scientific reports

Check for updates

OPEN

Loss functions in deep residual networks for short-term load forecasting: a systematic analysis

Junchen Liu[1], Faisul Arif Ahmad[1]✉, Khairulmizam Samsudin[1], Fazirulhisyam Hashim[1] & Mohd Zainal Abidin Ab Kadir[2]

The dependable and effective operation of contemporary electricity systems depends on short-term load forecasting (STLF). Although Deep Residual Networks (DRNs) have proven to be highly predictive, little is known about how loss functions affect forecasting performance and optimization behavior. This study presents a systematic evaluation of various loss functions within the original DRN and Principal Component Analysis–Deep Residual Network (PCA–DRN) frameworks under a unified experimental setting. Traditional, robust, and task-specific Penalized loss functions are examined using real-world datasets with distinct climatic and load characteristics. The results show that the Charbonnier loss consistently achieves the best overall performance across all evaluation metrics under the original DRN framework. However, under the PCA–DRN framework, a performance divergence emerges: while the Charbonnier loss remains optimal for point-forecast error metrics, the Penalized loss demonstrates superior performance in squared-error-based and correlation-oriented metrics. Moreover, PCA–DRN consistently outperforms the original DRN, highlighting the effectiveness of dimensionality reduction in improving feature representation and generalization capability. The statistical significance of the reported improvements is further confirmed using bootstrap-based statistical analysis. These findings suggest that loss function selection should be jointly considered with data characteristics and feature representation, rather than treated as an isolated design choice, to achieve robust and accurate STLF.

Keywords Charbonnier loss, DRN, Loss function, PCA, STLF

Abbreviations

Adam	Adaptive moment estimation
AI	Artificial intelligence
ANN	Artificial neural network
CI	Confidence interval
CNN	Convolutional neural network
CNN–LSTM–MMA	CNN-LSTM with multi-modal attention
Conv1D	One-dimensional convolution
CRN	Convolutional residual network
DL	Deep learning
DNN	Deep neural network
DRN	Deep residual network
ERN	Ensemble residual network
FC	Fully connected
ISO-NE	New England Independent System Operator
LSTM	Long short-term memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MSE	Mean squared error
MW	Megawatt
MyPJ	Malaysia Petaling Jaya

[1]Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia. [2]Advanced Lightning, Power and Energy Research Centre (ALPER), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia. ✉email: faisul@upm.edu.my

NMSE	Normalized mean squared error
PCA–DRN	Principal component analysis–deep residual network
R	Correlation coefficient
R^2	Coefficient of determination
ReLU	Rectified linear unit
ResNet	Residual network
RMSE	Root mean squared error
RNN	Recurrent neural network
SD	Standard deviation
SELU	Scaled exponential linear unit
STLF	Short-term load forecasting
TCN	Temporal convolutional networks
Log-Cosh	Logarithm of the hyperbolic cosine loss

Modern power systems depend on accurate Short-Term Load Forecasting (STLF) for stable and effective operation[1]. Utilities may optimize important decision-making processes, such as generation scheduling, unit commitment, and energy trading strategies, by using STLF, which provides accurate demand estimates for forecasting timeframes ranging from 24 h to one week ahead[2]. Improved grid planning, operation, and management are directly impacted by increased forecasting accuracy, which strengthens supply reliability, reduces operating costs, and increases total energy efficiency[3].

Yet, the complexity and significance of STLF have significantly increased due to the continuous rise in energy demand and the growing variety of consumption patterns[4,5]. From an economic standpoint, even slight increases in predicting accuracy can result in significant cash gains. It has been observed that even a 1% increase in accuracy can result in yearly savings of several million dollars for large-scale electricity providers, mainly due to more effective resource allocation and lower reserve needs[6,7]. As a result, ongoing improvements in forecasting techniques, no matter how big, can result in substantial operational and financial benefits.

Generally, STLF approaches can be broadly categorized into earlier statistical and machine learning-based methods and more recent data-driven approaches. Early studies predominantly relied on statistical models and conventional machine learning techniques[8], which established the foundation for STLF research and provided valuable insights into electricity consumption patterns. However, these early approaches exhibit inherent limitations in modeling complex and highly nonlinear load dynamics. Their reliance on simplified assumptions about underlying consumption patterns restricts their adaptability to real-world variability. In addition, as the dimensionality of input features increases, their generalization capability tends to degrade, making them more susceptible to overfitting[9,10].

To address these limitations, artificial neural networks (ANNs) have been increasingly introduced, leading to growing interest in data-driven STLF solutions. By leveraging deep learning (DL) principles, ANNs demonstrate improved capability in capturing complex nonlinear relationships within load data[11]. Nevertheless, increasing the number of input variables, hidden units, or network layers may still lead to overfitting, thereby compromising model generalization. To mitigate this issue, various enhanced ANN architectures have been explored in the context of STLF[12–14].

Despite these developments, the capacity of early ANNs and their variants, which are typically characterized by relatively shallow structures, remains limited in capturing highly complex load patterns. Because of their enhanced representational capacity and capacity to learn hierarchical features, deep neural networks (DNNs) with several hidden layers have become more popular. Complex temporal dependencies and geographical correlations seen in a variety of data sources may be more effectively modeled thanks to these structures. As a result, STLF research has progressively moved away from conventional ANN-based models and toward more complex DL frameworks that provide more flexibility and predictive power[15].

In STLF, shallow model designs have been gradually replaced by deeper neural network frameworks with a variety of input characteristics. Convolutional Neural Networks (CNNs) have been widely employed among them because to their ability to capture localized temporal patterns through convolutional processes. Nevertheless, they are still not very good at simulating long-range temporal connections, and deeper networks frequently result in more challenging training[16,17]. As an extension of traditional CNNs, Temporal Convolutional Networks (TCNs) have been proposed to overcome these constraints. TCNs allow for simultaneous modeling of longer temporal sequences by using dilated and causal convolutions, which increase the receptive field. However, their performance is largely reliant on architectural design decisions, and more intricate or deep configurations may still be needed to correctly capture very long-term dependencies[18,19]. Recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks, use memory cells and gating methods to explicitly express temporal connections and solve the vanishing gradient problem. Despite these benefits, managing lengthy input sequences reduces computing efficiency due to their sequential processing structure[20,21]. Transformer-based architectures have attracted a lot of interest lately because of their self-attention mechanisms, which enable completely parallel modeling of long-range dependencies[22–24]. However, when extended to deeper configurations, these models may show unstable training behavior and are frequently linked to high computing costs.

In an attempt to increase forecasting accuracy and generalization capability, hybrid architectures that integrate convolutional, recurrent, and attention-based techniques have been studied more and more. Transformer–LSTM[25] and CNN-LSTM with multi-modal attention (CNN–LSTM–MMA)[26] are two notable examples that seek to take use of the complementing advantages of many modeling paradigms. These hybrid approaches have greatly increased predicted accuracy, but they often have problems with training stability and processing

efficiency. These shortcomings highlight the necessity of developing more dependable, efficient, and scalable DL frameworks for STLF.

In order to solve the vanishing gradient problem and enhance training stability in DNNs, residual learning—which is backed by identity shortcut connections—was initially implemented in the Residual Network (ResNet) architecture[27]. By introducing the Deep Residual Network (DRN), Chen et al.[28] expanded this idea to the field of STLF and showed that a well-balanced design in terms of network depth, optimization stability, and representational capability can achieve dependable and high-accuracy forecasting performance. Following this ground-breaking work, other DRN-based enhancements have been developed to further increase forecasting accuracy and model durability. These include data-driven DRN models using specific residual architectures designed for STLF tasks[29], ensemble-enhanced DRN frameworks using snapshot learning or residual ensemble algorithms[30,31], and adaptive feature selection approaches[32]. Furthermore, structural improvements have been introduced to refine residual representations, including the redesign of convolutional residual blocks for enhanced local feature extraction, as well as the incorporation of inception-inspired modules and CNN-based feature extraction mechanisms[33–35], while hybrid DRN architectures incorporating recurrent networks or attention mechanisms have been investigated to better capture temporal dependencies[36–38]. Additionally, by using dimensionality reduction approaches, the Principal Component Analysis–Deep Residual Network (PCA–DRN) has been suggested to efficiently include various meteorological variables and enhance feature representation efficiency[39].

Building upon these advancements, it can be observed that most existing studies primarily focus on improving model architectures, feature representations, or ensemble strategies to enhance forecasting performance. However, comparatively less attention has been paid to the role of the loss function, which serves as a fundamental component guiding the optimization process in DL models. In addition, real-world load data are inherently non-stationary and often subject to regime changes, seasonal transitions, and distribution shifts caused by evolving consumption patterns and weather variability. Under such conditions, the effectiveness of a forecasting model is not only determined by its architecture but also by how the loss function shapes optimization under changing data distributions. Despite this close connection, the systematic investigation of loss function design in DRN-based STLF, particularly in relation to non-stationary and distributional variations, remains limited.

In particular, different categories of loss functions exhibit distinct optimization characteristics. Traditional error-driven formulations focus on minimizing prediction discrepancies, while robust loss functions are designed to enhance stability under noisy and nonlinear conditions. Furthermore, task-specific loss formulations, such as the Penalized loss proposed in DRN-based STLF, introduce additional structural constraints to better capture global load characteristics. However, the relative effectiveness of these loss functions under different data conditions, as well as their interaction with feature representation strategies such as PCA, has not been comprehensively explored.

To address this gap, this study conducts a systematic and controlled evaluation of various loss functions within DRN and PCA-DRN frameworks for STLF. By maintaining consistent model architectures and training settings, the impact of different loss formulations on forecasting performance is isolated and rigorously analyzed. Experiments are conducted on two real-world datasets from distinct regions with different climatic conditions to examine the generalizability of the findings across diverse load patterns. Furthermore, a bootstrap-based statistical analysis is employed to assess the significance of the observed performance differences.

The following is a summary of this study's primary contributions. First, a comprehensive comparison of traditional, robust, and Penalized loss functions is conducted within a unified DRN-based STLF framework. Second, the interaction between loss functions and PCA-based feature representation is systematically analyzed to reveal their complementary effects. Third, the reliability of the results is validated through bootstrap-based statistical testing, providing strong evidence for the observed performance differences. Finally, this study offers practical insights into the selection of loss functions for DRN-based forecasting models under diverse real-world conditions.

The remainder of this paper is organized as follows. Section "Literature review" introduces the related work on DRN-based STLF and loss function design. Section "Methodology of research" describes the proposed methodology, including the DRN and PCA-DRN frameworks as well as the considered loss functions. Section "Experimental results and analysis" presents the experimental results and analysis. Finally, Section "Conclusion" concludes the paper and outlines directions for future research.

Literature review

The increasing complexity, variability, and uncertainty of electricity consumption patterns have made STLF a more challenging task. Recent advances in deep learning-based STLF have also been summarized in comprehensive surveys, highlighting emerging learning paradigms and the growing role of advanced data-driven models in energy forecasting tasks[40]. Traditional statistical and shallow learning methods often struggle to effectively capture nonlinear relationships and intricate temporal dependencies in load data. In this context, DL approaches, particularly DNNs, have gained widespread attention due to their superior capability in modeling complex nonlinear dynamics and temporal correlations. Modern DL-based forecasting methods are generally categorized into convolution-based, recurrent-based, attention-driven, and hybrid modeling frameworks, each emphasizing different characteristics of load time-series data. Despite these advancements, existing DL-based STLF models still face challenges in maintaining stable optimization and robust performance under highly nonlinear and noisy conditions.

Because CNNs are so good at explaining short-term correlations and extracting localized temporal patterns, they have been thoroughly researched in STLF. For instance, Li et al.[16] enhanced forecasting performance over a range of prediction horizons by transforming load sequences into image-like representations, which enabled convolutional kernels to use spatial correlations. However, the dual-branch design and necessary data

transformation processes added more structural complexity, which could make real-time implementation more difficult. To improve prediction accuracy and quantify uncertainty, Jurado et al.[17] created an encoder-decoder CNN system with Monte Carlo Dropout and probabilistic density estimation. This method performed better than traditional recurrent models, but it was less resilient to sudden changes in demand due to its decreased accuracy at high load times.

In order to better capture long-range temporal correlations, TCNs are an extension of traditional convolutional models that include dilated and causal convolutions. In order to account for cumulative weather impacts and changing feature relevance, Tang et al.[18] integrated both channel-wise and temporal attention processes into a TCN framework, improving predicting accuracy. However, adding attention modules made the model much more complicated. Liu et al.[19] suggested a temporal depthwise convolutional network based on TCN to address computational efficiency. This network uses depthwise separable convolutions to minimize parameter redundancy while preserving temporal modeling capabilities. The model is still mostly convolution-driven despite these advancements, which might restrict its flexibility when handling extremely complicated temporal dynamics.

Sequence modeling remains a significant challenge in STLF, particularly when handling complex seasonal dynamics. Models based on recurrent architectures, such LSTM variants, have demonstrated strong prediction power when sufficient historical data is provided[20]. However, prediction accuracy may actually decline if significant external variables—such as weather-related factors—are not expressly considered. Bento et al.[21] attempted to get around this constraint by employing metaheuristic-driven hyperparameter tuning, but the resulting increase in computational demand presents scalability challenges.

Transformer-based models and attention mechanisms, in contrast to recurrent architectures, have garnered a lot of interest lately due to their capacity to capture long-range temporal correlations through global self-attention. For example, Ran et al.[22] used Transformer networks with empirical mode decomposition to enhance temporal feature extraction. Nevertheless, the model's high training time and reliance on preset breakdown parameters restrict its ability to adjust to new datasets. By expanding the attention receptive area, Jiang et al.[23] enhanced prediction accuracy, albeit at the expense of much higher memory usage. TS2ARCformer, which combines contextual encoding, cross-dimensional attention, and autoregressive components into a single framework, was developed by Li et al.[24] to further this area of research. The inclusion of autoregressive processes and hierarchical attention structures significantly increased architectural complexity, making practical deployment in STLF applications challenging, even if this model scored better on test datasets.

In an effort to mitigate the inherent limitations of single-paradigm systems, recent research has focused more and more on hybrid DL frameworks that employ complimentary modeling techniques. Chen et al.[25] introduced a Transformer–LSTM hybrid architecture for industrial STLF that employed self-attention to capture global temporal interactions and LSTM layers to mimic sequential dependencies. Despite continuous performance gains, the cascaded structure necessitated careful hyperparameter adjustment and increased processing costs. Similar to this, Guo et al.[26] developed an enhanced CNN–LSTM framework with multi-modal attention to adaptively fuse a variety of inputs, including load demand and meteorological data. While increasing prediction accuracy, the attention-based fusion approach increased computational complexity and imposed more restrictions on data quality and availability.

In addition to improving model architectures and loss formulations, another important direction for enhancing forecasting performance lies in ensemble and aggregation-based approaches. These methods combine predictions from multiple models or experts to improve robustness and generalization under complex and uncertain conditions. For example, Incremona et al.[41] proposed an aggregation framework of nonlinearly enhanced experts, demonstrating that combining multiple predictors can effectively improve forecasting accuracy and stability in electricity load forecasting tasks. Such approaches highlight that performance gains can also be achieved through model combination strategies, rather than solely modifying the training objective or network architecture.

Deeper and more complex network designs may exacerbate training instability, gradient degradation, and performance saturation even when convolution-based, recurrent-based, attention-based, and hybrid DL models yield notable performance gains. These challenges highlight the need for learning paradigms that can maintain deep structures while maintaining consistent optimization behavior. DRNs have emerged as a feasible solution in this scenario by incorporating residual connections that enable scalable depth extension and gradient propagation.

In STLF applications, DRN-based models have demonstrated notable advantages in modeling intricate and highly nonlinear load patterns. Early research focused mostly on feasibility evaluation and comparative analysis with conventional DL techniques. Chen et al.[28] coupled residual learning with ensemble methods to increase robustness and generalization, although at a larger computational cost. Kondaiah et al.[29] proposed a task-oriented DRN architecture and showed that customized residual designs effectively predict nonlinear load patterns across many datasets. However, these studies mainly concentrated on aggregate forecasting accuracy and paid little attention to highly variable or high-frequency load dynamics.

To improve stability and generalization even further, ensemble-enhanced DRN frameworks were studied. Xu et al.[30] presented an Ensemble Residual Network (ERN) that leverages snapshot ensemble learning to increase robustness without training several distinct models. Chen et al.[31] developed this concept further by including multi-scale feature extraction and snapshot ensembles into a ResNet-based system. Although ensemble-based DRN approaches are more effective at generalization, their real-time application may be limited due to the fact that they often demand greater processing power, lengthier training schedules, and accurate snapshot interval design.

