribution and consumption. The second group of data includes all external environmental data, gathered in real-time using either via APIs (Application Programming Interfaces) or via files—with weather forecasts from any number of meteorological offices; power grid topology and resource location data from GIS systems; and renewable resource (such as wind and solar farms) and carbon market quota information from various, often governmental, databases. The third and final type of source data received through the data layer includes all unstructured types of data related to the energy industry—including dispatch logs, inspection reports, and policy documents. From the collected data, unified data preprocessing creates a structured data stream that is then sent to the model layer for further processing. Such real-time integration technologies for large-scale heterogeneous data provide a solid and reliable foundation for subsequent intelligent evaluation and analysis [18].

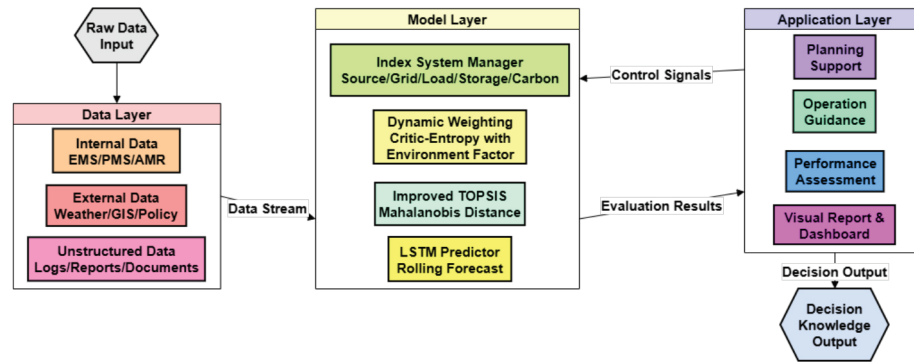


Figure 1 Three-layer architecture of the dynamic benchmark system.

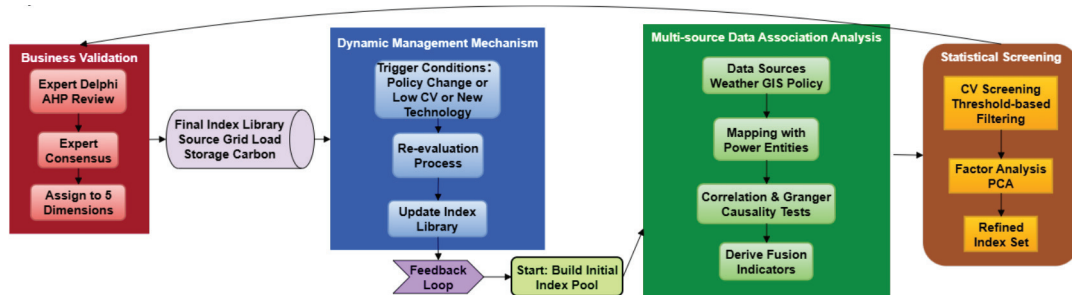


Figure 2 Framework diagram for generating a dynamic multidimensional indicator system.

The model layer contains the engine of this system. It comprises four distinct modules: the first module is the indicator management module; this module contains the dynamically updated “indicator library” created by the indicator system manager using a five-dimensional framework of source, grid, load, storage, and carbon, to create a dynamic library of indicators that are updated as new data becomes available. Next is the dynamic weighting module. This module uses an improved Critic-Entropy scheme to create dynamic weights for each indicator in the indicator library; however, a major enhancement of this module is that the outdoor conditions, such as season and climate and any policies that could affect an indicator or all indicators, are considered when establishing the weights for each indicator in the indicator library. Therefore, as outdoor conditions change (climate change or seasonal change), the weights of the indicators also change. Next, the comprehensive benchmark module integrates improvements to the model developed using TOPSIS and Mahalanobis distance. Based on the correlations between indicators in the indicator library, it comprehensively ranks the indicators. Essentially, the comprehensive benchmark module ranks indicators based on their correlation with all other indicators in the library. Finally, the prediction module uses historical time-series data to generate forward-looking or predictive decisions. The prediction module uses the historical time-series data of each indicator to perform rolling predictions of the future values of each indicator.

The evaluation results produced by the model layer of the intelligent evaluation system can be translated into meaningful action support for real-world business scenarios at the application layer. Among other things, its functions include providing data for long-term planning, providing

optimization recommendations for day-to-day operations, and evaluating and analyzing performance. Ultimately, all results of the analysis are displayed graphically as reports and dashboards, thus constituting the final decision knowledge output and ultimately fulfilling the transformation of data into tangible value.

The clearly established data flow and control flow between the respective layers ensures the integrity and closed-loop nature of the intelligent evaluation system as a whole. The data layer supplies the model layer with high quality input; the model layer supplies the application layer with intelligent analysis; the application layer uses feedback mechanisms to continuously improve the data and model layers. Collectively, these three elements of the intelligent evaluation system provide a self-evolving, intelligently responsive and dynamic evaluation system.

2.2 Generation of Dynamic Multi-Dimensional Indicator System

In this paper, a five-dimensional indicator system generation process (as shown in Figure 2) is created that is “data-driven, business-oriented, and is subject to continuous evolution” to solve the evaluation difficulties of a new power system with “multiple objectives”, “high dynamism”, and “very strong coupling”. The process enables the conversion of raw data into a measurable and comparative “source-grid-load-storage-carbon” indicator library that continuously evolves to indicate both the present state of the system and the potential future impact of the various systems on the environment.

To truly understand the power business from a macroeconomic perspective, the first step is to establish a meaningful index of all variables related to power business operations. This is achieved through a series of three key phases: building an initial index pool; multi-source data correlation analysis; and finally, generating a comprehensive set of indicators from the multi-source data correlation analysis. The building of the initial index pool should be done by acquiring an intimate knowledge and understanding of how the electric power business works and the many different aspects of it including the planning and operation of the power system and associated market. The first part of the initial index pool is built by obtaining comprehensive collections of those variables that are commonly used as indicators by the industry from sources such as industry standards and internal corporate performance indicators.

The second phase of producing an integrated indicator set through the association analysis should be seen as a key component of the production of data-driven operations. By determining the relationship between external environmental data and an entity in the electric power business such as a power plant, a multi-source data association analysis yields unique integrated indicators that cannot be represented by traditional internal business data.