In the meantime, learning stability in DRN-based models has been enhanced by data-driven augmentation techniques. A Deep-ResNet architecture that emphasizes adaptive feature selection was developed by Kondaiah

et al.[32] to capture important load characteristics under various operating situations. Transferability across various power systems is hampered by susceptibility to shifts in data distribution, even with improvements in predicting consistency. Additionally, structural changes have been implemented to improve feature representation. While Sheng et al.[33] proposed a convolutional residual network (CRN) by redesigning convolutional ResBlocks to enhance local pattern extraction for high-resolution load data, Ding et al.[34] integrated inception-inspired modules into the feature extraction stage to enable multi-scale learning, and Liu et al.[35] introduced CNN modules into the feature extraction stage to improve feature representation capability. These improvements increase architectural complexity and complicate hyperparameter tuning, even though they enhance representational capacity.

To enhance temporal dependency modeling, several studies have integrated DRNs with attention-based and recurrent processes. While Li et al.[36] employed attention techniques to highlight significant time steps, Tian et al.[37] combined sequential modeling and deep residual feature extraction in a ResNetPlus–LSTM framework. Sheng et al.[38] has introduced a Residual LSTM Plus architecture that tightly connects ResBlocks with recurrent layers. The simultaneous stacking of residual, recurrent, and attention components significantly increases model depth and computational cost, limiting scalability in large-scale or real-time STLF systems, even if these hybrid frameworks usually enhance forecasting reliability.

In addition to architectural and ensemble-based enhancements, prior DRN-based studies have explored input-level expansions to provide improved meteorological data. Liu et al.[39] developed the PCA-DRN framework, which integrates several meteorological variables using PCA, to extend temperature-based DRN models. Although PCA-DRN effectively reduces input dimensionality and feature redundancy while preserving the residual learning structure, its dependence on linear dimensionality reduction inevitably weakens the physical interpretability of meteorological data. However, most existing DRN-based studies primarily focus on architectural design and feature engineering, while the role of loss functions in shaping optimization behavior and training stability remains largely underexplored. In recent years, increasing attention has been paid to forecasting models that explicitly address non-stationarity, regime shifts, and distributional changes in time-series data[42,43]. These approaches often incorporate adaptive learning strategies, robust optimization objectives, or distribution-aware mechanisms to improve generalization under evolving conditions. Since loss functions directly shape the optimization landscape and error sensitivity, they play a critical role in determining how models respond to such variations. However, the interaction between loss function design and non-stationary load characteristics remains insufficiently investigated, particularly within DRN-based STLF frameworks.

In DL models, the loss function plays a fundamental role in guiding the optimization process by quantifying the discrepancy between predicted and actual values[44]. Recent studies have also explored task-specific and geometry-aware loss formulations to improve model robustness and predictive performance under complex conditions, further highlighting the importance of loss function design in load forecasting models[45]. It directly determines the gradient signals used for parameter updates and therefore significantly influences model convergence behavior, training stability, and final predictive performance[46]. Moreover, the smoothness of the loss function and the resulting optimization landscape play a critical role in ensuring stable training dynamics and reliable convergence in DNNs. More importantly, the choice of loss function shapes the optimization landscape, affecting gradient smoothness, sensitivity to large errors, and the balance between bias and variance during training[47]. As a result, an appropriately designed loss function is particularly critical when dealing with complex, nonlinear, and noisy data distributions[48]. This is especially important for STLF tasks, where load data often exhibit non-Gaussian noise and extreme variations, making the selection of an appropriate loss function more critical for achieving stable and reliable model performance[49].

In the context of STLF, load data often exhibit high variability, nonlinear patterns, and strong dependence on external factors such as weather conditions[50]. These characteristics introduce significant noise and irregular fluctuations, which can distort gradient updates and lead to unstable training or suboptimal convergence[51]. In particular, the presence of outliers or extreme load variations may disproportionately influence the optimization process, depending on the loss formulation used[52]. Therefore, beyond architectural design and feature representation, the selection of an appropriate loss function becomes an essential factor in ensuring robust learning and reliable forecasting performance[53]. While substantial progress has been achieved through advanced network architectures and feature engineering, most existing studies primarily focus on structural design and input representation. In contrast, the role of loss functions in shaping optimization behavior remains comparatively underexplored, particularly within DRN-based STLF frameworks.

In recent studies, a diverse range of loss formulations has been explored in DL-based STLF, which can be broadly categorized into three groups: traditional error-based loss functions, robust loss functions, and Penalized loss function specifically designed for DRN-based STLF frameworks.

The first category includes widely used traditional loss functions such as Mean Absolute Percentage Error (MAPE)[54], Mean Absolute Error (MAE)[55], and Mean Squared Error (MSE)[56]. These loss functions are commonly adopted due to their simplicity and clear interpretability. MAPE directly measures relative forecasting error and is frequently used as a loss function in forecasting and time-series prediction tasks. MAE applies a linear penalty on absolute deviations and is widely used as an L1 loss in DNN regression problems, whereas MSE, also known as L2 loss, imposes a quadratic penalty and is one of the most commonly used loss functions for regression in DL models. However, these loss functions may exhibit sensitivity to outliers or unstable gradient behavior under highly nonlinear and noisy conditions.

To enhance robustness under noisy and nonlinear data conditions, recent studies in DL have increasingly adopted several advanced loss functions, including Huber loss[57], Logarithm of the hyperbolic cosine (Log-Cosh) loss[58], Pseudo-Huber loss[59], Smooth L1 loss[60], and Charbonnier loss[61]. These formulations are designed to balance sensitivity to small errors with robustness against outliers. For instance, Huber loss combines the advantages of MAE and MSE by using a quadratic penalty for small errors and a linear penalty for large errors. Log-Cosh loss provides a smooth and differentiable approximation to MAE, while Pseudo-Huber and Smooth

L1 losses further enhance numerical stability. Among these, Charbonnier loss, as a differentiable variant of the L1-norm, offers stable gradients and strong robustness, making it particularly suitable for DL optimization under complex, nonlinear, and noisy data distributions commonly encountered in STLF tasks.

In addition to these general-purpose loss functions, a Penalized loss function[28] originally introduced in the DRN framework is also considered as a key baseline. This loss is constructed based on MAPE by incorporating additional penalty terms to emphasize large forecasting errors. Such a design is particularly relevant for STLF applications, where large prediction deviations may lead to significant operational and economic consequences. By assigning higher weights to critical errors, the Penalized loss enhances the model's sensitivity to extreme load variations and improves forecasting reliability.

Based on the above analysis, this study systematically investigates the impact of the aforementioned categories of loss functions on the training stability and forecasting performance of DRN-based models, with a particular focus on traditional error-based, robust, and Penalized loss functions, aiming to provide more effective optimization strategies for STLF under complex and challenging data conditions.

Methodology of research

Research data

In real-world power system applications, issues such as incomplete records, missing observations, noise contamination, and heterogeneous data formats are common, and they can significantly degrade forecasting performance if not properly addressed[62]. Therefore, careful data preprocessing is essential to improve the robustness and reliability of STLF models[63]. This work uses two real-world datasets, the the New England Independent System Operator (ISO-NE) dataset and the Malaysia Petaling Jaya (MyPJ) dataset, to assess model performance under various operational situations and climatic regions.

The ISO-NE dataset represents a temperate climate and covers six states in the New England region of the United States, including Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. It contains hourly electricity demand and temperature data from March 2003 to December 2006. The dataset exhibits significant seasonal and interannual variability, reflecting the influence of heating demand in winter and cooling demand in summer. The load data correspond to system-level electricity consumption, while the temperature data represent observed hourly regional measurements rather than forecasted values. Due to its consistency, completeness, and standardized format, the ISO-NE dataset has been widely adopted as a benchmark in STLF research. Since it has already undergone data quality control by the provider, no additional data cleaning procedures were required in this study.

In contrast, the MyPJ dataset represents a tropical climate and provides a valuable case for analyzing electricity consumption patterns in such regions. It includes hourly load data obtained from Tenaga Nasional Berhad's Malaysian Grid System Operator, along with daily meteorological observations collected in Petaling Jaya by the Malaysian Meteorological Department from January 2020 to December 2022. The weather variables include rainfall, mean temperature, minimum temperature, maximum temperature, mean wind speed, wind direction, and maximum wind speed. Compared to temperate regions, tropical load patterns exhibit less pronounced seasonal variation, with short-term weather conditions playing a more dominant role in influencing electricity demand. While the load data are complete, the meteorological variables contain a small proportion of missing values (approximately 0.59%). To address this issue, linear interpolation is applied to impute missing values, as it preserves temporal continuity and minimizes distortion to the underlying data patterns, with negligible impact on forecasting performance.

As illustrated in Fig. 1, the two datasets exhibit distinct load characteristics under different climatic conditions. The ISO-NE dataset shows substantial variability, with load values typically ranging from 10,000 to 27,500 megawatts (MW), reflecting strong seasonal fluctuations. In contrast, the MyPJ dataset demonstrates relatively stable load patterns, with values generally ranging from 10,000 to 18,000 MW and limited seasonal variation. This difference highlights the contrasting influence of climate on electricity demand, where temperate regions are dominated by seasonal effects, whereas tropical regions are more influenced by short-term weather variations.

To prevent information leakage during model training and evaluation, data normalization is performed using statistical parameters derived exclusively from the training set, and the same scaling factors are subsequently applied to the test set. This preprocessing strategy ensures a fair and realistic assessment of model performance under practical operational conditions.

Research framework

This study employs two representative residual-based forecasting models, namely the original DRN[28] and PCA-DRN[39], to enable a systematic evaluation of their training behavior and forecasting performance. Rather than introducing additional architectural modifications, a consistent network topology is maintained throughout the experiments, allowing differences in performance to be examined under a unified structural framework. Although a fixed architecture is adopted to enable a controlled comparison, the selected datasets inherently exhibit different degrees of variability and temporal dynamics. This allows the analysis to indirectly reflect how different loss functions behave under varying levels of non-stationarity and distributional complexity. Such a design facilitates a controlled and objective comparison while ensuring that variations in results are not attributed to structural discrepancies. The architecture of the original DRN model applied to the ISO-NE dataset is shown in Fig. 2, while Figs. 3 and 4 depict the DRN and PCA-DRN architectures used for the MyPJ dataset, respectively. The PCA-DRN model is applied only to the MyPJ dataset, as it contains multiple meteorological variables that may introduce redundancy in the input representation, whereas the ISO-NE dataset includes only a single meteorological variable (temperature), making dimensionality reduction unnecessary. Therefore, the original DRN architecture is used without PCA-based feature processing for the ISO-NE dataset. Despite these

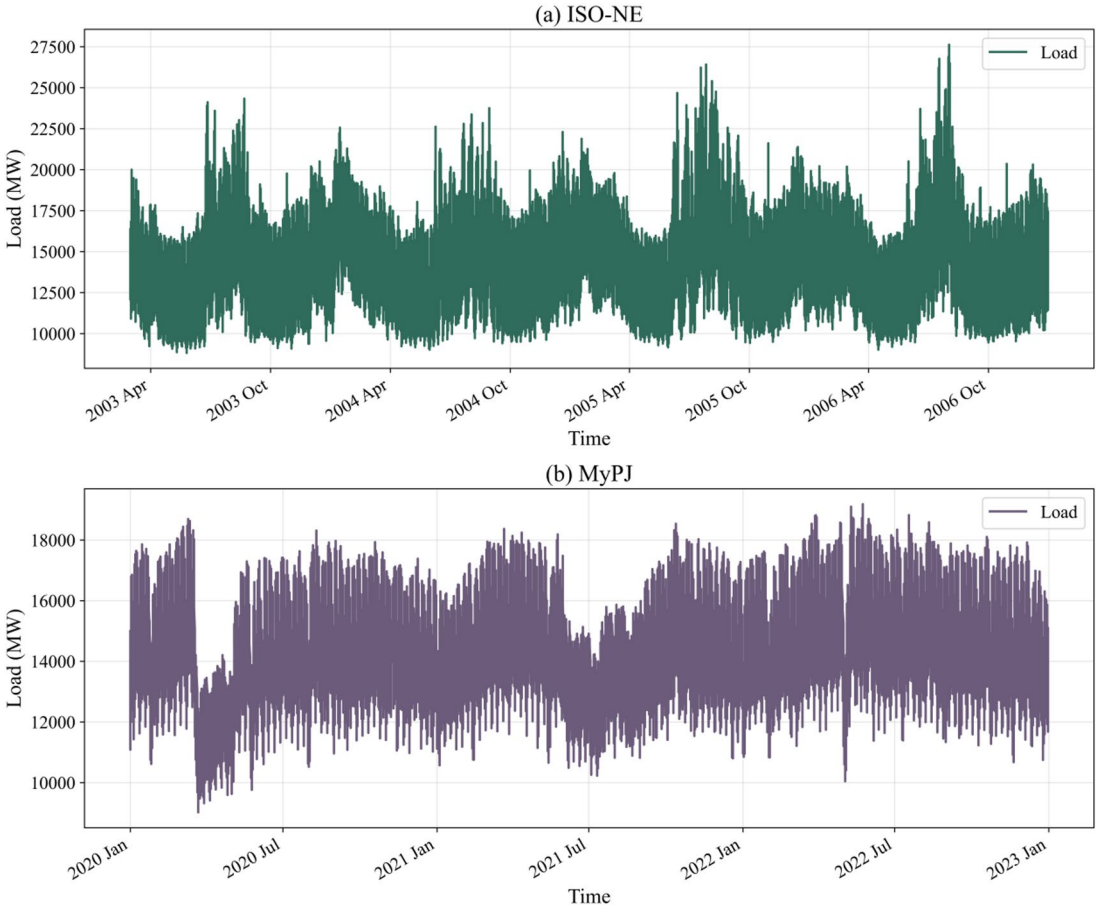

Fig. 1. Across the two datasets: load patterns observed in (**a**) ISO-NE; (**b**) MyPJ.

differences in input representation, both frameworks share the same overall model architecture, ensuring a fair comparison across different feature configurations.

The DRN and PCA-DRN architectures are composed of two principal components, namely a basic structure and a ResNetPlus module. Input features consist of historical load, time-related variables, and meteorological data, which are first processed by the basic structure to extract preliminary representations of electricity demand. These representations are subsequently refined through the ResNetPlus module—an enhanced residual learning framework derived from the conventional ResNet—thereby improving forecasting accuracy over the 24-hour horizon. While the original DRN demonstrates strong predictive performance using a limited set of meteorological inputs, real-world STLF scenarios are often influenced by multiple weather-related factors that may exhibit high correlation and redundancy. To address this limitation, PCA-DRN extends the baseline DRN by incorporating additional meteorological variables and applying PCA during data preprocessing and feature construction, effectively reducing dimensionality and eliminating redundant information while preserving the dual-component residual learning structure. For day-ahead forecasting, both models adopt a multi-submodel strategy, in which each sub-model is responsible for predicting the load of a specific future hour. To capture short-term inter-hour dependencies, outputs from preceding sub-models are recursively fed as inputs to subsequent ones. The resulting 24-hour predictions are then jointly optimized through the ResNetPlus module to produce the final load profile. In both architectures, the Scaled Exponential Linear Unit (SELU) activation function is employed across all hidden layers, whereas a linear activation is used in the output layer to generate continuous-valued predictions.

The basic structure is built utilizing a series of fully connected (FC) layers to produce initial 24-hour-ahead load projections. This is the first component of both the original DRN and PCA-DRN frameworks. In order to represent multi-scale temporal interactions, this component integrates historical load data from several time periods.

For the ISO-NE dataset, the basic structure of the DRN is designed such that each FC layer contains ten hidden units and processes multiple groups of input features, including $[L_h^{day}, T_h^{day}]$, $[L_h^{week}, T_h^{week}]$, $[L_h^{month}, T_h^{month}]$, as well as L_h^{hour} . Meanwhile, the FC layers associated with categorical temporal features-namely weekday and seasonal indicators encoded through one-hot representations (W and S)-are configured with five hidden neurons each. Additionally, the FC1 layer, FC2 layer, and the layer immediately preceding the aggregated output are all assigned ten hidden units. Except for the output layer, activation functions are applied throughout the network. Historical load information is organized at different temporal scales: L_h^{month} captures the load at

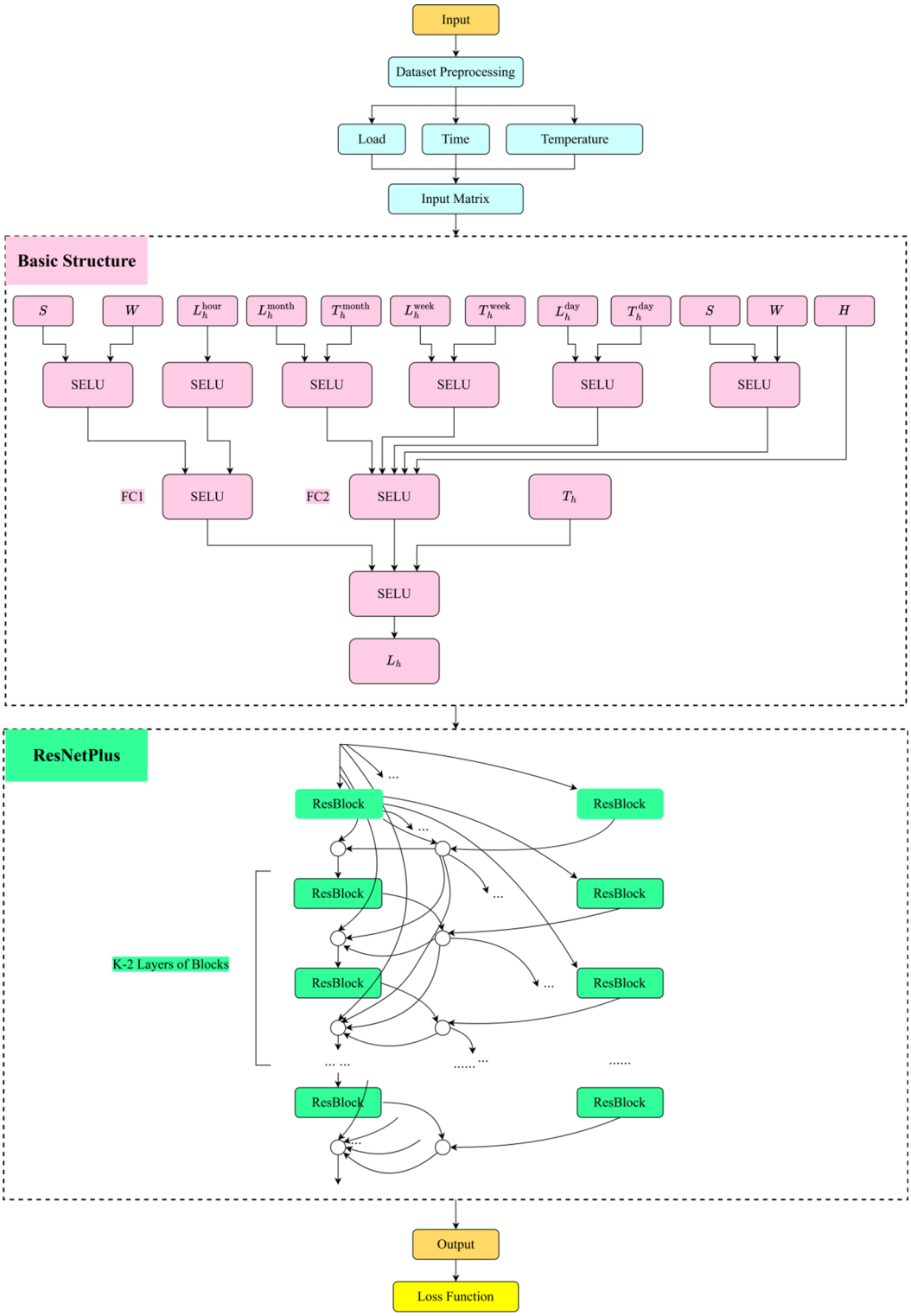

Fig. 2. Architecture of the original DRN for the ISO-NE dataset.

the same hour from one, two, and three months prior; L_h^{week} represents the load at the same hour over the previous one to eight weeks; L_h^{hour} records the load values from the preceding 24 h at the same hour; and L_h^{day} denotes the load at the same hour for each day within the previous week. Correspondingly, temperature variables T_h^{month}, T_h^{week} and T_h^{day} are aligned with their respective load components L_h^{month}, L_h^{week} and L_h^{day}, where T_h refers to the observed temperature on the target day. In addition, categorical variables S, W and H are incorporated via one-hot encoding to represent seasonal patterns, weekdays, and holiday conditions. Specifically, S encodes the four seasons (spring, summer, autumn, and winter), while H includes major public

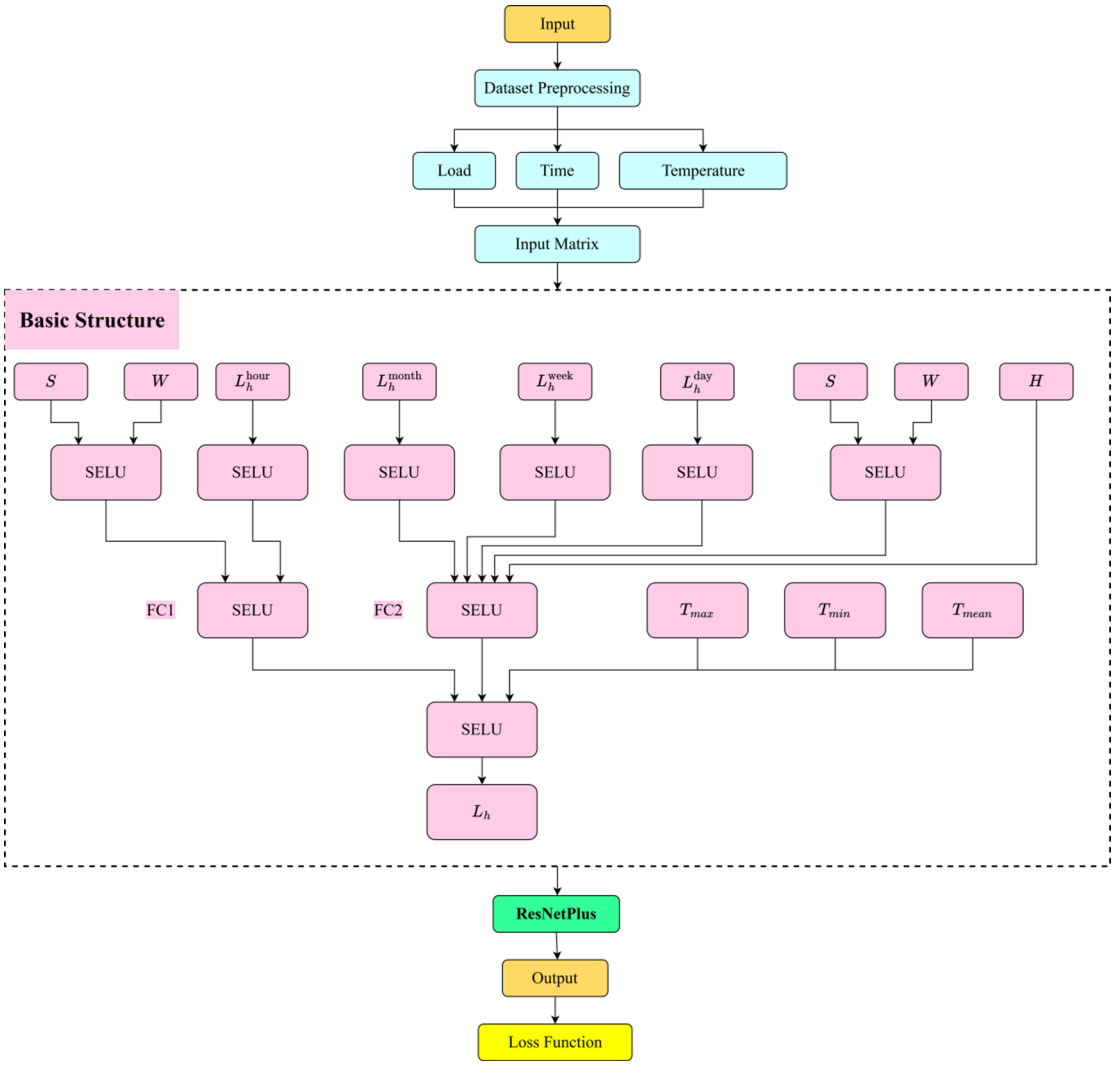

Fig. 3. Architecture of the original DRN for the MyPJ dataset.

holidays such as Christmas and Independence Day. The output generated from this basic structure is denoted as $\mathrm{L_h}$, which is subsequently forwarded to the ResNetPlus module for further feature refinement, ultimately enhancing forecasting performance.

Applied to the MyPJ dataset, the DRN framework is adapted by defining meteorological and categorical variables in accordance with region-specific climatic characteristics. The load recorded at the same hour within the most recent 24-hour window is denoted as $\mathrm{L_h^{hour}}$, while weekly periodic behavior is described by $\mathrm{L_h^{day}}$, which consists of hourly load values for each day over the previous week. To capture longer-term temporal dependencies, $\mathrm{L_h^{week}}$ aggregates load observations at the same hour across the preceding one to eight weeks, whereas $\mathrm{L_h^{month}}$ represents load patterns from one, two, and three months earlier. Each historical load component is independently processed through a dedicated FC layer comprising ten hidden neurons. Meteorological information is incorporated using daily temperature statistics, including mean, maximum, and minimum values ($\mathrm{T_{mean}}$, $\mathrm{T_{max}}$, $\mathrm{T_{min}}$), which are first concatenated and then transformed via an additional FC layer with ten hidden units. Temporal categorical variables, including season, weekday, and holiday indicators, are encoded using one-hot representations (S, W and H) and subsequently mapped through corresponding FC layers with five hidden neurons. In this context, H includes major public holidays such as Eid al-Fitr and Malaysia Independence Day, while S reflects local climatic conditions characterized by rainy and dry seasons. To strengthen feature representation, the intermediate FC layers (FC1 and FC2), along with the layer immediately preceding the aggregated output, are uniformly configured with ten hidden neurons. The final output generated by this component is denoted as $\mathrm{L_h}$, which is then forwarded to the ResNetPlus module for further refinement, thereby enhancing overall forecasting accuracy.

To overcome the limitations arising from the use of limited meteorological inputs, the PCA-DRN framework enhances the original DRN by introducing a PCA-based feature preprocessing mechanism prior to model training. Within this framework, the first component-corresponding to the basic structure-retains the same architectural configuration as the original DRN, while the representation of weather-related information is

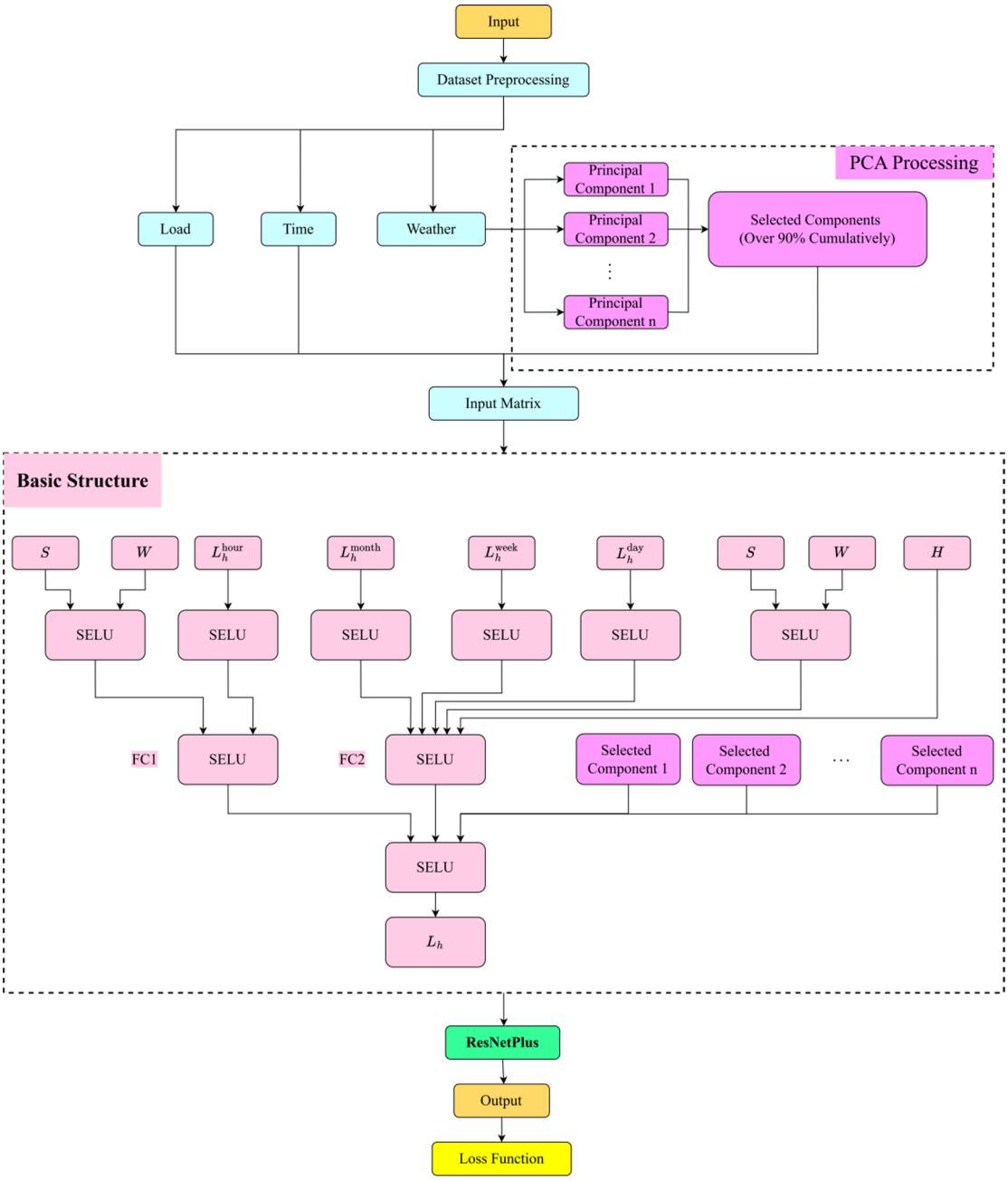

Fig. 4. Architecture of the PCA-DRN for the MyPJ dataset.

reformulated. Instead of directly employing raw meteorological variables, PCA is utilized to transform multiple weather features into a reduced-dimensional feature space. This transformation effectively mitigates redundancy while preserving the principal components that together explain approximately 90% of the cumulative variance. The extracted principal components are subsequently used as weatherrelated inputs to the basic structure. All other elements of this component remain consistent with the original DRN design. This includes the FC layers associated with L_h^{day}, L_h^{week}, L_h^{month} and L_h^{hour}, as well as the one-hot encoded categorical variables S, W and H. The aggregated output generated from this component is denoted as L_h, which is then forwarded to the ResNetPlus module for further refinement and improved forecasting performance.

ResNetPlus, the second component shared by the DRN and PCA-DRN frameworks, is an improved residual learning module that incorporates task-oriented structural modifications while adhering to the fundamental ideas of traditional ResNet designs. Each of the module's many stacked building units has a linear projection layer that guarantees dimensional compatibility for residual aggregation after a nonlinear hidden layer with 20 neurons controlled by SELU. ResNetPlus creates a deep hierarchical structure that can gradually improve feature representations by stacking these units repeatedly. In the chosen design, a ResBlock is created by successively grouping four of these units, and 10 hierarchical levels repeat this grouping pattern. In deep designs, shortcut

links are included across the network to support stable optimization and effective gradient propagation. ResNetPlus allows for more efficient use of residual learning inside the DRN-based forecasting framework by rearranging the internal block composition without changing the initial hyperparameter values, in contrast to the normal ResNet architecture. This design enables progressive feature refinement while maintaining stable gradient propagation, thereby improving training stability in deep residual learning for STLF.

To ensure a systematic and comprehensive evaluation, a diverse set of loss functions is considered in this study, including traditional error-based losses, robust loss functions, and the Penalized loss originally proposed within the DRN framework. Specifically, traditional loss functions comprise MAPE, MAE, and MSE, whereas the robust category includes Huber loss, Log-Cosh loss, Pseudo-Huber loss, Smooth L1 loss, and Charbonnier loss. These loss functions exhibit fundamentally different optimization characteristics that directly influence model behavior. In particular, the MSE loss imposes a quadratic penalty, making it highly sensitive to large errors and potentially leading to unstable gradient updates under high-variance conditions. In contrast, the MAE loss provides robustness to outliers through linear penalization but suffers from non-smooth gradients around zero, which may hinder convergence. The Charbonnier loss, as a smooth approximation of the L1 norm, offers continuous and bounded gradients, thereby improving optimization stability while maintaining robustness to extreme deviations. Furthermore, the Penalized loss introduces additional structural constraints by incorporating range-aware terms, enabling the model to better capture global load patterns. By capturing these distinct properties—such as gradient smoothness, sensitivity to outliers, and structural constraints—these loss formulations enable a comprehensive investigation of their influence on convergence stability and predictive performance under different data conditions. Figure 5 presents the overall experimental framework adopted in this study, illustrating the complete pipeline from data preprocessing and feature construction to model development using DRN and PCA-DRN, followed by loss function selection for performance evaluation.

To enable a controlled and fair comparison, only the loss function is varied across all experiments, while the network architecture, hyperparameter configurations, and training procedures are kept identical for both DRN and PCA-DRN models. This design ensures that any observed differences in forecasting performance can be attributed exclusively to the choice of loss function rather than structural or training-related variations. Furthermore, experiments are conducted on two real-world datasets, ISO-NE and MyPJ, which represent distinct climatic conditions (temperate and tropical), thereby enabling a comprehensive assessment of the robustness and generalizability of different loss functions in STLF.