The method for identifying correlation includes statistical procedures such as the Granger causality test and correlation analysis, both of which allow one to analyze quantitatively how external factors relate to internal operating indicators (load curve, line loss rate, voltage qualification rate). As a step toward building new integrated indicator systems, very strong correlations are identified between meteorological data, the locations of new energy power plants and the “theoretical power generation potential”, leading to the derivation of the “climate adaptability for new energy generation”. The integration of historical load data, temperature and holiday data leads to additional metrics such as “holiday baseline load” and “temperature-sensitive load”. The extensive pathways of complex data correlations identified through this process lead to the creation of new innovative metrics in the future. Once the first round of candidates has been established, they are filtered further to ensure that they are both scientifically valid and efficient to use within a well-rounded format.

Initial Screening: The coefficient of variation (CV) is used as the criterion for the initial screening of candidates for indicator variables [19–20]. The coefficient of variation of candidate variables has been calculated for many evaluation objects (time periods); based on this data, cutoffs are applied to variables that can adequately separate two or more different state objects by their coefficient of variation, and variables larger than the cutoff are selected for evaluation.

Dimensionality Reduction: For the set of indices that pass through the first selection process, several comprehensive components are generated from PCA or Factor Analysis to create additional information to assist in understanding and utilizing the original set of variables. This creation of additional components has eliminated redundancy [21–22] and has increased the effectiveness and explanatory strength of the indices system and through this process, producing a

structured/concise version of the index known as the “Refined Index Set”.

The approved index set must undergo a validation process by testing the business model via practical application in order to provide evidence of its operational capacity.

Expert review: Several rounds of ranking/scoring and discussion are conducted by experts in the field using the Delphi method or the Analytic Hierarchy Process (AHP) [23–25] to review the physical meaning, data availability, computational complexity, and business significance of the indicators.

Classification and Consensus: The indicators to be included in the library are agreed upon by the consensus of experts and are assigned accurately to the appropriate categories within the five dimensions of source, grid, load, storage and carbon, thereby creating the final “Final Index Library”.

A mechanism for dynamic management, so that updates to the indicator library can be performed in a closed-loop, is established to preserve the sustainability of the system over time.

Triggering Events: If any significant changes occur in the external environment (e.g. adjustments to national policies or large-scale implementation of new technologies), the dynamic management system triggers a review whenever the long-term coefficient of variation of an indicator approaches zero.

Review and Update: The “Re-evaluation Process” is initiated, re-executing multi-source data correlation analysis, statistical screening, and business verification, ultimately completing the “Update Index Library.”

Feedback Loop: The updated indicator library is fed back into the entire process, forming a continuously self-optimizing “Feedback Loop,” ensuring that the evaluation system remains synchronized with the latest technologies, policies, and operational needs.

2.3 Adaptive Intelligent Evaluation Model Design

To achieve the integrated evaluation goal of “assessable current status, diagnosable shortcomings, and predictable trends” for the new power system, this paper designs an adaptive intelligent evaluation model at the model level. This model consists of three core modules connected in series: dynamic weighting, comprehensive ranking, and trend prediction, forming a complete technical chain of “adaptive adjustment of indicator weights—multi-dimensional precise benchmarking of objects—forward-looking prediction of key indicators.” Its overall architecture is shown in Figure 3.

The model starts with the Input Layer, receiving three types of key information from the Data Layer: (1) standardized multi-dimensional index values of “source-grid-load-storage-carbon”; (2) external environmental factors (such as quarterly codes, regional average temperature deviation, carbon price index, etc.) used to drive dynamic weight adjustment; (3) historical multi-dimensional index sequences used for time-series modeling. These inputs collectively provide data support for subsequent modules.

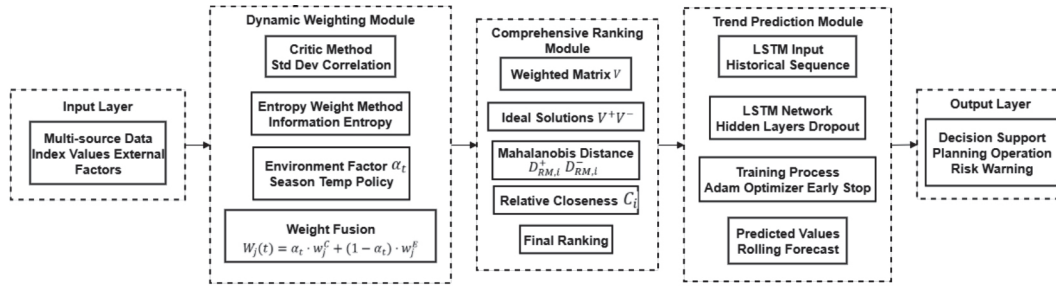


Figure 3 Structure of the adaptive intelligent evaluation model.

The Dynamic Weighting Module aims to overcome the limitations of traditional static weights, enabling the evaluation system to respond to changes in the external environment. This integration approach provides an enhanced way to use both energy and vector information by creating two types (energy and vector) of objective weights and assigning each a dynamic adjustment factor or coefficient to allow for adaptive fusion.

First, based on the standardized indicator data matrix, the weights w_j^C (critic method) and w_j^E (entropy weight method) are calculated respectively. The critic method weight calculation involves the standard deviation of the indicators (representing the intensity of comparison) and the correlation coefficient between indicators (representing conflict). The entropy weight method weight calculation is based on the information entropy e_j of the indicators; the smaller the entropy value, the greater the amount of information provided by the indicator, and the higher the weight. Second, a dynamic adjustment coefficient α_t is constructed. This coefficient is generated by a set of normalized external environmental factors (such as quarterly feature codes, regional average temperature deviation, and policy intensity index) through a preset linear or nonlinear mapping function, ensuring that $\alpha_t \in [0, 1]$. Finally, within the evaluation period t , the dynamic combination weight $W_j(t)$ of the j -th indicator is calculated by the following formula:

$$W_j(t) = \alpha_t \cdot w_j^C + (1 - \alpha_t) \cdot w_j^E \quad (1)$$

The standardized matrix X is weighted using $W_j(t)$ to obtain the weighted normalized matrix $V = (v_{ij})$. Then, the positive ideal solution $V^+ = (v_1^+, \dots, v_n^+)$ and the negative ideal solution $V^- = (v_1^-, \dots, v_n^-)$ are determined, where $v_j^+ = \max_i v_{ij}$ and $v_j^- = \min_i v_{ij}$.