Among the considered loss functions, particular emphasis is placed on the Penalized loss, which was originally introduced in the DRN framework for STLF and is therefore adopted as a baseline in this study. This loss function is specifically designed for STLF tasks by incorporating additional penalty terms to emphasize large forecasting errors. By jointly constraining point-wise prediction accuracy and deviations from realistic load ranges, the Penalized loss enhances training stability and improves forecasting reliability. As expressed in

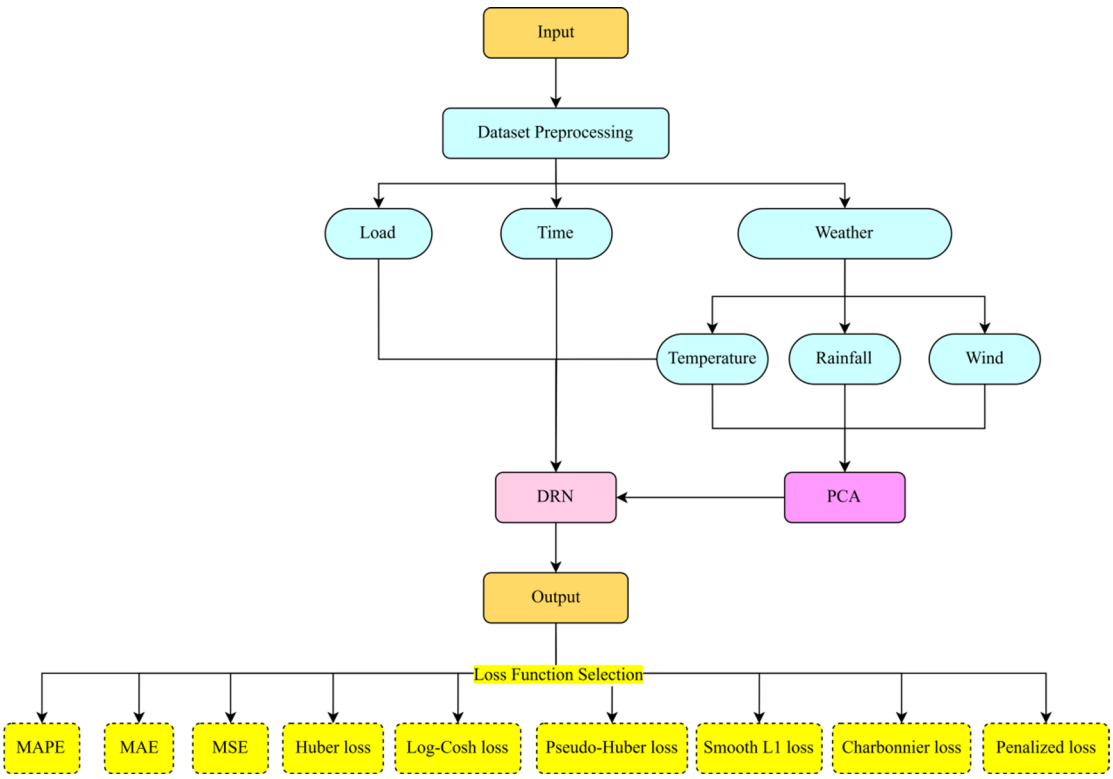

Fig. 5. Overall experimental framework.

Eq. (1), the total loss is composed of an accuracy-oriented term ($\mathrm{Loss_E}$) together with a range-aware penalty term ($\mathrm{Loss_R}$).

$$\mathrm{Loss} = \mathrm{Loss_E} + \mathrm{Loss_R} \tag{1}$$

In Eq. (2), the error component $\mathrm{Loss_E}$ is formulated based on the MAPE, by which forecasting accuracy is quantified through the relative deviation between the actual normalized load $y_{j,h}$ and the predicted normalized load $\hat{y}_{j,h}$ at the h-th hour of the j-th day. Here, Num represents the total number of samples, while H is fixed at 24 to correspond to the number of hourly load observations within a day. Defined in Eq. (3), the range-aware component $\mathrm{Loss_R}$ introduces penalties on unrealistic daily extrema by evaluating discrepancies between predicted and observed maximum and minimum load values, thereby discouraging overestimation of peak loads as well as underestimation of trough values. By simultaneously enforcing predictive accuracy and range consistency, improved robustness and training stability are achieved, enabling the Penalized loss to effectively guide the learning process within both DRN and PCA-DRN frameworks.

$$\mathrm{Loss_E} = \frac{1}{\mathrm{NumH}} \sum_{j=1}^{N} \sum_{h=1}^{H} \frac{\left|\hat{y}_{(j,h)} - y_{(j,h)}\right|}{y_{(j,h)}} \tag{2}$$

$$\begin{aligned} \mathrm{Loss_R} = \frac{1}{2\mathrm{Num}} \sum_{j=1}^{\mathrm{Num}} & \max\left(0, \max_h \hat{y}_{(j,h)} - \max_h y_{(j,h)}\right) \\ & + \max\left(0, \min_h y_{(j,h)} - \min_h \hat{y}_{(j,h)}\right) \end{aligned} \tag{3}$$

Design of experiments

This study utilizes real-world power load datasets obtained from ISO-NE and MyPJ. Table 1 summarizes the data partitioning strategy for both datasets. These datasets capture diverse load variations influenced by weather-related factors and provide a reliable basis for evaluating forecasting performance across both temperate and tropical environments.

To establish a meaningful performance reference, several widely adopted DL models in STLF are included as benchmark methods. These comprise convolution-based architectures, such as CNN and TCN, recurrent models such as LSTM, attention-based models such as Transformer, and representative hybrid frameworks including Transformer–LSTM[25] and CNN–LSTM–MMA[26], which reflect commonly adopted modeling paradigms in the literature. It should be noted that this benchmark comparison is conducted under the original experimental setting based on the Penalized loss adopted in the DRN framework. The primary objective of incorporating these baseline models is to highlight the relative advantages of the DRN and PCA-DRN frameworks in capturing nonlinear dependencies, modeling complex load patterns, and achieving superior forecasting performance under a consistent and commonly used training setting. Rather than evaluating the effect of different loss functions across heterogeneous architectures, this comparison is intended to demonstrate the effectiveness of the proposed residual-based frameworks in relation to mainstream DL approaches reported in the STLF literature.

To maintain consistency across different approaches, the CNN baseline uses a one-dimensional convolution (Conv1D) architecture with 32 filters, a kernel size of 3, Rectified Linear Unit (ReLU) activation, and He normal initialization. For every other variable, the default setup is used. The TCN model is built using stacked Conv1D layers with dilated causal convolutions, where the dilation rates are set to {1, 2, 4, 8}, to represent temporal relationships across different time scales. The TCN's Conv1D layers employ the same activation function, kernel size, and number of filters as the CNN baseline; all other parameters are kept constant. All other architectural and training parameters are left as default, and the LSTM baseline contains 64 hidden units. The Transformer baseline employs a normal encoder-only architecture with the embedding dimension set to 64 in order to give sufficient modeling capacity for self-attention methods without restricting it to convolutional filter sizes. The model consists of a single encoder layer with eight attention heads and a feedforward network of size 2048; all other elements, including dropout (set to 0.1) and positional encoding, are specified by default. The Transformer–LSTM and CNN–LSTM–MMA hybrid baselines are thoroughly executed in accordance with the designs and parameter values specified in their initial study.

To systematically investigate the influence of loss function selection, the DRN and PCA-DRN frameworks are implemented according to their original architectural configurations, without introducing any additional structural modifications. In addition to the Penalized loss originally adopted in the DRN framework, a set of representative loss functions is considered, including conventional error-based losses such as MAPE, MAE, and MSE, together with robust alternatives including Huber loss, Log-Cosh loss, Pseudo-Huber loss, Smooth L1 loss, and Charbonnier loss. This inclusion allows the Penalized loss to serve as a reference baseline for evaluating the relative effectiveness of alternative loss formulations. To ensure a controlled and fair comparison, all experiments

Dataset	Training period	Testing period	Training samples	Testing samples
ISO-NE	2003.03–2005.12	2006.01–2006.12	24,888	8760
MyPJ	2020.01–2021.12	2022.01–2022.12	17,544	8760

Table 1. Data partition of ISO-NE and MyPJ.

are conducted under identical settings, including input features, network architecture, optimizer, learning rate, batch size, and training epochs. For loss functions involving additional parameters, commonly adopted parameter settings are used without additional tuning to ensure a fair and consistent comparison across different loss formulations. Specifically, the transition parameter for Huber-type losses (including Huber, Pseudo-Huber, and Smooth L1 losses) is set to 1.0, while the smoothing parameter of the Charbonnier loss is set to $\varepsilon = 0.001$. The Log-Cosh loss is implemented in its standard form without additional parameters. Under this unified experimental protocol, any observed differences in forecasting performance can be primarily attributed to the characteristics of the employed loss functions.

After an initial training phase of 600 epochs, the model is trained using a mini-batch size of 32 for a total of 700 epochs, consisting of two additional training periods of 50 epochs each[28]. A snapshot ensemble learning approach is adopted to improve model robustness and mitigate overfitting. Specifically, model weights are saved at the end of each 50-epoch phase, and the final prediction is obtained by averaging the outputs of these snapshots[64]. This method enhances prediction stability and generalization performance while offering a computationally efficient alternative to training multiple independent models[65]. The Adaptive Moment Estimation (Adam) optimizer[66] is used with its default learning rate.

To ensure reproducibility, the experimental protocol described above is further clarified as follows. The datasets are partitioned using a chronological split, where earlier data are used for training and later data are reserved for testing, without random shuffling. No separate validation set is introduced, following the commonly adopted DRN-based STLF setting. The snapshot ensemble strategy is implemented by saving model weights at predefined intervals (every 50 epochs after the initial 600 epochs), resulting in two snapshot models whose outputs are averaged for final prediction. The training process is conducted under fixed initialization settings to ensure consistency across runs. Identical training configurations are applied across all models to ensure fair comparison.

The observed performance differences are carefully evaluated for statistical significance using a nonparametric bootstrap resampling method with 10,000 iterations. Compared to the paired Student's t-test, which assumes normally distributed paired differences, the bootstrap technique provides a more dependable foundation for model comparison since it is distribution-free[67]. Statistical significance is determined by two complementary criteria. First, if the 95% confidence interval (CI) for the mean performance difference is entirely above zero, the improvement is considered statistically significant; if the interval contains zero, the difference is considered irrelevant. Second, statistical significance at the 95% confidence level is shown by a bootstrap-derived p-value of less than 0.05. Remember that a stated value of $p \approx 0$ does not indicate a perfect zero, but rather an exceedingly minimal possibility (often < 0.0001).

Every test was conducted in a Python 3.8 environment using TensorFlow 2.10.0 and Keras 2.10.0 as the DL backends. A Lenovo laptop with an AMD Ryzen 7 6800 H CPU, 16 GB DDR5 4800 MHz RAM, and an NVIDIA GeForce RTX 3050 Ti Laptop GPU (4 GB) was used for the computations.

Metrics for assessment

Consistent with previous DRN-based STLF research, this study assesses predicting performance using a set of complementing indicators. Because MAPE is widely used in STLF investigations and is easy to comprehend, it is chosen as the main assessment criterion. Additionally, it provides the foundation for model comparison and further statistical research.

The evaluation measures are divided into three sections in order to offer a thorough review. Point-forecast error metrics, such as MAPE and MAE, which assess average relative and absolute deviations, respectively, make up the first group. The second category consists of squared-error-based metrics that emphasize error dispersion and penalize significant deviations more severely. These metrics include MSE, Root Mean Squared Error (RMSE), and Normalized Mean Squared Error (NMSE). The Correlation Coefficient (R) and the Coefficient of Determination (R^2), two correlation-based metrics that evaluate the overall agreement and goodness-of-fit between predicted and observed load levels, make up the third category.

The evaluation metrics employed in this study can be organized into three categories in accordance with the previously described framework. The first category comprises point-forecast error metrics, namely MAPE and MAE, as given in Eqs. (4) and (5), where y_i and $\hat{y}_i$ denote the actual and predicted values of the -th sample, respectively, and N represents the total number of samples; these metrics capture the average relative and absolute discrepancies between predictions and observations.

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{4}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i-1}^{N}|y_i - \hat{y}_i| \tag{5}$$

The second category includes squared-error-based measures, specifically MSE, RMSE, and NMSE, as defined in Eqs. (6)–(8), which highlight the dispersion of prediction errors by imposing larger penalties on significant deviations and thus provide additional insight into error variability; in Eq. (8), σ_y^2 denotes the variance of the observed values.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i-1}^{N}(y_i - \hat{y}_i)^2 \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}\right)^2} \tag{7}$$

$$\mathrm{NMSE} = \frac{\sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}\right)^2}{\mathrm{N} \cdot \sigma_{\hat{\mathrm{y}}}^2} \tag{8}$$

The third category consists of correlation-oriented metrics, including the R and R^2, as expressed in Eqs. (9) and (10), where $\bar{y}$ and $\hat{\bar{y}}$ represent the mean values of the actual and predicted series, respectively; these metrics evaluate the overall consistency and goodness-of-fit between predicted and actual load profiles.

$$\mathrm{R} = \frac{\sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \bar{\mathrm{y}}\right)\left(\hat{\mathrm{y}}_\mathrm{i} - \hat{\bar{\mathrm{y}}}\right)}{\sqrt{\sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \mathrm{y}\right)^2 \sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\hat{\mathrm{y}}_\mathrm{i} - \hat{\bar{\mathrm{y}}}\right)^2}} \tag{9}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}\right)^2}{\sum_{\mathrm{i}-1}^{\mathrm{N}} \left(\mathrm{y_i} - \bar{\mathrm{y}}_\mathrm{i}\right)^2} \tag{10}$$

In general, smaller values of MAPE, MAE, MSE, RMSE, and NMSE indicate higher predictive accuracy, whereas larger values of R and R^2 imply stronger correlation and improved model fitting capability.

Experimental results and analysis

Comparison with benchmark methods under penalized loss

Comparison with benchmark methods (ISO-NE dataset)

Table 2 presents the forecasting performance of the considered models on the ISO-NE dataset under a unified training setting. As observed, all models achieve strong predictive performance, with MAPE values concentrated within a relatively narrow range, indicating stable learning behavior across different DL architectures.

Among all models, the proposed DRN achieves the best performance among the considered models, attaining the lowest MAPE value of 0.017182, marginally outperforming the Transformer–LSTM hybrid (0.017280). This result demonstrates that the DRN framework is capable of achieving superior accuracy while maintaining a more structured and efficient residual learning mechanism compared with more complex hybrid architectures.

Compared with conventional models such as CNN, TCN, and LSTM, the DRN consistently achieves lower error values across all evaluation metrics, including MAE, MSE, RMSE, NMSE, R and R^2. The performance improvement over standalone CNN and LSTM models indicates that residual learning enhances feature representation and facilitates stable gradient propagation in deep networks. Although TCN demonstrates competitive performance due to its ability to model temporal dependencies, it is still slightly outperformed by DRN, suggesting that residual connections provide additional benefits in capturing nonlinear load dynamics.

From the perspective of correlation-based metrics, DRN achieves the highest values of R (0.989767) and R^2 (0.979568), reflecting a strong agreement between predicted and actual load values. This confirms that the DRN framework is capable of accurately capturing both the magnitude and variation patterns of electricity demand in a temperate climate.

Taken together, the results demonstrate that the DRN model achieves the best overall performance on the ISO-NE dataset, highlighting the effectiveness of residual learning in capturing complex and highly variable load patterns. The consistently strong performance across all evaluation metrics further indicates that the DRN framework provides a robust and efficient modeling strategy for STLF under temperate climatic conditions.

Comparison with benchmark methods (MyPJ dataset)

The forecasting results for the MyPJ dataset are summarized in Table 3. Compared with the ISO-NE dataset, larger variations in model performance can be observed, reflecting the increased difficulty of forecasting under tropical climate conditions, where load patterns are more strongly influenced by short-term weather variations rather than pronounced seasonal trends.

Among all models, the PCA-DRN achieves the best overall performance among the evaluated models, with the lowest MAPE value of 0.049994, followed by the original DRN model with a MAPE of 0.052514. This

Method	MAPE	MAE	MSE	RMSE	NMSE	*R*	R^2
CNN	0.020587	0.013364	0.000435	0.020862	0.028878	0.986425	0.971122
TCN	0.018899	0.012143	0.000332	0.018216	0.022018	0.989094	0.977982
LSTM	0.020754	0.013331	0.000416	0.020397	0.027605	0.986165	0.972395
Transformer	0.021925	0.014352	0.000517	0.022749	0.034337	0.982763	0.965663
Transformer-LSTM	0.017280	0.011185	0.000311	0.017624	0.020609	0.989707	0.979391
CNN-LSTM-MMA	0.019086	0.012157	0.000348	0.018645	0.023067	0.988523	0.976933
DRN	0.017182	0.011138	0.000308	0.017548	0.020432	0.989767	0.979568

Table 2. Performance comparison of benchmark models on the ISO-NE dataset under the penalized loss setting.