The key improvement lies in the distance metric: calculating the regularized Mahalanobis distance from each object to the ideal solution:

$$\begin{aligned} D_{RM,i}^+ &= \sqrt{(V_i - V^+)(\Sigma + \lambda I)^{-1}(V_i - V^+)^+}, D_{RM,i}^- \\ &= \sqrt{(V_i - V^-)(\Sigma + \lambda I)^{-1}(V_i - V^-)^+} \end{aligned} \quad (2)$$

$\lambda = 10^{-4}$. Preliminary experiments have verified that it can effectively balance numerical stability and distance fidelity. I is the identity matrix of the same dimension as Σ . Finally, the relative proximity is calculated:

$$C_i = \frac{D_{RM,i}^-}{D_{RM,i}^+ + D_{RM,i}^-} \quad (3)$$

Arranged in descending order of C_i , the comprehensive sorting is completed.

The trend prediction module uses a long short-term memory (LSTM) architecture to generate rolling forecasts of important indicators (e.g., the new energy utilization ratio) to assist with forward-looking decision-making. The model receives historical multidimensional time series data as input (e.g., load data, weather data, output data, etc.) and then passes the data through a sliding window to create training instances. The network architecture consists of multiple stacked layers of LSTMs and dropout layers and minimizes the mean squared error loss function using the Adam optimizer. To reduce the chances of overfitting and enhance generalizability, early stopping is employed throughout training. For each forecast period, the predicted output consists of the forecast indicator value and confidence interval, which serves as guidelines for resource scheduling and risk warnings.

The outputs of the model result in three types of decision support tools: (1) comprehensive benchmarking report (for planning and assessment) (2) weakness list (for optimizing operations) (3) trend predictions and early warning (for control of proactive actions).

3. CASES AND EVALUATION DESIGN

3.1 Case Selection and Data Preparation

This study compares the effectiveness and universality of the evaluation system developed in this research by studying two provincial power grids: Province A (which represents the renewable resource-rich mountainous regions of China) and Province B (which represents eastern China, as strong load centers). The rationale for using these two examples is that they represent common examples in building a new power system across China. Province A has an installed capacity of more than 50% for renewable energy sources and has a predominance of mountainous terrain along with outstanding wind and solar resources with large variation in their outputs, so its operation is significantly influenced by weather conditions. Province B, conversely, is located within a region with a relatively mature grid structure, strong demand-side response resources and has implemented large amounts of flexible resources (e.g., gas and energy storage) and thus can serve as a good example for testing the comprehensive judgement ability of the evaluation system for

Table 1 Detailed table of multi-source data for empirical research.

Data Category	Data Type / Example	Source	Update Frequency	Primary Usage
Internal Operational	Power output, load, line loss, voltage data	EMS, PMS, AMI systems	15-min (real-time)	Core input for “source-grid-load” evaluation
	Event logs, outage records	Dispatch & Operations and Maintenance logs	Daily	Support reliability and safety indicators
External Environmental	Temperature, solar irradiance, wind speed	National Meteorological Center	Hourly / Fore-cast	Derive climate-sensitive indicators
	GIS: terrain, land use, resource distribution	National Geospatial Information Platform	Annual / As updated	Compute spatial risk and capacity indices
	Carbon price, renewable quota, policy texts	National Carbon Market, Energy Policy Database	Daily / Monthly	Construct “carbon”-related metrics and policy proxy
Derived Features	Theoretical renewable generation potential	Combined meteorological + plant location data	15-min	Input for “climate adaptability” indicator
	Temperature-sensitive load profile	Load + weather + holiday data	Daily	Enhance demand-side responsiveness evaluation
	GIS-based transmission corridor risk index	GIS + fault history + terrain data	Monthly	Reflect grid vulnerability in complex terrain
	Carbon-adjusted generation cost	Generation cost + carbon price	Monthly	Integrate economic and carbon dimensions

complex collaborative “source-grid-load-storage” scenarios. The two provinces present a stark contrast in terms of energy structure, power grid development stage, and external environmental pressures. Furthermore, their key operation and management data are highly accessible, meeting the basic requirements of empirical research.

The empirical data spans from January 1, 2019 to December 31, 2023, covering the key stages of policy transformation at the beginning of the 14th Five-Year Plan period, the leapfrog development of new energy, and the launch of the carbon market. Spatial granularity is based on provincial power grids as the basic unit, with some external data refined to the prefecture-level city level, and spatially matched to key nodes of the power grid.

This study integrates three types of multi-source heterogeneous data, the specific composition and sources of which are shown in Table 1. All data come from authentic and verifiable official or industry channels:

Internal operational data primarily originates from publicly available annual dispatch operation reports of the State Grid Corporation of China and China Southern Power Grid, anonymized operation records provided by provincial power dispatch control centers, and annual power reliability index reports published by the National Energy Administration. Specifically, this includes 15-minute time-series data such as power generation output, load curves, line loss rate, and voltage qualification rate, as well as event logs for equipment outages and fault trips [26–27].

External environmental data comes from authoritative public platforms: meteorological data originates from the National Meteorological Information Center of the China Meteorological Administration, including its managed China Meteorological Administration Data Broadcast

System (CMACast) and historical meteorological database; geospatial data is obtained from the Ministry of Natural Resources’ Geographic Information Public Service Platform (“Tianditu”) and the Resource and Environmental Science Data Center of the Chinese Academy of Sciences; carbon market data is derived from the daily transaction prices of national carbon emission trading published by the Shanghai Environment and Energy Exchange (SEEE); the renewable energy quota and policy texts are based on the policy document library published on the official websites of the National Development and Reform Commission and the National Energy Administration [28–30].

Derivative feature data is generated from the aforementioned basic data through feature engineering. For example, “Climate Adaptability of New Energy Power Generation” is calculated based on station-level meteorological matching and output data, while “Carbon-Adjusted Power Generation Cost” is synthesized by combining coal consumption of thermal power units and SEEE carbon price data, ensuring that all indicators have physical meaning and reproducibility.

Before entering the model, all raw data strictly follow the preprocessing procedure in Section 2.1. Time series data is uniformly calibrated to the UTC+8 time zone; frequency is aligned to 15-minute granularity; missing values are filled using cubic spline interpolation or a similar daily pattern-based filling strategy. Spatial data achieves accurate mapping from meteorological stations to power grid substations through WGS84 (World Geodetic System 1984) coordinate system matching and ordinary kriging. Finally, the indicator values are normalized using Min-Max (positive/negative indicators) or Z-score standardization (moderate indicators) according to the indicator attributes to eliminate the influence of dimensions.

Table 2 Core model parameter configuration and evaluation index definition table.

Module / Component	Parameter / Metric	Value / Definition
LSTM Prediction Model	Input sequence length (time steps)	96 (corresponding to 24 hours at 15-min granularity)
	Forecast horizon	4 steps (1-hour ahead) / 96 steps (24-hour ahead) for rolling prediction
Improved TOPSIS	Hidden layers	2
	Hidden units per layer	128
	Dropout rate	0.2
	Optimizer	Adam
	Learning rate	0.001
	Loss function	MSE
	Training epochs	200 (with early stopping patience = 15)
Critic-Entropy Weighting	Distance metric	Mahalanobis distance (using inverse covariance matrix of weighted indicators)
	Regularization coefficient (λ)	1e-4
	Ideal solutions	Positive: max for benefit-type; min for cost-type Negative: opposite
Evaluation Metrics	Entropy calculation	$e_j = -\frac{1}{\ln m} \sum_{i=1}^m p_{ij} \ln p_{ij}, p_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} \quad (4)$
	Critic weight	$w_j^C \sigma_j \sum_{k=1}^n (1-r_{jk}) \quad (5)$
	Dynamic blending coefficient (α_t)	Linear combination of normalized external factors: season, temperature deviation, policy index
	RMSE (Root Mean Squared Error)	$\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (6)$
	MAE (Mean Absolute Error)	$\frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i \quad (7)$
	Expert Consistency Score (ECS)	Average pairwise Kendall's W across =5 domain experts (scale: 0–1, higher = more consensus)

3.2 Experimental Setup and Evaluation Metrics

In order to have dependable model performance validation and to be able to accurately interpret evaluation results, a defined experimental environment has been created, and an extensive method of evaluation has been put in place. All of the experiments performed for this study are done on a consistent hardware and software platform, which also have many replications to account for any randomness that may have affected the results. The mean of each metric represents the reliability and reproducibility of the results.

The experimental environment is developed on a high-performance computing platform that is considered to be the standard for the current technology. The configuration of the hardware is a dual Intel Xeon Gold 6248R processor server with 256GB of RAM and two 32GB NVIDIA Tesla V100 GPU processors, which fits the needs for computational power to train the deep learning models. On the software side, this environment runs Ubuntu 20.04; all core algorithms are written in Python 3.8; the deep learning framework is PyTorch 1.12.1; it includes data manipulation tools and traditional algorithms, such as the Critic and Entropy Weight methods using libraries like Scikit-learn, Pandas, and NumPy. By following this setup, a stable environment is created to act as a strong and efficient computing environment while enabling the tight integration of the various modules.

The hyperparameters of the core model are not arbitrarily selected, but optimized based on extensive pre-experiments and the experience of experts in the power system field. For example, the number of hidden layers and units in the LSTM network need to be balanced between model complexity and the risk of overfitting; the calculation of the covariance matrix of Mahalanobis distance in TOPSIS needs to consider data dimensionality and sample size; the mapping function form (linear/nonlinear) of the adjustment coefficient α_t in dynamic weighting is also tested comparatively. The specific values of these key decisions are detailed in Table 2.

Note: LSTM parameters are a general configuration; the monthly forecasting task in Section 4.5 uses the same network structure, but the input sequence is 12 months of historical monthly data.

The evaluation system is designed to comprehensively examine the entire evaluation system from two dimensions: business rationality and forecast accuracy.

Business Rationality: The “Expert Consistency Score” (ECS) is used to measure whether the overall ranking results conform to the actual understanding of domain experts. This process invites no fewer than five senior experts from power grid dispatching, planning, and marketing departments to independently score the provincial rankings output by the model. Finally, Kendall's W coefficient is used to calculate the level of consensus among experts [31–32].

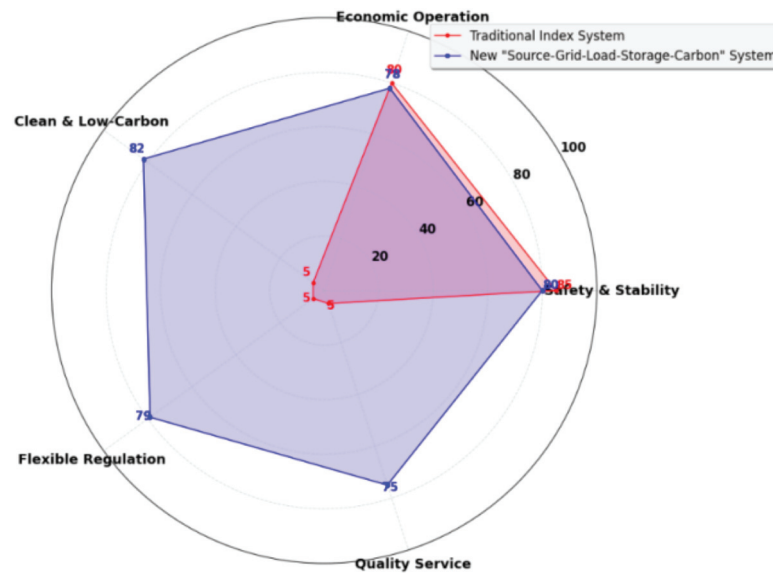


Figure 4 Radar chart comparison of dimensions between the old and new indicator systems.

The second aspect of prediction accuracy refers to the two standard forms of metrics associated with trend prediction — RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) - which allow researchers to quantify the model's skill at predicting key future indicators (e.g. utilization rate of renewable energy) [33]. Lower values for these metrics indicate a model's ability to more accurately identify historical measurements and predict future outcomes, thus providing robust data support for developing ahead of time (advance) strategies for management function. Every experiment is performed using the fixed environment outlined above, with the average score from each of the five trials used to determine the reported result for each evaluation metric. By repeating this process, the robustness of the experiments is enhanced, providing a robust technical framework to guide future researchers in replicating the experiments and building more reliable results through multiple iterations.