Method	MAPE	MAE	MSE	RMSE	NMSE	*R*	R^2
CNN	0.054093	0.028479	0.002152	0.046394	0.075599	0.963564	0.924401
TCN	0.057486	0.030419	0.002431	0.049303	0.085378	0.957301	0.914622
LSTM	0.056978	0.029211	0.002547	0.050470	0.089469	0.954934	0.910531
Transformer	0.054016	0.026741	0.002139	0.046247	0.075124	0.961710	0.924876
Transformer-LSTM	0.051774	0.026462	0.002056	0.045343	0.072215	0.963724	0.927785
CNN-LSTM-MMA	0.056267	0.028622	0.002458	0.049577	0.086330	0.956972	0.913670
DRN	0.052514	0.026467	0.002050	0.045278	0.072007	0.964032	0.927993
PCA-DRN	0.049994	0.026254	0.001866	0.043193	0.065527	0.967339	0.934473

Table 3. Performance comparison of benchmark models on the MyPJ dataset under the penalized loss setting.

indicates that incorporating additional meteorological variables through PCA-based feature extraction can significantly enhance forecasting accuracy in environments where weather-related factors play a dominant role.

Compared with traditional and hybrid benchmark models, both DRN and PCA-DRN consistently outperform CNN, TCN, LSTM, Transformer, and CNN–LSTM–MMA across all evaluation metrics. In particular, the performance gap between residual-based models and conventional architectures is more pronounced than that observed in the ISO-NE dataset, suggesting that the advantages of residual learning become more evident when modeling more complex and less regular load patterns.

The Transformer–LSTM model achieves the best performance among the benchmark models, with a MAPE of 0.051774, but still falls short of the DRN and PCA-DRN frameworks. This suggests that while hybrid architectures can effectively capture temporal dependencies, they may not fully exploit complex interactions among multiple meteorological variables as efficiently as residual-based frameworks, especially when combined with dimensionality reduction techniques.

In terms of correlation metrics, PCA-DRN achieves the highest values of R (0.967339) and R^2 (0.934473), indicating superior consistency between predicted and actual load values. The improvement in NMSE further confirms that PCA-DRN effectively reduces error dispersion, leading to more stable forecasting performance.

Taken together, the results on the MyPJ dataset highlight the importance of both model architecture and input representation in STLF tasks under tropical conditions. The superior performance of PCA-DRN demonstrates that combining residual learning with dimensionality reduction is an effective strategy for handling complex meteorological influences and improving forecasting reliability.

When compared with the ISO-NE results, a clearer performance gap among models can be observed in the MyPJ dataset, indicating that forecasting under tropical conditions poses greater challenges due to higher variability and stronger dependence on short-term weather factors. In this context, while the DRN framework maintains strong performance across both datasets, the PCA-DRN further enhances prediction accuracy by effectively exploiting additional meteorological information, making it particularly suitable for more complex and data-rich environments.

Comparison under different loss functions

DRN performance On the ISO-NE dataset

Figure 6 illustrates the comparison of MAPE across different loss functions for the DRN model on the ISO-NE dataset, providing an intuitive overview of the relative performance trends. It can be observed that loss functions with smoother gradient properties, particularly the Charbonnier loss and Log-Cosh loss, consistently achieve lower MAPE values, indicating more stable and effective optimization behavior.

To further examine the distributional characteristics of forecasting errors, the boxplot visualization is presented in Fig. 7. The Charbonnier loss exhibits a more concentrated interquartile range and a lower median MAPE compared with other loss functions, indicating reduced variability and more stable prediction performance. In contrast, loss functions such as MSE and Pseudo-Huber display wider interquartile ranges and longer whiskers, suggesting higher sensitivity to large deviations and less consistent optimization behavior. From an optimization perspective, this difference can be attributed to the smooth and bounded gradient properties of the Charbonnier loss, which enable stable parameter updates across different error magnitudes. In contrast, the quadratic nature of the MSE loss amplifies large errors, leading to unstable gradients, while the piecewise structure of Pseudo-Huber may introduce suboptimal transitions between linear and quadratic regions. These observations further confirm that loss functions with smoother gradient behavior contribute to more reliable convergence and improved robustness.

Table 4 presents the detailed quantitative results corresponding to these observations. Overall, the results demonstrate that the choice of loss function has a noticeable impact on both optimization behavior and final forecasting accuracy, even under a fixed network architecture.

Among all considered loss functions, the Charbonnier loss achieves the best performance among the evaluated loss functions, yielding the lowest MAPE value of 0.016707. It consistently outperforms other loss functions across point-forecast error metrics (MAPE, MAE), squared-error-based measures (MSE, RMSE, NMSE), and correlation-oriented metrics (R, R^2), indicating a well-balanced performance in terms of prediction accuracy, error distribution, and goodness of fit. This superior performance can be attributed to its formulation as a smooth and differentiable approximation of the L1 norm, which leads to more stable gradient behavior during optimization. Unlike MAE, which suffers from non-differentiability at zero and may cause abrupt gradient

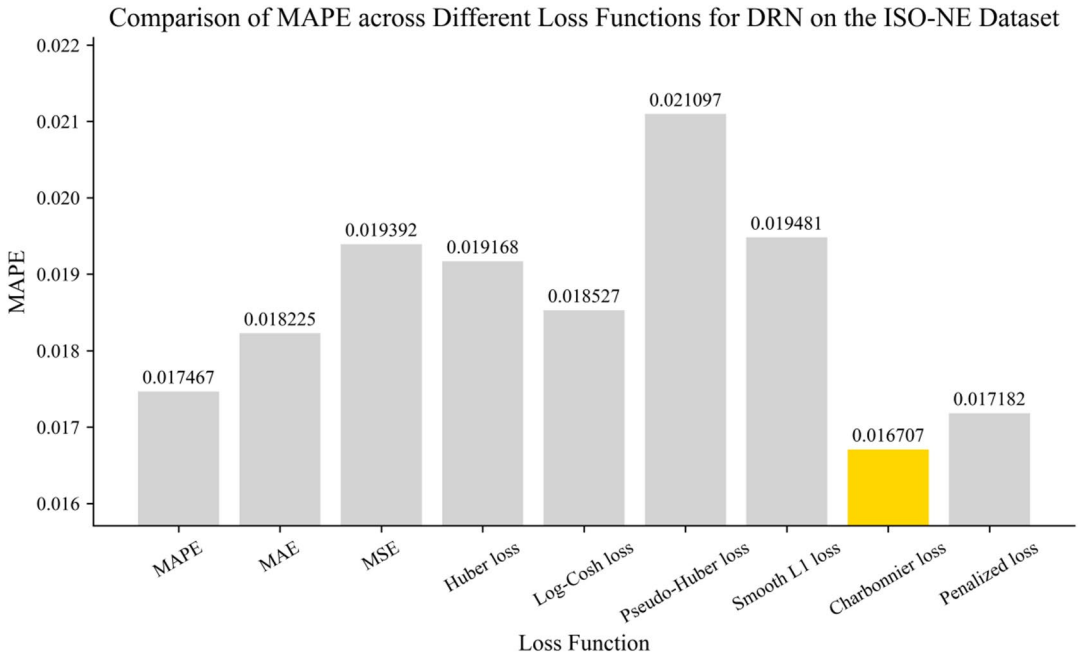

Fig. 6. Comparison of MAPE across different loss functions in DRN (ISO-NE).

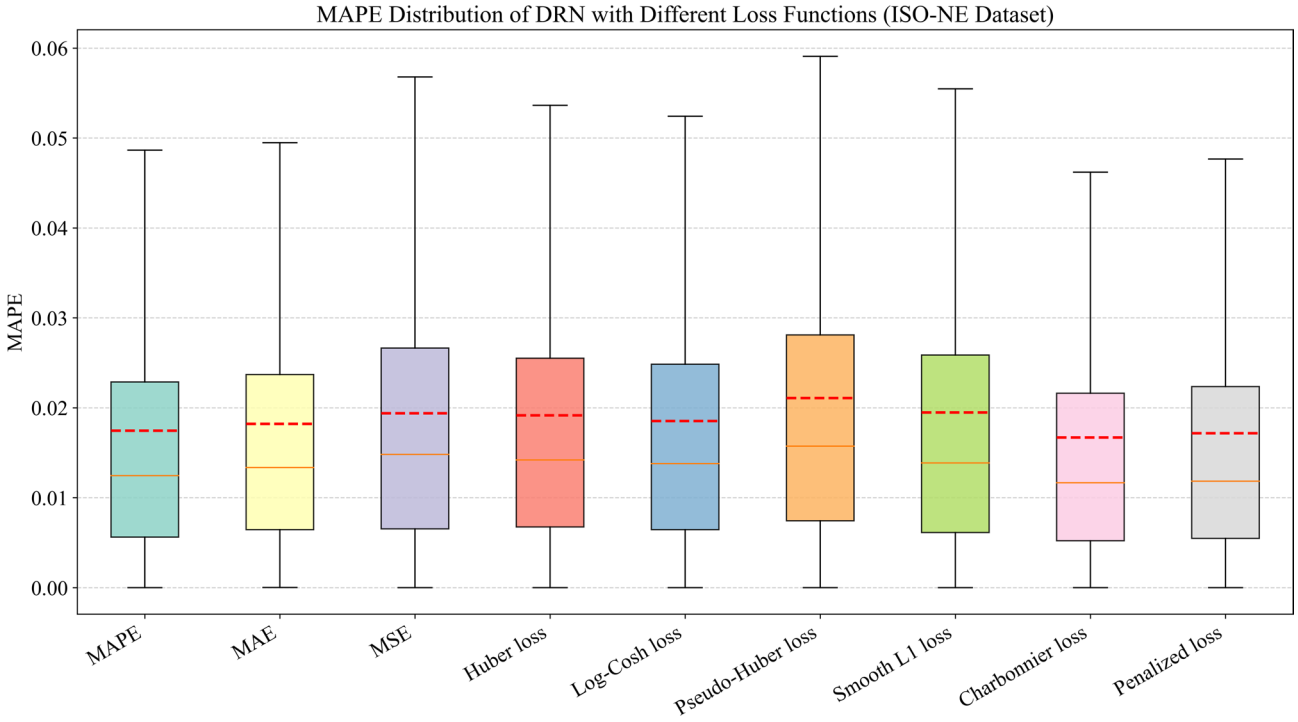

Fig. 7. Distribution of MAPE for DRN under different loss functions on the ISO-NE dataset.

changes, and MSE, which amplifies large errors due to its quadratic form and may result in unstable updates, the Charbonnier loss provides continuous and bounded gradients across the entire error range. This property facilitates stable backpropagation and prevents excessive parameter updates caused by extreme deviations, thereby enhancing robustness to outliers. In addition, the smoothness of the Charbonnier loss contributes to a more well-behaved optimization landscape, enabling more reliable convergence and improved generalization performance. This balance is particularly effective for the ISO-NE dataset, where load patterns exhibit strong seasonal variability but relatively structured temporal dynamics. Furthermore, from a convergence perspective, the smooth gradient profile of the Charbonnier loss helps reduce oscillations during training and enables

Loss function	MAPE	MAE	MSE	RMSE	NMSE	R	R^2
MAPE	0.017467	0.011252	0.000313	0.017698	0.020782	0.989672	0.979218
MAE	0.018225	0.011583	0.000312	0.017661	0.020695	0.990025	0.979305
MSE	0.019392	0.012427	0.000326	0.018047	0.021610	0.989612	0.978390
Huber loss	0.019168	0.012160	0.000332	0.018233	0.022057	0.989345	0.977943
Log-Cosh loss	0.018527	0.011847	0.000300	0.017333	0.019935	0.990011	0.980065
Pseudo-Huber loss	0.021097	0.013439	0.000386	0.019650	0.025620	0.987907	0.974380
Smooth L1 loss	0.019481	0.012575	0.000358	0.018934	0.023787	0.989227	0.976213
Charbonnier loss	0.016707	0.010716	0.000291	0.017055	0.019299	0.990346	0.980701
Penalized loss	0.017182	0.011138	0.000308	0.017548	0.020432	0.989767	0.979568

Table 4. Performance comparison of DRN under different loss functions on the ISO-NE dataset.

more stable parameter updates. This leads to a more well-behaved optimization process, facilitating reliable convergence and improving overall training stability.

The Log-Cosh loss also demonstrates competitive performance, achieving the second-best MAPE and the lowest MSE among all loss functions. Its smooth approximation to the absolute error enables continuous gradient updates, which helps avoid abrupt changes during optimization. As a result, Log-Cosh provides a favorable compromise between sensitivity to small errors and robustness to moderate deviations.

Among traditional loss functions, the MAPE-based loss performs relatively well due to its alignment with the evaluation metric. However, its formulation introduces instability when actual values approach small magnitudes, leading to potentially large gradient fluctuations. The MAE loss achieves moderate performance, reflecting its robustness to outliers but limited ability to emphasize larger deviations. In contrast, the MSE loss yields inferior MAPE performance, as its quadratic penalty overemphasizes large errors, causing the optimization process to be dominated by extreme deviations.

For other robust loss functions, the Huber loss provides stable but not optimal performance, suggesting that its piecewise formulation may not fully capture the characteristics of the dataset. The Pseudo-Huber loss, despite being a smooth approximation, exhibits degraded performance, indicating that excessive smoothing may reduce sensitivity to meaningful error variations. Similarly, the Smooth L1 loss does not offer additional benefits, likely due to suboptimal balancing between linear and quadratic regions.

The Penalized loss, originally proposed in the DRN framework and extending the MAPE formulation by incorporating a range-aware penalty term, achieves strong performance with a MAPE of 0.017182, closely approaching that of the Charbonnier loss. Its effectiveness lies in the incorporation of a range-aware penalty term, which constrains unrealistic daily extrema by penalizing deviations in peak and trough values. By jointly optimizing point-wise accuracy and global load consistency, the Penalized loss enhances robustness to extreme prediction errors. However, its performance remains slightly inferior to the Charbonnier loss, suggesting that, for the ISO-NE dataset, smooth gradient behavior and balanced error handling are more critical than explicit range constraints.

Taken together, the results indicate that loss functions characterized by smooth gradients and robust error handling, such as Charbonnier and Log-Cosh, are particularly well-suited for temperate-climate datasets with structured variability. These findings highlight the importance of selecting loss functions that ensure stable optimization while maintaining sensitivity to meaningful prediction errors.

DRN performance on the MyPJ dataset

Figure 8 presents the comparison of MAPE across different loss functions for the DRN model on the MyPJ dataset, highlighting more pronounced performance variations compared with the ISO-NE dataset. It is evident that loss functions designed to handle noisy and irregular data distributions, particularly the Charbonnier loss, achieve superior performance, reflecting the increased importance of robust optimization under tropical conditions.

To further examine the distributional characteristics of forecasting errors, the boxplot visualization is presented in Fig. 9, where more pronounced variability can be observed compared with the ISO-NE dataset. The Charbonnier loss maintains a relatively concentrated interquartile range and a lower median MAPE, indicating strong robustness and stable prediction performance under highly variable conditions. In contrast, loss functions such as MSE and Smooth L1 exhibit wider interquartile ranges and longer whiskers, reflecting increased sensitivity to extreme deviations and less consistent optimization behavior. This difference becomes more evident in the MyPJ dataset, where load patterns are influenced by irregular and short-term fluctuations. From an optimization perspective, the smooth and bounded gradient properties of the Charbonnier loss help mitigate the impact of noise and outliers, whereas the quadratic penalty of MSE amplifies large errors, and the piecewise structure of Smooth L1 may introduce instability under rapidly changing conditions. These observations further indicate that robust loss functions with smooth gradient behavior are more suitable for datasets characterized by higher variability and non-stationary patterns.

The corresponding quantitative results are summarized in Table 5. Compared with the ISO-NE dataset, a more significant divergence in performance across loss functions can be observed, indicating that loss function selection plays a more critical role in datasets characterized by higher variability and stronger short-term fluctuations.

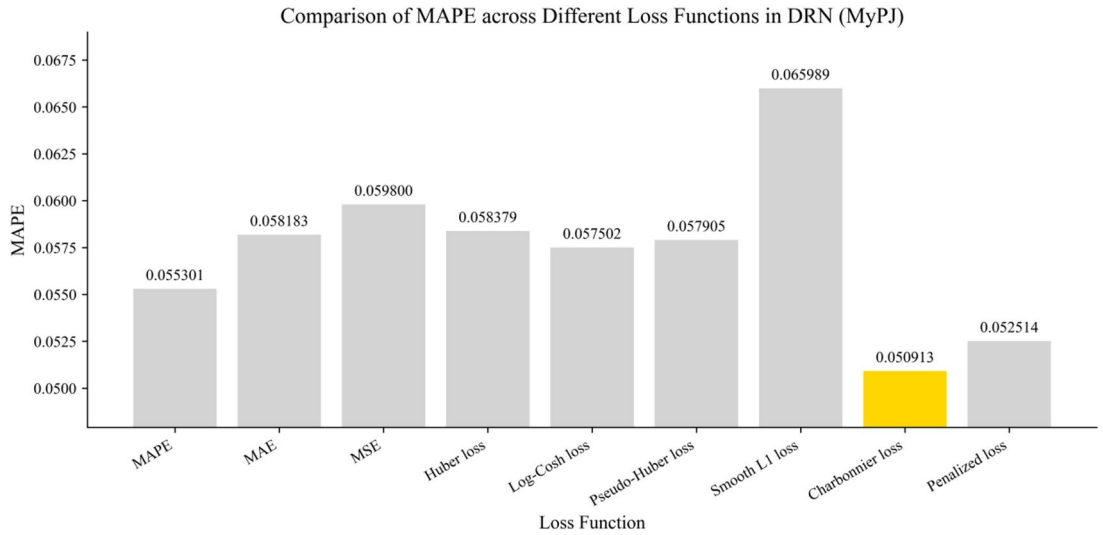

Fig. 8. Comparison of MAPE across different loss functions in DRN (MyPJ).