4. RESULTS DISCUSSION

4.1 Completeness of The Multi-Dimensional Indicator System

This paper quantitatively assesses how well the new metrics system captures each of the three dimensional areas of coverage by comparing it to the traditional metrics system, which contains only two metrics areas of coverage (Safety/Stability, Economic Operation), and the total number of the two systems of coverage (2 vs. 12 metrics). Both systems (traditional metrics and new metrics) are compared relative to both systems. Traditional metrics only represent 12 of the total potential number of available performance metrics, while the new metrics, built using the "Source, Grid, Load, Storage, and Carbon" Framework, represent three core areas of performance metrics (clean and low carbon, flexible adjustment and quality service). By introducing

27 indicators such as "new energy penetration rate" and "equivalent available hours of energy storage," it constructs a more comprehensive and three-dimensional evaluation capability.

Figure 4 uses the five evaluation dimensions as coordinate axes, plotting the standardized scores of the traditional system (red outline) and the new system (blue outline) on each dimension.

Figure 4 clearly shows that the evaluation outline (red) of the traditional indicator system only excels in the dimensions of 'Safety and Stability' (85 points) and 'Economic Operation' (80 points), while its scores in the dimensions of 'Clean and Low-Carbon,' 'Flexible Adjustment,' and 'High-Quality Service' are all close to zero (5 points). This reveals a significant blind spot in the evaluation perspective of the traditional system, failing to effectively respond to the core demands of the new power system in terms of green transformation and enhanced flexibility, and failing to effectively capture the newly established "source-grid-load-storage-carbon" system of the new power system.

In contrast, the evaluation outline of the new indicator system covers all five dimensions, forming a full and approximately pentagonal closed graph. Especially in the "clean and low-carbon" dimension, thanks to the inclusion of indicators such as "new energy utilization rate" and "carbon emission intensity," the score reaches as high as 82 points. Meanwhile, in the "flexible adjustment" dimension, through the introduction of indicators such as "ramp speed adequacy," the newly built system also demonstrates an excellent performance of 79 points—an evaluation capability completely lacking in the traditional system. Experimental data strongly demonstrates that the newly built system not only achieves full coverage of dimensions in breadth but also achieves dynamic balance of evaluation focus through the synergistic effect of multi-dimensional indicators in depth, thus enabling a more comprehensive and accurate characterization of the overall development level of the new power system.

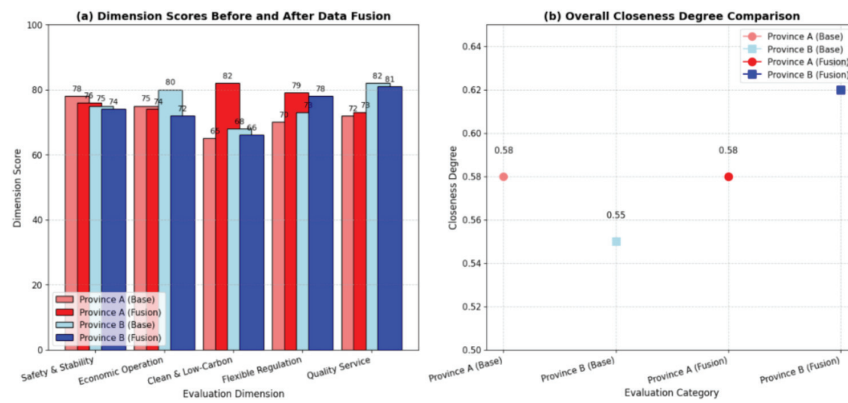


Figure 5 Comparison of scores and overall similarity of the power grids of the two provinces before and after multi-source data fusion. (a) Dimensional scores before and after data fusion, (b) Overall proximity comparison.

4.2 Impact of Multi-Source Data Fusion on Evaluation Results

To quantitatively reveal the impact of external environmental data and multi-source information fusion on the evaluation results, this study uses an indicator system containing only internal operational data (basic scheme) and a complete indicator system integrating multi-source data such as meteorological, geographical, and policy data (fusion scheme) as inputs. An improved TOPSIS model is used to comprehensively evaluate the power grids of Province A and Province B. The scores and overall ranking of each province in the five core dimensions under the two schemes are calculated, as shown in Figure 5.

As can be clearly observed from Figure 5(a), multi-source data fusion has a structural impact on the evaluation scores of each province across different dimensions. Under the basic scheme, Province A performs well in the dimensions of “safety and stability” (78 points) and “economic operation” (75 points), while Province B has an advantage in the dimension of “high-quality service” (82 points). However, after the introduction of multi-source data, the evaluation pattern has changed significantly.

For Province A, its score in the “clean and low-carbon” dimension has achieved a qualitative leap, increasing significantly from 65 points in the basic scheme to 82 points in the fused scheme. Meanwhile, its score in the “flexible adjustment” dimension also improves from 70 points to 79 points, further enhancing its system flexibility. However, its score in the “safety and stability” dimension decreases slightly to 76 points.

For Province B, its score in the “Economic Operation” dimension decreases from 80 points in the basic scheme to 72 points in the integrated scheme, and its score in the “Clean and Low-Carbon” dimension also slightly decreases from 68 points to 66 points. Besides, its score in the “Flexible Adjustment” dimension increases from 73 points to 78 points, reflecting a more accurate assessment of the adjustment value of its large-scale demand-side response resources after incorporating meteorologically sensitive load forecast data.

These changes in dimension scores are ultimately reflected in the overall proximity score shown in Figure 5(b). Under the basic scheme, Province A’s overall proximity score is

0.58, slightly higher than Province B’s 0.55. However, under the integrated scheme, Province B’s overall proximity score jumps to 0.62, surpassing Province A’s 0.58 and becoming the new leader. This reversal in rankings profoundly reveals the limitations of traditional evaluation systems.

4.3 Effectiveness of Dynamic Weighting Mechanism

To verify the effectiveness of the dynamic weighting mechanism and demonstrate its adaptability to evaluation decisions, this study calculates the dynamic weights of each core evaluation dimension over five consecutive annual evaluation periods from 2019 to 2023 based on the improved Critic-Entropy Weight Combination Model. Figure 6 visually presents this trend as a stacked area diagram, revealing how the evaluation orientation dynamically adjusts over time and with changes in the external environment.

In Figure 6, the weight of the “Safety and Stability” dimension shows a trend of first decreasing and then stabilizing over the five years, dropping significantly from 35% in 2019 to 28% in 2021, and then stabilizing in the 26%–27% range between 2022 and 2023. This trend closely matches the shift in the focus of system development. The weight of the “Clean and Low-Carbon” dimension, however, shows a strong upward trend, continuously climbing from 18% in 2019 to 30% in 2023, and in 2023, it surpasses all other dimensions, becoming the highest-weighted evaluation indicator. Meanwhile, the weight of the “flexible adjustment” dimension steadily increases from 15% in 2019 to 22% in 2023. The weight of the “economic operation” dimension gradually declines from 25% in 2019 to 18% in 2023. The weight of the “high-quality service” dimension remains relatively stable, fluctuating slightly within the 5%–10% range, reflecting its solid position as a fundamental and baseline evaluation dimension.

4.4 Comprehensive Benchmarking and Shortcoming Diagnosis Results

Based on the adaptive intelligent evaluation model designed in Section 2.3, a complete indicator system integrating

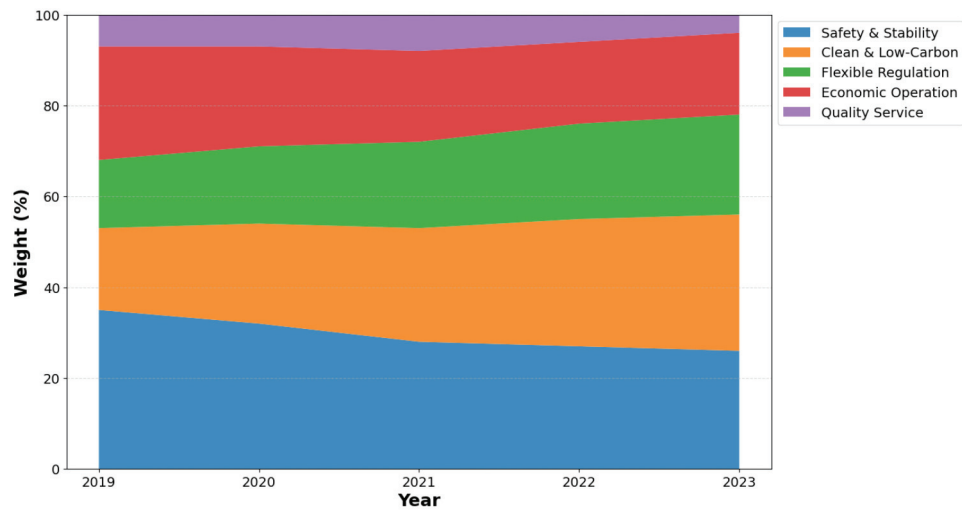


Figure 6 Annual trends in core evaluation dimension weights (2019–2023).

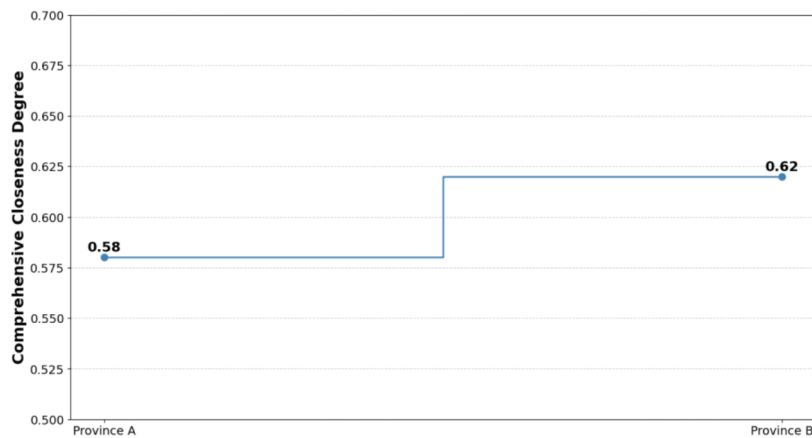


Figure 7 Annual comprehensive evaluation scores and rankings of the two provincial power grids based on the improved TOPSIS.

Table 3 Comparison of key secondary indicator scores.

Indicator	Unit / Description	Province A	Province B
Renewable Energy Utilization	%	93.5	95.8
ESS Equivalent Utilization Hours	Hours	1850	2100
Demand Response Matching Degree	%	82.4	88.1
Carbon Intensity per Unit Generation	tCO ₂ /MWh	0.412	0.368
Peak-shaving Resource Adequacy	% (available/required)	76.3	85.7
Ramping Capacity Sufficiency	% (available/required)	72.9	83.5
System Frequency Compliance Rate	%	99.96	99.94
Composite Line Loss Rate	%	4.8	5.2

multi-source data is adopted, and the dynamic weights for 2023 derived from the analysis in Section 4.3 are applied to conduct an annual comprehensive benchmarking of the power grids of Province A and Province B. The improved TOPSIS model is used to calculate the comprehensive closeness of each province to the ideal solution as the final score, and the ranking is determined based on the scores.

As shown in Figure 7, Province B's comprehensive score is 0.62, while Province A's is 0.58, placing Province B first in this year's comprehensive evaluation. This ranking result is consistent with the conclusions obtained using the fusion scheme in Section 4.2, verifying the consistency and stability

of the evaluation system under different data input scenarios.

To further analyze the root causes of the differences in comprehensive rankings, the scores of key secondary indicators for the two provinces under the five core dimensions are analyzed in detail. The specific comparative data is shown in Table 3.