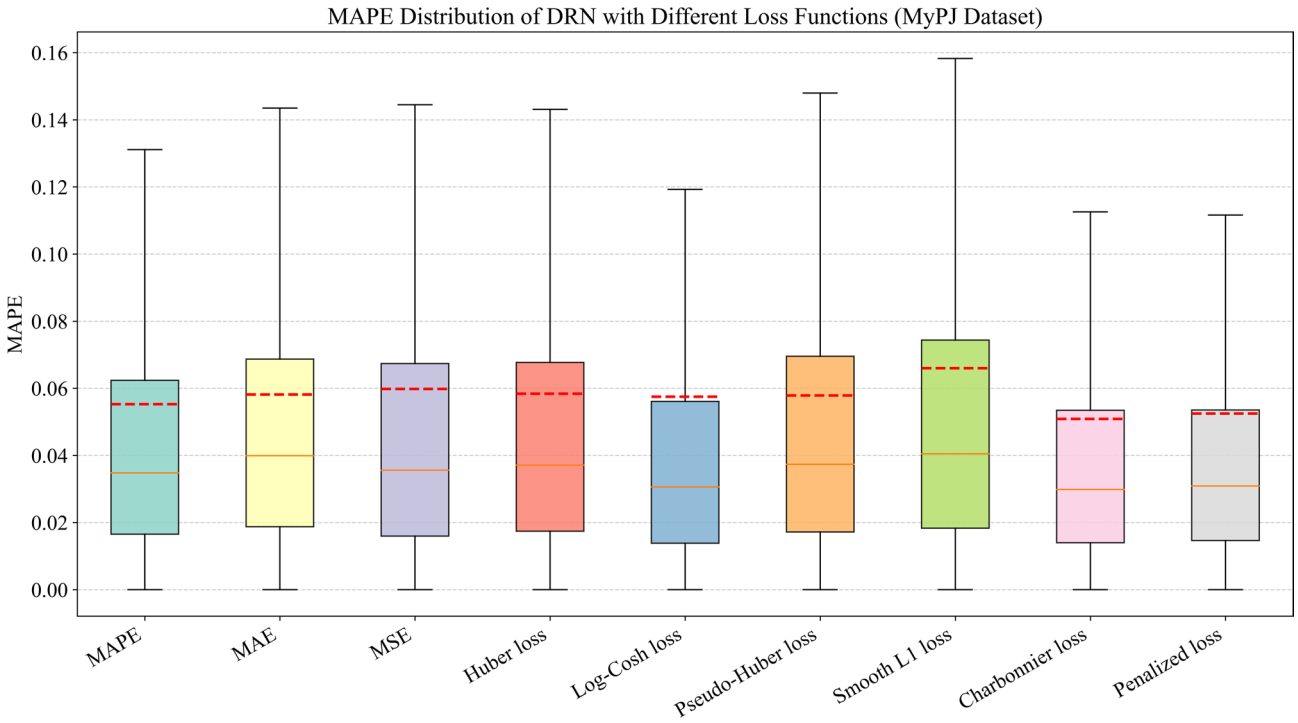

Fig. 9. Distribution of MAPE for DRN under different loss functions on the MyPJ dataset.

Among all evaluated loss functions, the Charbonnier loss again achieves the best overall performance, with the lowest MAPE value of 0.050913 and consistent improvements across all evaluation metrics. This confirms its strong capability in handling complex and noisy load data. In the MyPJ dataset, where load patterns are less regular and more influenced by short-term weather variations, the smooth formulation of the Charbonnier loss enables stable gradient updates while suppressing the impact of extreme deviations. From an optimization perspective, this behavior can be further understood by comparing it with conventional loss functions. The MSE loss imposes a quadratic penalty on large errors, which can lead to disproportionately large gradients and unstable updates under highly variable conditions. In contrast, the MAE loss applies a constant gradient but suffers from non-smoothness around zero, which may hinder convergence. The Charbonnier loss effectively balances these two extremes by maintaining smooth and continuous gradients while limiting the influence of outliers, making it particularly suitable for datasets characterized by irregular fluctuations and non-Gaussian noise, as observed in the MyPJ dataset. This property is particularly beneficial under non-Gaussian noise

Loss function	MAPE	MAE	MSE	RMSE	NMSE	R	R^2
MAPE	0.055301	0.029915	0.002433	0.049327	0.085461	0.958622	0.914539
MAE	0.058183	0.032292	0.002434	0.049335	0.085491	0.961745	0.914509
MSE	0.059800	0.031858	0.002669	0.051661	0.093739	0.953545	0.906261
Huber loss	0.058379	0.032204	0.002603	0.051022	0.091438	0.956596	0.908562
Log-Cosh loss	0.057502	0.028416	0.002652	0.051500	0.093159	0.952426	0.906841
Pseudo-Huber loss	0.057905	0.031166	0.002280	0.047744	0.080066	0.961509	0.919934
Smooth L1 loss	0.065989	0.036059	0.003260	0.057095	0.114499	0.946844	0.885501
Charbonnier loss	0.050913	0.026212	0.002031	0.045070	0.071347	0.964356	0.928653
Penalized loss	0.052514	0.026467	0.002050	0.045278	0.072007	0.964032	0.927993

Table 5. Performance comparison of DRN under different loss functions on the MyPJ dataset.

conditions and irregular load fluctuations. This observation highlights that the effectiveness of a loss function is closely related to data characteristics, where datasets with higher variability and noise levels tend to benefit more from loss formulations that balance robustness and gradient smoothness.

The Penalized loss ranks as the second-best performer, achieving a MAPE of 0.052514. Compared with its performance on the ISO-NE dataset, the relative advantage of the Penalized loss becomes less pronounced. This can be explained by the nature of tropical load patterns, where short-term variability dominates over strict daily extrema constraints. Although the range-aware penalty component effectively limits unrealistic peak and trough predictions, it may introduce additional optimization constraints that reduce flexibility in capturing rapid local variations.

Among traditional loss functions, the MAPE-based loss performs better than MAE and MSE, reflecting its focus on relative error minimization. However, its sensitivity to small denominator values still introduces instability under irregular load conditions. The MAE loss provides moderate performance due to its robustness to outliers, while the MSE loss again underperforms, as its quadratic penalty amplifies extreme errors and leads to less balanced optimization.

For other robust loss functions, the Pseudo-Huber loss shows relatively competitive performance, particularly in RMSE and NMSE, suggesting that its smooth transition between L1 and L2 behaviors can effectively handle moderate noise levels. However, its MAPE remains higher than that of the Charbonnier loss, indicating limited effectiveness in minimizing relative forecasting errors. The Huber loss provides stable but suboptimal performance, while the Log-Cosh loss, despite its smoothness, does not achieve significant improvements, possibly due to insufficient sensitivity to larger deviations.

A notable observation is the inferior performance of the Smooth L1 loss, which records the highest MAPE among all evaluated loss functions. This suggests that its piecewise linear–quadratic structure may not be well-suited for datasets with highly irregular fluctuations, where abrupt transitions in gradient behavior can hinder effective optimization.

A cross-dataset comparison between the ISO-NE and MyPJ results reveals several important insights regarding the behavior of different loss functions under varying data characteristics. For the ISO-NE dataset, which exhibits relatively structured seasonal patterns and smoother temporal variations, loss functions with stable gradient properties, such as Charbonnier and Log-Cosh, demonstrate clear advantages by enabling consistent optimization and balanced error minimization. In contrast, the MyPJ dataset is characterized by higher variability, irregular fluctuations, and stronger short-term weather influences. Under such conditions, the robustness of the loss function becomes more critical than strict adherence to structured error patterns, leading to more pronounced performance differences across loss functions.

Another notable observation is that the effectiveness of the Penalized loss varies across datasets. While it performs competitively on the ISO-NE dataset by enforcing global load consistency through range-aware constraints, its relative advantage diminishes in the MyPJ dataset, where short-term local variations dominate over daily extrema patterns. This indicates that task-specific loss formulations may not generalize equally well across datasets with fundamentally different statistical properties.

Taken together, these findings highlight that the selection of loss functions in DRN-based STLF models is inherently dataset-dependent. Loss functions that provide a balanced trade-off between gradient smoothness and robustness to outliers, particularly the Charbonnier loss, demonstrate strong generalization capability across both temperate and tropical environments, making them more suitable for practical applications involving diverse load patterns.

PCA results for PCA-DRN on the MyPJ dataset

To mitigate redundancy among meteorological variables and uncover their underlying variance structure in the MyPJ dataset, PCA was performed prior to model training. A cumulative explained variance threshold of 90% was adopted as the selection criterion, leading to the retention of five principal components. Together, these components account for 94.12% of the total variance, indicating that most of the original information is preserved despite substantial dimensionality reduction. Among them, the first component captures 37.88% of the variance, followed by the second (19.62%), third (14.23%), fourth (12.08%), and fifth (8.54%) components. This distribution suggests that the major variability of the meteorological data can be effectively represented within a compact and low-redundancy feature space.

To further interpret the retained components, the contribution of each meteorological variable is examined through their corresponding loading values. For the first principal component, loadings of rainfall, mean temperature, minimum temperature, maximum temperature, mean wind speed, maximum wind speed, and wind direction are 0.326146, − 0.579888, − 0.503639, − 0.442409, − 0.293040, 0.105115, and 0.105159, respectively, indicating that this component is primarily dominated by temperature-related variability. In contrast, the second component is characterized by stronger contributions from wind intensity and precipitation, with loadings of 0.442154, 0.082323, − 0.203754, 0.365551, 0.318846, 0.701837, and − 0.168348.

The third component is largely governed by wind direction, as evidenced by its dominant loading of − 0.946270, while other variables exhibit relatively minor contributions. The fourth component reflects a mixed influence of wind speed, temperature, and precipitation, with loadings of 0.381735, 0.141974, 0.147226, 0.413782, − 0.792360, − 0.087986, and − 0.075194. Meanwhile, the fifth component highlights a coupled relationship between rainfall and wind-related factors under varying wind strength conditions, as indicated by loadings of 0.709699, 0.086094, 0.123478, − 0.051551, 0.387480, − 0.512842, and 0.240546.

Taken together, the retained principal components not only preserve the dominant physical characteristics of the original meteorological variables but also transform them into an orthogonal and low-dimensional representation. Such a transformation effectively reduces redundancy while maintaining interpretability, thereby providing a compact and stable input representation for subsequent analyses of training behavior and forecasting performance in the PCA-DRN framework.

From a data representation perspective, PCA transforms the original correlated meteorological variables into an orthogonal feature space, thereby reducing redundancy and suppressing noise in the input distribution. This transformation leads to a more compact and structured feature representation, where dominant patterns are preserved while minor fluctuations and irrelevant variations are mitigated. As a result, the statistical properties of the input data become more stable, which can significantly influence the optimization behavior of different loss functions.

PCA-DRN performance on the MyPJ dataset

Figure 10 presents the comparison of MAPE across different loss functions for the PCA-DRN model on the MyPJ dataset. Compared with the DRN results discussed earlier, an overall improvement in forecasting accuracy can be clearly observed, indicating that the incorporation of PCA-based feature representation enhances the model's ability to extract informative patterns from multi-dimensional meteorological variables.

To further examine the distributional characteristics of forecasting errors, the boxplot visualization is presented in Fig. 11, where a noticeable reduction in variability can be observed across most loss functions compared with the DRN results. The interquartile ranges are generally more compact, and the whiskers are shorter, indicating improved stability and reduced dispersion of prediction errors under the PCA-DRN framework. This suggests that PCA-based feature transformation produces a more structured and less noisy input space, leading to more consistent optimization behavior.

Among the evaluated loss functions, the Charbonnier loss continues to exhibit a concentrated distribution with a lower median MAPE, reflecting its strong capability in minimizing point-wise prediction errors. Meanwhile, the Penalized loss demonstrates enhanced consistency compared with its performance in the DRN setting, with reduced dispersion and more stable error distribution. This indicates that the range-aware constraints of the Penalized loss become more effective when the input feature space is regularized through PCA.

From an optimization perspective, the improved performance of the Penalized loss can be attributed to the alignment between its structural constraints and the transformed feature space. By reducing redundancy and suppressing noise, PCA enables the model to better capture global load patterns, thereby enhancing the

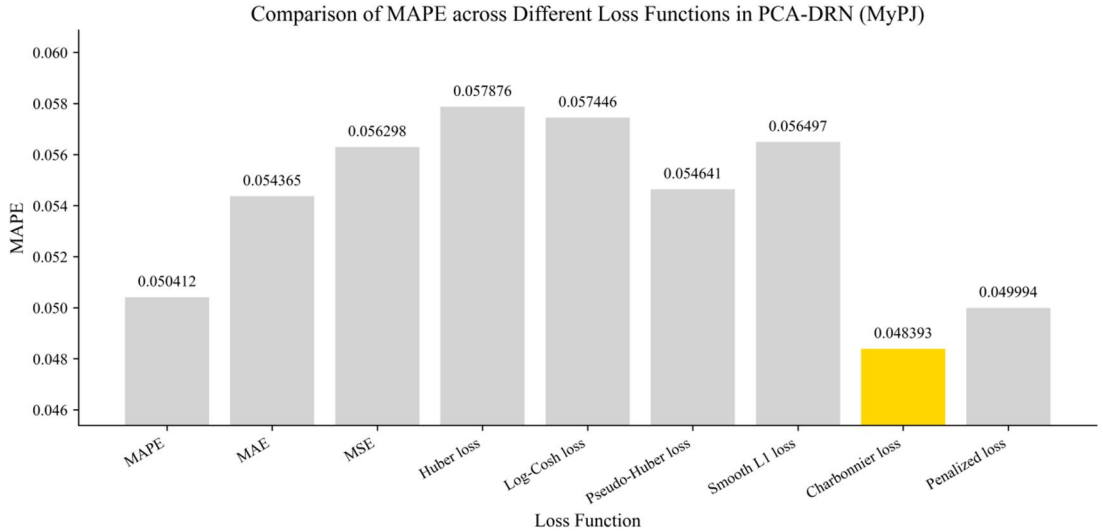

Fig. 10. Comparison of MAPE across different loss functions in PCA-DRN (MyPJ).

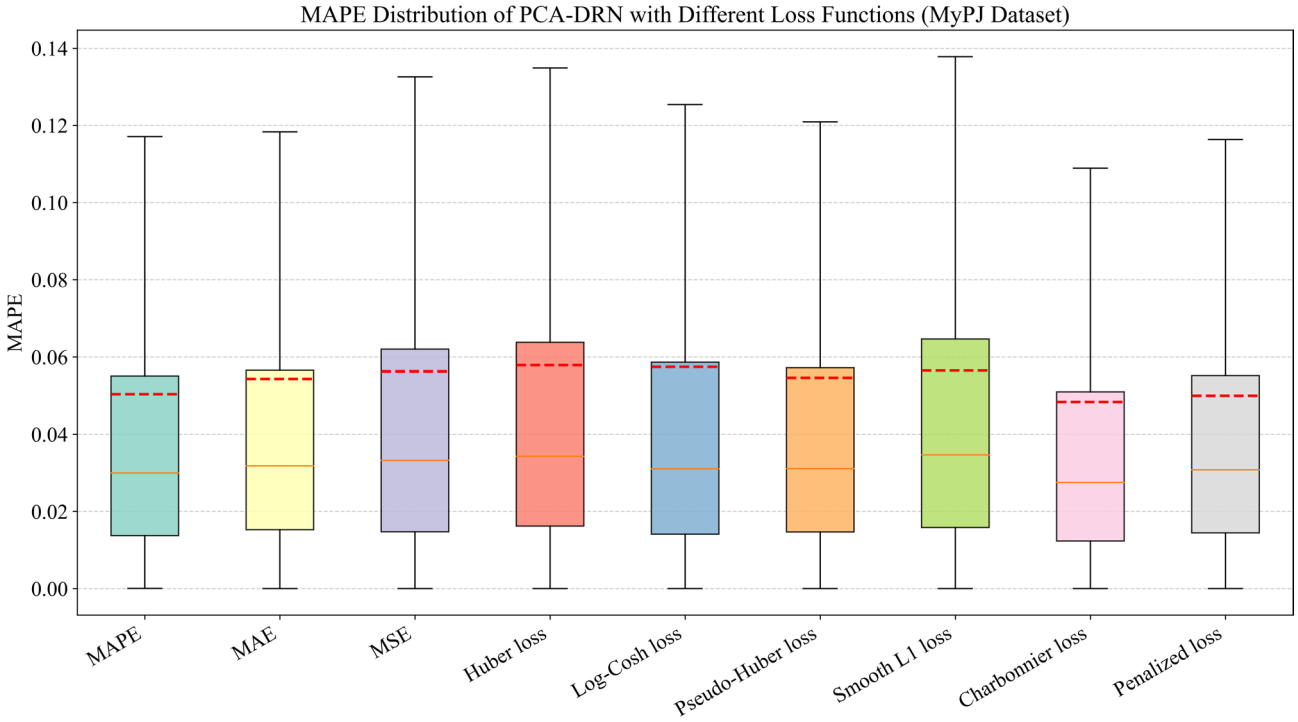

Fig. 11. Distribution of MAPE for DRN under different loss functions on the MyPJ dataset.