Data analysis shows that Province A lags significantly behind Province B in several key indicators reflecting the core capabilities of the new power system. In the "clean and low-carbon" dimension, Province A's "carbon intensity per unit of power generation" is 0.412 tCO₂/MWh, higher than Province B's 0.368 tCO₂/MWh, revealing deficiencies in the

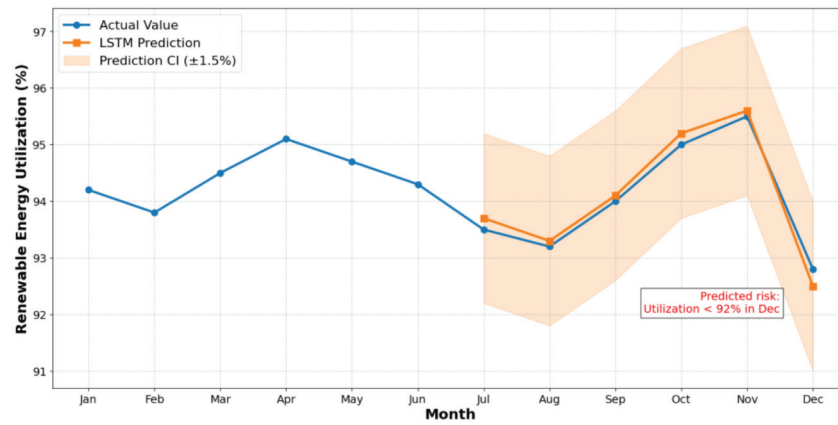


Figure 8 Comparison of actual and LSTM predicted values of new energy utilization rate.

synergistic optimization of its power structure regarding the proportion of clean energy and power generation efficiency. In the “flexible adjustment” dimension, Province A’s “peak-shaving resource sufficiency rate” (76.3%) and “ramp-up capacity sufficiency rate” (72.9%) are both lower than Province B’s corresponding levels (85.7% and 83.5%), indicating that the effectiveness of its physical reserves or dispatch strategies for adjusting resources in response to rapid fluctuations in renewable energy output has room for improvement. Furthermore, in terms of source-load coordination, Province A’s indicators such as “new energy utilization rate” (93.5%), “equivalent utilization hours of energy storage” (1850 hours), and “demand-side response matching degree” (82.4%) are also lower than Province B’s, reflecting its shortcomings in the integrated coordinated operation of source, grid, load, and storage.

It is worth noting that Province A performs relatively well in some traditional evaluation dimensions. For example, in the “safety and stability” dimension, its “system frequency qualification rate” is as high as 99.96%, better than Province B; in the “economic operation” dimension, its “comprehensive line loss rate” is 4.8%, also better than Province B’s 5.2%. However, these advantages have not been able to offset its significant gap in the two emerging core capabilities of “clean and low-carbon” and “flexible adjustment.”

4.5 Trend Forecasting Accuracy and Decision-Making Foresight

To meet the forward-looking requirements of monthly-scale indicators for medium- and long-term planning and scheduling decisions, this study configures the LSTM model in a monthly forecasting mode: the input is a series of monthly indicators from the past 12 months (generated by aggregating 15-minute-level raw data using an arithmetic mean); the forecast target is the average utilization rate of new energy sources in the first month of the future; the forecast is updated monthly using a rolling window approach. The model uses historical monthly data from January to June 2023 as the training set, performs rolling forecasts for monthly values from July to December, and compares these forecasts with actual observations to evaluate its forecasting

accuracy and decision-support capabilities.

Figure 8 presents a comparison of the actual observed values (solid blue line, January–December) and the predicted values (orange square line, July–December) of the “New Energy Utilization Rate” for the entire year of 2023, along with their corresponding confidence intervals (light orange shaded areas). The model’s predicted curve closely matches the actual value’s trajectory, successfully capturing the slight decline in utilization rate in July–August due to fluctuations in wind and solar resources, and the upward trend in utilization rate in October–November due to load growth and scheduling optimization.

The predictive performance is evaluated using the quantitative indicators defined earlier. On the test set (July–December 2023), the model’s root mean square error (RMSE) is 0.18%, and the mean absolute error (MAE) is 0.17%. These accuracy indicators demonstrate that the LSTM model can accurately capture the monthly variation patterns of this indicator and possesses reliable predictive capabilities.

The annotation in Figure 8 clearly indicates that the model forecast shows that the utilization rate of renewable energy in Province A in December has a high probability (based on the lower limit of the confidence interval) falling below 92%. Combined with the analysis of other data sources, this forecast risk stems mainly from historical winter weather patterns (lower wind speeds and less sunshine) and the high load pressure at the end of the year. This forecast information can be directly translated into specific decision support: dispatching departments can develop contingency plans in advance, appropriately increase the allocation of flexible resources (such as energy storage and gas turbine units) in December, or optimize inter-provincial power trading plans to offset the risk of declining renewable energy output and ensure that the utilization rate remains above the target level.

5. CONCLUSIONS

In this paper, a dynamic benchmarking evaluation system is constructed based on multi-source data fusion. By integrating multi-source data to establish a multi-dimensional index library of “source-grid-load-storage-carbon,” a modified critic-entropy weighting method is used for dynamic weighting.

A comprehensive ranking based on Mahalanobis distance is applied using TOPSIS, and trend prediction is performed using LSTM, forming a complete methodological chain. Empirical research shows that the traditional system scores only 5 points for the three dimensions of “clean and low-carbon”, “flexible adjustment”, and “high-quality service”, while the proposed system scores 82, 79, and 75 points in these three dimensions, respectively. Multi-source data fusion significantly changes the evaluation ranking, revealing the impact of the external environment; dynamic weighting enables weights to respond to policies and climate evolution, guiding the evolution from “safe economy” to “clean and low-carbon” and “flexible adjustment”; the diagnosis of shortcomings accurately identifies deficiencies in new energy utilization and energy storage call-up; the LSTM trend prediction model has a monthly root mean square error of 0.18% and an average absolute error of 0.17% for predicting new energy utilization rates, supporting proactive scheduling and risk warning. This system achieves collaborative innovation in terms of dimensions, data, mechanisms, and time scales, providing full-chain data-driven support for the planning and operation of new power systems, from current status analysis to future prediction, and therefore has both theoretical and practical value.