Loss function	MAPE	MAE	MSE	RMSE	NMSE	R	R^2
MAPE	0.050412	0.026604	0.002015	0.044886	0.070767	0.964799	0.929233
MAE	0.054365	0.027332	0.002152	0.046394	0.075601	0.961848	0.924399
MSE	0.056298	0.029551	0.002431	0.049301	0.085373	0.957148	0.914627
Huber loss	0.057876	0.030607	0.002537	0.050367	0.089103	0.956180	0.910897
Log-Cosh loss	0.057446	0.028477	0.002421	0.049203	0.085033	0.957144	0.914967
Pseudo-Huber loss	0.054641	0.027600	0.002224	0.047159	0.078116	0.960673	0.921884
Smooth L1 loss	0.056497	0.030254	0.002378	0.048767	0.083533	0.958668	0.916467
Charbonnier loss	0.048393	0.024879	0.001942	0.044071	0.068218	0.965513	0.931782
Penalized loss	0.049994	0.026254	0.001866	0.043193	0.065527	0.967339	0.934473

Table 6. Performance comparison of PCA-DRN under different loss functions on the MyPJ dataset.

effectiveness of range-based penalties. In contrast, loss functions such as MSE and Smooth L1 still exhibit relatively larger variability, suggesting that their sensitivity to large deviations or piecewise gradient behavior remains less suitable even under a more structured input space. These observations highlight that feature representation and loss function design jointly influence optimization stability and forecasting performance, rather than acting as independent components.

Table 6 summarizes the corresponding quantitative results. Similar to the DRN case, the selection of loss function continues to have a significant impact on forecasting performance. However, a notable difference emerges in the relative behavior of different loss functions after applying PCA.

Furthermore, a consistent and systematic improvement is observed when comparing PCA-DRN with DRN across all loss functions. For every evaluated loss formulation, the PCA-DRN framework achieves lower MAPE values than its DRN counterpart, demonstrating that PCA-based feature transformation enhances both representation efficiency and optimization effectiveness. This result highlights that reducing redundancy in meteorological variables enables the model to learn more informative patterns, thereby improving overall forecasting accuracy.

Among all evaluated loss functions, the Charbonnier loss achieves the best performance in terms of point-forecast error metrics, yielding the lowest MAPE (0.048393) and MAE (0.024879). This indicates that the smooth and robust formulation of the Charbonnier loss remains highly effective in minimizing point-wise prediction errors under the PCA-DRN framework.

In contrast, the Penalized loss demonstrates superior performance in squared-error-based measures (MSE, RMSE, NMSE) and correlation-oriented metrics (R, R^2). In particular, it achieves the lowest NMSE (0.065527)

1st model	2nd model	MAPE ± SD (1st model)	MAPE ± SD (2nd model)	Mean difference	CI (95%)	*p*-value
DRN (Penalized loss)	DRN (Charbonnier loss)	0.017182 ± 0.019726	0.016707 ± 0.020032	0.000476	[0.000189, 0.000754]	0.000800

Table 7. Bootstrap-based statistical comparison of DRN models under different loss functions on the ISO-NE dataset.

1st model	2nd model	MAPE ± SD (1st model)	MAPE ± SD (2nd model)	Mean difference	CI (95%)	*p*-value
DRN (Penalized loss)	DRN (Charbonnier loss)	0.052514 ± 0.106031	0.050913 ± 0.103926	0.001601	[0.000930, 0.002262]	≈ 0
PCA-DRN (Penalized loss)	PCA-DRN (Charbonnier loss)	0.049994 ± 0.088990	0.048393 ± 0.097656	0.001601	[0.000925, 0.002283]	≈ 0
DRN (Penalized loss)	PCA-DRN (Penalized loss)	0.052514 ± 0.106031	0.049994 ± 0.088990	0.002520	[0.001416, 0.003690]	0.000100
DRN (Charbonnier loss)	PCA-DRN (Charbonnier loss)	0.050913 ± 0.103926	0.048393 ± 0.097656	0.002520	[0.001599, 0.003469]	≈ 0

Table 8. Bootstrap-based statistical comparison of DRN and PCA-DRN models under different loss functions on the MyPJ dataset.

and the highest correlation values ($R = 0.967339$, $R^2 = 0.934473$), suggesting that it provides better global fitting consistency and more accurately captures the overall structure of load variations.

This divergence in performance between the Charbonnier and Penalized losses can be attributed to the impact of PCA on the input feature space. More specifically, by reducing noise and inter-feature correlations, PCA produces a more regularized input distribution with clearer global structure. Under such conditions, loss functions that incorporate structural constraints, such as the Penalized loss, can more effectively enforce consistency between predicted and actual load ranges. In contrast, when the input space contains redundant or noisy features, such constraints may be less effective due to increased variability in the optimization process. Therefore, the improvement of the Penalized loss under the PCA-DRN framework can be attributed to the enhanced alignment between its formulation and the structured feature space produced by PCA. From an optimization standpoint, PCA reduces noise and redundancy in the input space, leading to a smoother and more structured loss landscape. Under such conditions, the range-aware constraints of the Penalized loss can be more effectively enforced, allowing it to better capture global load patterns and improve overall fitting consistency. By transforming the original meteorological variables into orthogonal principal components, PCA reduces redundancy and suppresses noise, resulting in a more compact and structured representation. Under such conditions, the advantage of the Penalized loss in enforcing global consistency through its range-aware constraint becomes more pronounced, as the underlying signal is less affected by redundant or noisy features.

On the other hand, although the Charbonnier loss excels in minimizing local prediction errors due to its smooth gradient properties, it does not explicitly impose constraints on global extrema or distributional consistency. As a result, its advantage becomes concentrated on point-wise error minimization (MAPE and MAE), rather than overall fitting performance.

Compared with earlier DRN results, where the Charbonnier loss dominates across most evaluation metrics, the PCA-DRN results reveal a clearer functional differentiation among loss functions. This indicates that feature transformation plays a critical role in shaping the effectiveness of loss formulations. In particular, PCA enhances the compatibility of structure-aware loss functions, such as the Penalized loss, by providing a cleaner and more interpretable feature space.

Among the remaining loss functions, the MAPE-based loss continues to show relatively competitive performance in point-wise error minimization but remains sensitive to small denominator values. The MSE loss again underperforms due to its excessive emphasis on large errors, while Huber, Log-Cosh, and Pseudo-Huber losses provide moderate performance without achieving clear advantages. The Smooth L1 loss exhibits limited effectiveness, consistent with the observations in the DRN setting.

Taken together, the results demonstrate that, under the PCA-DRN framework, different loss functions exhibit complementary strengths across different categories of evaluation metrics. While the Charbonnier loss is more suitable for minimizing point-wise prediction errors, the Penalized loss is more effective in capturing global load patterns and improving overall model consistency. These findings further suggest that the selection of loss functions should be jointly considered with feature representation strategies, rather than optimized independently, in STLF tasks.

Bootstrap-based significance analysis in statistics

To determine whether the observed differences in forecasting accuracy are statistically significant rather than arising from random fluctuations, a non-parametric bootstrap resampling approach is employed. Tables 7 and 8 summarize the statistical comparison results for the ISO-NE and MyPJ datasets, respectively. For each pair of models, the mean and standard deviation (SD) of the MAPE distribution, the mean difference, the corresponding 95% CI, and the p-value are reported.

In these comparisons, a positive mean difference indicates that the second model achieves lower prediction errors. The CI reflects the range within which the true performance difference is expected to lie with a given level of confidence, while the p-value quantifies the probability that the observed difference arises under the null hypothesis of no performance difference. A result is considered statistically significant when the CI does not include zero and the p-value is less than 0.05.

For the ISO-NE dataset, the comparison between DRN with Penalized loss and DRN with Charbonnier loss yields a mean difference of 0.000476, with a 95% CI of [0.000189, 0.000754] and a p-value of 0.000800. Since the CI excludes zero and the p-value is well below the significance threshold, the superiority of the Charbonnier loss is statistically significant. This result further confirms that the performance advantage observed under smooth and robust loss formulations is not due to random variation, but reflects a consistent improvement in model optimization and prediction accuracy.

For the MyPJ dataset, more pronounced differences are observed across both loss functions and model structures. When comparing DRN with Penalized loss and DRN with Charbonnier loss, the mean difference is 0.001601, with a 95% CI of [0.000930, 0.002262] and a p-value approximately equal to zero, indicating a statistically significant improvement achieved by the Charbonnier loss. A similar pattern is observed for the PCA-DRN model, where the comparison between Penalized loss and Charbonnier loss yields the same mean difference of 0.001601, with a 95% CI of [0.000925, 0.002283] and a p-value close to zero. These results demonstrate that the advantage of the Charbonnier loss in minimizing point-forecast error metrics remains statistically robust across different model configurations.

In addition, cross-model comparisons further highlight the effectiveness of the PCA-DRN framework. When comparing DRN (Penalized loss) with PCA-DRN (Penalized loss), a mean difference of 0.002520 is obtained, with a 95% CI of [0.001416, 0.003690] and a p-value of 0.000100. Similarly, for the Charbonnier loss, the comparison between DRN and PCA-DRN also produces a mean difference of 0.002520, with a 95% CI of [0.001599, 0.003469] and a p-value approximately equal to zero. In both cases, the CIs exclude zero and the p-values indicate strong statistical significance, confirming that the PCA-DRN model consistently outperforms the original DRN regardless of the loss function used.

Overall, the bootstrap-based statistical analysis provides strong evidence that the observed performance differences are not attributable to random variation but reflect genuine improvements in model capability. In particular, the superiority of the Charbonnier loss in point-wise prediction accuracy and the consistent advantage of PCA-based feature transformation are both statistically validated. These findings reinforce the conclusions drawn from the experimental results and highlight the importance of jointly considering loss function design and feature representation in enhancing the performance of DRN-based STLF models.

Key findings and analysis

Based on the comprehensive experimental results and statistical analysis, several key findings can be drawn regarding the role of loss functions and feature representation in DRN-based STLF.

First, the choice of loss function is shown to have a substantial impact on both optimization behavior and forecasting performance. Across both datasets, loss functions characterized by smooth gradient properties and robustness to outliers, particularly the Charbonnier loss, generally achieve better performance in point-forecast error metrics. This highlights the importance of stable gradient propagation and balanced error handling in deep residual learning, especially under nonlinear and noisy data conditions commonly encountered in STLF tasks. More importantly, these results demonstrate that the inherent properties of loss functions, including gradient smoothness and sensitivity to outliers, directly influence model behavior in terms of convergence stability and predictive accuracy. Loss functions with smooth gradients tend to facilitate more stable optimization by reducing oscillations during training, whereas robust formulations help mitigate the impact of extreme deviations. In addition, the effectiveness of different loss functions is closely related to data characteristics, where datasets with higher variability and irregular fluctuations tend to benefit more from loss formulations that balance robustness and gradient continuity.

Second, the effectiveness of different loss functions is found to be strongly dependent on dataset characteristics. For the ISO-NE dataset, which exhibits relatively structured seasonal patterns and smoother temporal variations, loss functions with stable and smooth optimization behavior, such as Charbonnier and Log-Cosh, demonstrate clear advantages across all evaluation metrics. In contrast, for the MyPJ dataset, where load patterns are more irregular and influenced by short-term weather fluctuations, the performance differences among loss functions become more pronounced. Under such conditions, robustness to noise and adaptability to local variations play a more critical role in determining forecasting accuracy.

Third, the introduction of PCA-based feature transformation significantly enhances model performance in data-rich environments. The PCA-DRN framework consistently outperforms the original DRN across all evaluated loss functions, demonstrating that reducing redundancy and extracting orthogonal feature representations can effectively improve both learning efficiency and generalization capability. This improvement is particularly evident in the MyPJ dataset, where multiple meteorological variables are involved and feature redundancy is more prominent. This further indicates that PCA not only improves feature representation but also reshapes the input data distribution by reducing noise and inter-feature correlations, thereby influencing how different loss functions interact with the data and contributing to the observed variation in their effectiveness.

Moreover, the interaction between loss functions and feature representation reveals a complementary relationship. While the Charbonnier loss excels in minimizing point-wise prediction errors, the Penalized loss demonstrates stronger performance in squared-error-based and correlation-oriented metrics under the PCA-DRN framework. This suggests that, when the input feature space is refined through PCA, structure-aware loss formulations that incorporate global constraints can better exploit the underlying data distribution, leading to improved overall model consistency.

In addition, the distributional analysis based on boxplot visualizations further supports this observation. Across the DRN and PCA-DRN frameworks, loss functions with smoother gradient properties, particularly the Charbonnier loss, consistently exhibit more concentrated error distributions with reduced variability. This indicates not only lower average prediction errors but also improved stability and robustness during optimization. Furthermore, under the PCA-DRN framework, a noticeable reduction in dispersion is observed across most loss functions, suggesting that PCA-based feature transformation contributes to a more structured input space and more consistent training behavior.

In addition, the bootstrap-based statistical analysis provides strong evidence supporting the reliability of the observed performance differences. The superiority of the Charbonnier loss over the Penalized loss, as well as the consistent advantage of PCA-DRN over DRN, are both confirmed to be statistically significant. This indicates that the improvements are not attributable to random variation but reflect genuine differences in model capability and optimization effectiveness.

Taken together, these findings highlight that the selection of loss functions in DRN-based STLF models should not be treated as an isolated design choice. Instead, it should be jointly considered with dataset characteristics and feature representation strategies. Loss functions with smooth gradients and robust error handling are generally more suitable for achieving stable and accurate predictions, while structure-aware formulations may provide additional benefits when combined with refined feature spaces. In addition, alternative strategies such as ensemble and aggregation-based approaches offer complementary pathways for improving forecasting performance by combining multiple predictors to enhance robustness and generalization. Therefore, a holistic approach that integrates loss function design, feature engineering, model architecture, and prediction combination strategies is essential for developing reliable and high-performance STLF models under diverse real-world conditions.

Conclusion

This study presents a systematic investigation into the role of loss functions in DRN-based STLF, with a particular focus on their impact on optimization behavior, forecasting accuracy, and model stability under diverse data conditions. These findings further suggest that loss functions implicitly influence how DRN-based models adapt to non-stationary patterns and distributional variations in load data. In this sense, the selection of an appropriate loss function can be viewed as a complementary strategy to enhance model robustness under dynamic real-world conditions. By evaluating traditional error-based, robust, and Penalized loss functions within a unified experimental framework, this work provides a comprehensive analysis of how different loss formulations influence the performance of residual learning models.

The experimental results demonstrate that loss functions with smooth gradient properties and strong robustness to outliers, particularly the Charbonnier loss, generally achieve improved performance in point-forecast error metrics across both temperate and tropical datasets. This highlights the importance of stable gradient propagation and balanced error handling in deep residual learning, especially when dealing with nonlinear and noisy load data. At the same time, the Penalized loss, originally developed for DRN-based STLF, exhibits competitive performance by incorporating range-aware constraints, showing advantages in capturing global load characteristics under certain conditions.

Furthermore, the integration of PCA into the DRN framework significantly enhances forecasting performance in data-rich environments. The PCA-DRN model consistently outperforms the original DRN across all evaluated loss functions, confirming that reducing feature redundancy and transforming meteorological variables into a compact and orthogonal representation can effectively improve both learning efficiency and generalization capability. In particular, the combination of PCA and structure-aware loss formulations reveals a complementary relationship, where refined feature representations enable more effective utilization of loss-driven optimization strategies.

The bootstrap-based statistical analysis further validates the reliability of these findings. The observed performance improvements, including the superiority of the Charbonnier loss and the consistent advantage of PCA-DRN over DRN, are confirmed to be statistically significant, indicating that these improvements are not due to random variation but reflect genuine differences in model capability.

Overall, this study highlights that the selection of loss functions in DRN-based STLF models is inherently dependent on both dataset characteristics and feature representation strategies. Rather than treating loss functions as independent design components, a more effective approach is to jointly consider loss formulation, feature engineering, and model architecture. Such an integrated design paradigm is essential for achieving robust and accurate forecasting performance in real-world power systems.