REFERENCES

- Iweh, C. D., Gyamfi, S., Tanyi, E., & Effah-Donyina, E. (2021). Distributed generation and renewable energy integration into the grid: Prerequisites, push factors, practical options, issues and merits. *Energies*, 14(17), 5375–5409.
- Mehedi, H., Wang, X., Ye, S., Xue, G., et al. (2023). Power generation expansion planning approach considering carbon emission constraints. *Global Energy Interconnection*, 6(2), 127–140.
- Tan, Q., Li, X., Fan, W., Wang, H., et al. (2023). Some key issues in building a “source network load storage” complementary energy internet in China. *Environmental Science and Pollution Research*, 30(35), 83513–83529.
- Chen, X., Tian, G., Huang, Y., Yang, Y., et al. (2023). New power system development path mechanism design. *Global Energy Interconnection*, 6(2), 166–174.
- Xialin, L., Yixin, L., Zhongguan, W., Lin, Z., et al. (2023). Design and teaching application of comprehensive experimental platform for new electricity power system. *Experimental Technology and Management*, 40(4), 192–199.
- Yan, H., Zhang, W., Kang, J., & Yuan, T. (2023). The necessity and feasibility of hydrogen storage for large-scale, long-term energy storage in the new power system in China. *Energies*, 16(13), 4837–4858.
- Yu Liu, Weihua He, Jianfei Zhu, Ran Yang and Xinrui Yang. (2026). Intelligent decision-making system for power grid dispatching based on generative AI. *Engineering Intelligent Systems*, 34(1).
- Han, T., Gao, Z., Du, W., & Hu, S. (2022). Multi-dimensional evaluation method for new power system. *Energy Reports*, 8(6), 618–635.
- Liu, Y., Wang, Y., Li, T., Ma, R., et al. (2022). Evaluation of new power system based on entropy weight-TOPSIS method. *Mathematical Problems in Engineering*, 2022(1), 1–10.
- Shi, H., Li, Y., Jiang, Z., & Zhang, J. (2021). Comprehensive power quality evaluation method of microgrid with dynamic weighting based on CRITIC. *Measurement and Control*, 54(5–6), 1097–1104.
- Cao, J., He, B., Qu, N., Zhang, J., et al. (2022). Benefits evaluation method of an integrated energy system based on a fuzzy comprehensive evaluation method. *Symmetry*, 15(1), 84–109.
- Li, M., Shan, R., Virguez, E., Patiño-Echeverri, D., et al. (2022). Energy storage reduces costs and emissions even without large penetration of renewable energy: The case of China Southern Power Grid. *Energy Policy*, 161(1), 1–39.
- Chen, Y., Bhutta, M. S., Abubakar, M., Xiao, D., et al. (2023). Evaluation of machine learning models for smart grid parameters: performance analysis of ARIMA and Bi-LSTM. *Sustainability*, 15(11), 8555–8580.
- Pierre, A. A., Akim, S. A., Semenyo, A. K., & Babiga, B. (2023). Peak electrical energy consumption prediction by ARIMA, LSTM, GRU, ARIMA-LSTM and ARIMA-GRU approaches. *Energies*, 16(12), 4739–4751.
- Li, G., Zhang, L., Wang, Y., & Kang, Z. (2023). Critical node identification method of power grid based on the improved entropy weight method. *Electronics*, 12(11), 2439–2455.
- Dong, H., Yang, K., & Bai, G. (2022). Evaluation of TPGU using entropy-improved TOPSIS-GRU method in China. *PloS one*, 17(1), 1–24.
- Mathebula, J., & Mbuli, N. (2025). Application of TOPSIS for Multi-Criteria Decision Analysis (MCDA) in Power Systems: A Systematic Literature Review. *Energies*, 18(13), 3478–3506.
- Jiawei Zhang. (2025). Real-time integration technology of large-scale heterogeneous data based on big data and artificial intelligence. *Engineering Intelligent Systems*, 33(2).
- Jalilibal, Z., Amiri, A., Castagliola, P., & Khoo, M. B. (2021). Monitoring the coefficient of variation: A literature review. *Computers & Industrial Engineering*, 161(1), 1–55.
- Yadav, S. K., Misra, S., & Gupta, S. (2024). Improved family of estimators of population coefficient of variation under simple random sampling. *Communications in Statistics-Theory and Methods*, 53(2), 727–747.
- Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *International journal of Innovative Science and Research Technology*, 8(5), 2028–2032.
- Zhao, J., Cui, S. & Xu, Z. Research on the closed-loop supply chain of intelligent products considering government subsidies in the context of the internet of things. *Discov Internet Things* 5, 34 (2025).
- Pant, S., Kumar, A., Ram, M., Klochkov, Y., et al. (2022). Consistency indices in analytic hierarchy process: a review. *Mathematics*, 10(8), 1206–1221.
- Sablitzky, T. (2022). The Delphi method. Hypothesis: Research Journal for Health Information Professionals, 34(1), 1–6.
- Tavana, M., Soltanifar, M., & Santos-Arteaga, F. J. (2023). Analytical hierarchy process: Revolution and evolution. *Annals of Operations Research*, 326(2), 879–907.
- Auffhammer, M., Wang, M., Xie, L., & Xu, J. (2021). Renewable electricity development in China: Policies, performance, and challenges. *Review of Environmental Economics and Policy*, 15(2), 323–339.
- Sospiro, P., Amarnath, L., Di Nardo, V., Talluri, G., et al. (2021). Smart grid in China, EU, and the US: state of implementation. *Energies*, 14(18), 5637–5652.
- Du, Y., Shi, H., Zhang, J., Xia, X., et al. (2022). Evaluation of MERRA-2 hourly surface solar radiation across China. *Solar Energy*, 234(1), 103–110.

29. Li, Y., Wang, R., Zhou, W., &Gao, B. (2022). Evaluation of dispatching results of power system with high penetration of renewable energy based on Pythagorean fuzzy set and TOPSIS. *Energy Reports*, 8(13), 524–532.
30. Zhou, K., Zhang, Z., Liu, L., &Yang, S. (2022). Energy storage resources management: Planning, operation, and business model. *Frontiers of Engineering Management*, 9(3), 373–391.
31. Franceschini, F., &Maisano, D. (2021). Aggregating multiple ordinal rankings in engineering design: the best model according to the Kendall's coefficient of concordance. *Research in Engineering Design*, 32(1), 91–103.
32. Xu, C., Li, W., Rao, Y., Qi, B., Yang, B., & Wang, Z. (2023). Coordinative energy efficiency improvement of buildings based on deep reinforcement learning. *Cyber-Physical Systems*, 9(3), 260–272.
33. Xiang, X., Li, X., Zhang, Y., &Hu, J. (2024). A short-term forecasting method for photovoltaic power generation based on the TCN-ECANet-GRU hybrid model. *Scientific Reports*, 14(1), 6744–6756.