For future work, several directions can be explored. First, the development of adaptive or data-driven loss functions tailored to specific load characteristics may further enhance model performance. Second, nonlinear dimensionality reduction techniques could be investigated as alternatives to PCA to better preserve complex feature interactions. In addition, a more fine-grained diagnostic analysis of forecasting errors will be conducted, including stratified evaluation under specific operating conditions such as peak load periods, troughs, holidays, and extreme weather events. Such analysis is expected to provide deeper insights into when and where different loss functions perform better, thereby further improving the interpretability and practical applicability of DRN-based STLF models. Finally, extending the proposed framework to probabilistic forecasting and uncertainty quantification would provide additional practical value for real-world power system operations.

Data availability

The datasets analyzed during this study are publicly available from official sources, subject to relevant access conditions or registration requirements. The ISO-NE dataset is accessible via the New England Independent

System Operator (https://www.iso-ne.com/isoexpress/web/reports/load-and-demand). The MyPJ load data are available from the Grid System Operator (GSO), Malaysia (https://www.gso.org.my/SystemData/SystemDemand.aspx), and the corresponding meteorological data can be obtained from the Malaysian Meteorological Department (https://www.met.gov.my/). Processed data supporting the findings of this study are available from the corresponding author upon reasonable request.

Received: 25 March 2026; Accepted: 30 April 2026
Published online: 06 May 2026

References

1. Zhou, S. et al. UniLF: A novel short-term load forecasting model uniformly considering various features from multivariate load data. *Sci. Rep.* **15**, 4282. https://doi.org/10.1038/s41598-025-88566-4 (2025).
2. Hasan, M. et al. A state-of-the-art comparative review of load forecasting methods: Characteristics, perspectives, and applications. *Energy Convers. Manag. X* **26**, 100922. https://doi.org/10.1016/j.ecmx.2025.100922 (2025).
3. Li, K. et al. Short-term load forecasting for electricity spot markets across different seasons based on a hybrid VMD-LSTM-random forest model. *Energies* **18**, 6097. https://doi.org/10.3390/en18236097 (2025).
4. Ahmad, F. A., Liu, J., Hashim, F. & Samsudin, K. Short-term load forecasting utilizing a combination model: A brief review. *Int. J. Technol.* **15**, 121–129. https://doi.org/10.14716/ijtech.v15i1.5543 (2024).
5. Liu, J., Ahmad, F. A., Samsudin, K., Hashim, F. & Ab Kadir, M. Z. A. A comprehensive review of deep residual networks for short-term load forecasting. *Pertanika J. Sci. Technol.* **34**, 1 (2026).
6. Hobbs, B. F. et al. Analysis of the value for unit commitment of improved load forecasts. *IEEE Trans. Power Syst.* **14**, 1342–1348. https://doi.org/10.1109/59.801894 (1999).
7. Wang, Y. & Wu, L. Improving economic values of day-ahead load forecasts to real-time power system operations. *IET Gener. Transm. Distrib.* **11**, 4238–4247. https://doi.org/10.1049/iet-gtd.2017.0517 (2017).
8. Akhtar, S. et al. Short-term load forecasting models: A review of challenges, progress, and the road ahead. *Energies* **16**, 4060. https://doi.org/10.3390/en16104060 (2023).
9. Hippert, H. S., Pedreira, C. E. & Souza, R. C. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Trans. Power Syst.* **16**, 44–55. https://doi.org/10.1109/59.910780 (2001).
10. Ceperic, E., Ceperic, V. & Baric, A. A strategy for short-term load forecasting by support vector regression machines. *IEEE Trans. Power Syst.* **28**, 4356–4364. https://doi.org/10.1109/TPWRS.2013.2269803 (2013).
11. Kuster, C., Rezgui, Y. & Mourshed, M. Electrical load forecasting models: A critical systematic review. *Sustain. Cities Soc.* **35**, 257–270. https://doi.org/10.1016/j.scs.2017.08.009 (2017).
12. Cecati, C., Kolbusz, J., Różycki, P., Siano, P. & Wilamowski, B. M. A novel RBF training algorithm for short-term electric load forecasting and comparative studies. *IEEE Trans. Ind. Electron.* **62**, 6519–6529. https://doi.org/10.1109/TIE.2015.2424399 (2015).
13. Chen, Y. et al. Short-term load forecasting: Similar day-based wavelet neural networks. *IEEE Trans. Power Syst.* **25**, 322–330. https://doi.org/10.1109/TPWRS.2009.2030426 (2009).
14. Zhao, Y., Luh, P. B., Bomgardner, C. & Beerel, G. H. Short-term load forecasting: Multi-level wavelet neural networks with holiday corrections. In *Proc. IEEE Power Energy Soc. Gen. Meeting* 1–7. https://doi.org/10.1109/PES.2009.5275304 (2009).
15. Eren, Y. & Küçükdemiral, İ. A comprehensive review on deep learning approaches for short-term load forecasting. *Renew. Sustain. Energy Rev.* **189**, 114031. https://doi.org/10.1016/j.rser.2023.114031 (2024).
16. Li, L., Ota, K. & Dong, M. Everything is image: CNN-based short-term electrical load forecasting for smart grid. In *Proc. 14th Int. Symp. Pervasive Syst. Algorithms Netw. (ISPAN-FCST-ISCC)* 344–351. https://doi.org/10.1109/ISPAN-FCST-ISCC.2017.78 (2017).
17. Jurado, M., Samper, M. & Rosés, R. An improved encoder-decoder-based CNN model for probabilistic short-term load and PV forecasting. *Electr. Power Syst. Res.* **217**, 109153. https://doi.org/10.1016/j.epsr.2023.109153 (2023).
18. Tang, X., Chen, H., Xiang, W., Yang, J. & Zou, M. Short-term load forecasting using channel and temporal attention based temporal convolutional network. *Electr. Power Syst. Res.* **205**, 107761. https://doi.org/10.1016/j.epsr.2021.107761 (2022).
19. Liu, M., Xia, C., Xia, Y., Deng, S. & Wang, Y. TDCN: A novel temporal depthwise convolutional network for short-term load forecasting. *Int. J. Electr. Power Energy Syst.* **165**, 110512. https://doi.org/10.1016/j.ijepes.2025.110512 (2025).
20. Narayan, A. & Hipel, K. W. Long short-term memory networks for short-term electric load forecasting. In *Proc. IEEE Int. Conf. Syst. Man Cybern.* 2573–2578. https://doi.org/10.1109/SMC.2017.8123012 (2017).
21. Bento, P., Pombo, J., Mariano, S. & Calado, M. R. Short-term load forecasting using optimized LSTM networks via improved bat algorithm. In *Proc. Int. Conf. Intell. Syst.* 351–357. https://doi.org/10.1109/IS.2018.8710498 (2018).
22. Ran, P., Dong, K., Liu, X. & Wang, J. Short-term load forecasting based on CEEMDAN and transformer. *Electr. Power Syst. Res.* **214**, 108885. https://doi.org/10.1016/j.epsr.2022.108885 (2023).
23. Jiang, B., Liu, Y., Geng, H., Zeng, H. & Ding, J. A transformer-based method with wide attention range for enhanced short-term load forecasting. In *Proc. 4th Int. Conf. Smart Power Internet Energy Syst.* 1684–1690 (2022).
24. Li, S., Zhang, W. & Wang, P. TS2ARCformer: A multi-dimensional time series forecasting framework for short-term load prediction. *Energies* **16**, 5825. https://doi.org/10.3390/en16155825 (2023).
25. Chen, Y., Wang, Y., Liu, X. & Huang, J. Short-term load forecasting for industrial users based on transformer–LSTM hybrid model. In *Proc. IEEE 5th Int. Electr. Energy Conf. (CIEEC)* 2470–2475. https://doi.org/10.1109/CIEEC54735.2022.9846658 (2022).
26. Guo, W., Liu, S., Weng, L. & Liang, X. Power grid load forecasting using a CNN–LSTM network based on a multi-modal attention mechanism. *Appl. Sci.* **15**, 2435. https://doi.org/10.3390/app15052435 (2025).
27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 770–778. https://doi.org/10.48550/arXiv.1512.03385 (2016).
28. Chen, K. et al. Short-term load forecasting with deep residual networks. *IEEE Trans. Smart Grid* **10**, 3943–3952. https://doi.org/10.1109/TSG.2018.2844307 (2018).
29. Kondaiah, V. Y. & Saravanan, B. Short-term load forecasting with deep learning. In *Proc. Innov. Power Adv. Comput. Technol. (i-PACT)* 1–5. https://doi.org/10.1109/i-PACT52855.2021.9696634 (2021).
30. Xu, Q., Yang, X. & Huang, X. Ensemble residual networks for short-term load forecasting. *IEEE Access* **8**, 64750–64759. https://doi.org/10.1109/ACCESS.2020.2984722 (2020).
31. Chen, W., Han, G., Zhu, H., Liao, L. & Zhao, W. Deep ResNet-based ensemble model for short-term load forecasting in protection system of smart grid. *Sustainability* **14**, 16894. https://doi.org/10.3390/su142416894 (2022).
32. Kondaiah, V. Y. & Saravanan, B. A modified deep residual network for short-term load forecasting. *Front. Energy Res.* **10**, 1038819. https://doi.org/10.3389/fenrg.2022.1038819 (2022).
33. Sheng, Z., Wang, H., Chen, G., Zhou, B. & Sun, J. Convolutional residual network for short-term load forecasting. *Appl. Intell.* **51**, 2485–2499. https://doi.org/10.1007/s10489-020-01932-9 (2021).
34. Ding, A., Liu, T. & Zou, X. Integration of ensemble GoogLeNet and modified deep residual networks for short-term load forecasting. *Electronics* **10**, 2455. https://doi.org/10.3390/electronics10202455 (2021).
35. Liu, J. et al. Deep residual networks with convolutional feature extraction for short-term load forecasting. *Sci. Rep.* **16**, 6382. https://doi.org/10.1038/s41598-026-35410-y (2026).

36. Li, H., Zhang, P. & Li, C. Short-term load forecasting for distribution substations based on residual neural networks and long short-term memory neural networks with attention mechanism. In *J. Phys. Conf. Ser.* Vol 2030, 012087. https://doi.org/10.1088/1742-6596/2030/1/012087 (2021).
37. Tian, Y., Yu, S., Wen, M., Zhang, K. & Chen, Y. Short-term load forecasting scheme based on improved deep residual network and LSTM. In *Proc. CIRED Berlin Workshop* 117–120. https://doi.org/10.1049/oap-cired.2021.0257 (2020).
38. Sheng, Z., An, Z., Wang, H., Chen, G. & Tian, K. Residual LSTM-based short-term load forecasting. *Appl. Soft Comput.* **144**, 110461. https://doi.org/10.1016/j.asoc.2023.110461 (2023).
39. Liu, J., Ahmad, F. A., Samsudin, K., Hashim, F. & Ab Kadir, M. Z. A. Optimizing deep residual networks for short-term load forecasting with multidimensional weather data and principal component analysis. *IEEE Access* **13**, 158614–158631. https://doi.org/10.1109/ACCESS.2025.3607330 (2025).
40. Zhao, P. et al. A review on deep learning-based electrical load forecasting: From perspectives of learning paradigms and foundation models. *J. Mod. Power Syst. Clean Energy* https://doi.org/10.35833/MPCE.2025.000740 (2026).
41. Incremona, A., De Nicolao, G., Fusco, F., Eck, B. J. & Tirupathi, S. Aggregation of nonlinearly enhanced experts with application to electricity load forecasting. *Appl. Soft Comput.* **112**, 107857. https://doi.org/10.1016/j.asoc.2021.107857 (2021).
42. Liu, P. et al. TimeBridge: Non-stationarity matters for long-term time series forecasting. arXiv. https://doi.org/10.48550/arXiv.2410.04442 (2024).
43. Lu, J. et al. Towards non-stationary time series forecasting with temporal stabilization and frequency differencing. In *Proc. AAAI Conf. Artif. Intell.* Vol 40, 24070–24078. https://doi.org/10.1609/aaai.v40i29.39585 (2026).
44. Terven, J. et al. A comprehensive survey of loss functions and metrics in deep learning. *Artif. Intell. Rev.* **58**, 195. https://doi.org/10.1007/s10462-025-11198-7 (2025).
45. Zhao, P. et al. Geometric loss-enabled complex neural network for multi-energy load forecasting in integrated energy systems. *IEEE Trans. Power Syst.* **39**, 5659–5671. https://doi.org/10.1109/TPWRS.2023.3345328 (2023).
46. Dwivedi, P., Islam, B. & Kajal, M. Smooth gradient loss: A loss function for gradient regularization in deep learning optimization. *J. Supercomput.* **81**, 1457. https://doi.org/10.1007/s11227-025-07954-9 (2025).
47. Liu, C., Zhu, L. & Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Appl. Comput. Harmon. Anal.* **59**, 85–116. https://doi.org/10.1016/j.acha.2021.12.009 (2022).
48. Achour, E. M., Malgouyres, F. & Gerchinovitz, S. The loss landscape of deep linear neural networks: A second-order analysis. *J. Mach. Learn. Res.* **25**, 1–76 (2024).
49. Li, S. et al. Short-term load forecasting system based on sliding fuzzy granulation and equilibrium optimizer. *Appl. Intell.* **53**, 21606–21640. https://doi.org/10.1007/s10489-023-04599-0 (2023).
50. Hong, T., Pinson, P. & Fan, S. Global energy forecasting competition 2012. *Int. J. Forecast.* **30**, 357–363. https://doi.org/10.1016/j.ijforecast.2013.07.001 (2014).
51. Li, C., Liu, K. & Liu, S. A survey of loss functions in deep learning. *Mathematics* **13**, 2417. https://doi.org/10.3390/math13152417 (2025).
52. Wang, Q. et al. A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* **9**, 187–212. https://doi.org/10.1007/s40745-020-00253-5 (2022).
53. Ly, A. & Gong, P. Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning. *Nat. Commun.* **16**, 3252. https://doi.org/10.1038/s41467-025-58532-9 (2025).
54. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001 (2006).
55. Qi, J., Du, J., Siniscalchi, S. M., Ma, X. & Lee, C. H. On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal. Process. Lett.* **27**, 1485–1489. https://doi.org/10.1109/LSP.2020.3016837 (2020).
56. Jadon, A., Patil, A. & Jadon, S. A comprehensive survey of regression-based loss functions for time series forecasting. In *Data Management, Analytics and Innovation. ICDMAI 2024. Lecture Notes in Networks and Systems* Vol 998 (eds Sharma, N., Goje, A. C., Chakrabarti, A. & Bruckstein, A. M.) (Springer, 2024). https://doi.org/10.1007/978-981-97-3245-6_9.
57. Aravkin, A. Y., Burke, J. V. & Pillonetto, G. Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *J. Mach. Learn. Res.* **14**, 2689–2728 (2013).
58. Noel, M. M., Banerjee, A., Oswal, Y. & Muthiah-Nakarajan, V. Alternate loss functions for classification and robust regression can improve the accuracy of artificial neural networks. arXiv preprint. https://doi.org/10.48550/arXiv.2303.09935 (2023).
59. Shen, Y. & Xia, D. Quantile and pseudo-huber tensor decomposition. arXiv preprint. https://doi.org/10.48550/arXiv.2309.02698 (2023).
60. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* 1440–1448. https://doi.org/10.1109/ICCV.2015.169 (2015).
61. Barron, J. T. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4331–4339. https://doi.org/10.1109/CVPR.2019.00446 (2019).
62. Fan, C., Chen, M., Wang, X., Wang, J. & Huang, B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front. Energy Res.* **9**, 652801. https://doi.org/10.3389/fenrg.2021.652801 (2021).
63. Ortiz, B. L. et al. Data preprocessing techniques for AI and machine learning readiness: Scoping review of wearable sensor data in cancer care. *JMIR mHealth uHealth* **12**, e59587. https://doi.org/10.2196/59587 (2024).
64. Huang, G. et al. Snapshot ensembles: Train 1, get m for free. arXiv. https://doi.org/10.48550/arXiv.1704.00109 (2017).
65. Khoshkangini, R., Tajgardan, M., Lundström, J., Rabbani, M. & Tegnered, D. A snapshot-stacked ensemble and optimization approach for vehicle breakdown prediction. *Sensors* **23**, 5621. https://doi.org/10.3390/s23125621 (2023).
66. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint. https://doi.org/10.48550/arXiv.1412.6980 (2014).
67. Johnston, M. G. & Faulkner, C. A bootstrap approach is a superior statistical method for the comparison of non-normal data with differing variances. *New Phytol.* **230**, 23–26. https://doi.org/10.1111/nph.17159 (2021).

Author contributions

J.L. and F.A.A. contributed to the conceptualization of the study. J.L. developed the methodology and conducted the investigation with F.A.A. J.L. prepared the original draft. J.L., F.A.A., K.S., F.H., and M.Z.A.A.K. contributed to the review and editing of the manuscript. F.A.A., K.S., F.H., and M.Z.A.A.K. provided supervision. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-52040-6.

Correspondence and requests for materials should be addressed to F.A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.