

Article

Multi-Horizon Short-Term Load Forecasting for Electric Power System Dispatch Optimization: A Bidirectional Gated Recurrent Unit Framework with Forecast-Origin Prior Integration and Aligned Evaluation

Feng Cao ¹, Jun-Hao Zhang ² and Gui-Hong Jia ^{3,*}

¹ Engineering Training and Innovation Center, North China Electric Power University, Baoding 071003, China; feng.cao@ncepu.edu.cn

² School of Electricity Transmission Engineering, North China Electric Power University, Baoding 071003, China; junhao.zhang@ncepu.edu.cn

³ Department of Mechanical Engineering, North China Electric Power University, Baoding 071003, China

* Correspondence: guihong.jia@ncepu.edu.cn

Abstract

Accurate short-term load forecasting is essential for power system operations, including unit commitment, economic dispatch, and reserve scheduling. Regional load series exhibit temporal variation and horizon-dependent dynamics that make multi-horizon forecasting challenging. This paper proposes a bidirectional gated recurrent unit-based multi-head joint forecasting framework (EA-BiGRU-Multi) for one-step, six-step, and 24-step-ahead electricity load prediction. A shared recurrent encoder captures temporal dependencies from historical observations, while horizon-specific forecasting heads produce separate predictions for each lead time. A terminal-anchor injection strategy incorporates day-ahead demand information only at the forecast origin, preserving the distinction between historical observations and known-future priors. Feature-channel recalibration adjusts heterogeneous input contributions before temporal encoding. Experiments on three regional load zones from a regional independent system operator in the northeastern United States show that the proposed model achieves competitive multi-horizon performance. In the Southeast Massachusetts six-step-ahead case, the proposed model increases the mean coefficient of determination from 0.7604 to 0.8346 and reduces the mean absolute error from 100.44 MW to 79.07 MW compared with the linear baseline. Ablation analysis confirms that the terminal-anchor prior is essential for medium- and long-horizon accuracy. The results indicate that model suitability depends jointly on forecast horizon, regional load characteristics, and availability of forecast-origin demand information.

Keywords: short-term load forecasting; power system optimization; multi-horizon forecasting algorithm; bidirectional GRU; terminal-anchor injection; day-ahead demand integration; operational decision support; smart grid



Academic Editor: Jiaqiang E

Received: 22 June 2026

Revised: 5 July 2026

Accepted: 9 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Accurate electricity load forecasting is a fundamental requirement for secure, economical, and flexible power system operation. With the increasing penetration of renewable generation, distributed energy resources, flexible demand, and market-based dispatch, system operators are required to make short-term decisions under more uncertain and variable operating conditions. Short-term electricity load forecasting (STLF), which generally

covers prediction from the next hour to the next day, provides essential information for unit commitment, economic dispatch, reserve scheduling, demand response, market clearing, and reliability assessment [1–3]. In practice, different lead times support different operational decisions: near-term forecasts are closely related to real-time balancing, whereas several-hour and day-ahead forecasts are more relevant to reserve scheduling, market preparation, and operational planning [1,2]. Inaccurate forecasts may lead to excessive reserve allocation, inefficient generation scheduling, increased operating costs, and reduced operational reliability. Therefore, improving the accuracy and reliability of STLF is not only a data-driven modeling task, but also an important requirement for real-time power system operation and market management.

Regional load forecasting is particularly challenging because regional demand is jointly affected by weather conditions, calendar effects, local consumption habits, industrial structure, and seasonal demand regimes [1,4]. Compared with system-level aggregated load, regional load series often exhibit stronger local fluctuations, sharper seasonal transitions, and more heterogeneous temporal patterns. These characteristics make the forecasting problem nonlinear, non-stationary, and dependent on both forecast horizon and regional demand behavior. A model that performs well at one horizon may not necessarily remain effective at another horizon, especially when the dominant temporal patterns and available prior information change with lead time. This horizon-dependent behavior is particularly important in regional STLF, where local demand regimes may amplify the difference between short-term extrapolation and day-ahead prediction.

A wide range of methods has been developed for electricity load forecasting, from early neural-network-based STLF studies to statistical, machine-learning, and deep-learning approaches [5]. Classical statistical models, such as autoregressive integrated moving average (ARIMA), exponential smoothing, and regression-based models, have been widely used because of their simple structures, clear assumptions, and good interpretability [6–8]. These methods are effective when load series exhibit relatively stable patterns and the relationship between historical load and future demand can be approximated by linear or low-order temporal structures. Semi-parametric and additive models further incorporate weather, calendar, and seasonal variables to improve explanatory capability [8]. In representative studies, Taylor [7] reported that double seasonal exponential smoothing achieved mean absolute percentage error (MAPE) below 2% for half-hourly electricity demand forecasting, demonstrating the effectiveness of parsimonious statistical structures under periodic load patterns. Fan and Hyndman [8] showed that a semi-parametric additive model attained leading accuracy in the Global Energy Forecasting Competition 2012, with normalized error approximately 5% lower than benchmark regression methods. However, statistical models usually rely on predefined functional forms and may have limited ability to represent nonlinear interactions, abrupt regime shifts, and complex multi-source dependencies in regional load data.

Machine learning methods have been introduced to overcome some of these limitations. Support vector regression, radial basis function neural networks, random forests, gradient boosting, extreme gradient boosting, and Gaussian process based methods have been applied to STLF and related energy forecasting tasks [9–12]. Compared with classical statistical models, these methods provide stronger nonlinear fitting ability and can handle a broader range of explanatory variables. Ko and Lee [9] reported that their SVR-based approach with dual extended Kalman filtering achieved a MAPE of approximately 1.32% on the tested electricity load dataset. Lloyd [12] demonstrated that gradient boosting combined with Gaussian processes achieved MAPE of approximately 4.1% in the hierarchical load forecasting track of GEFCom2012. Nevertheless, their effectiveness often depends on manually designed features, model-specific preprocessing, and careful pa-

parameter tuning. When the forecasting task involves multiple regions, multiple horizons, and changing seasonal regimes, the transferability of such feature-dependent models may become limited.

Deep learning methods have become increasingly important in load forecasting because they can learn temporal representations directly from multivariate sequences. Recurrent neural networks, especially long short-term memory (LSTM) networks, gated recurrent units (GRUs), and bidirectional recurrent structures, have been widely used to model sequential dependencies in load and energy time series [13–15]. Convolutional neural networks and temporal convolutional networks extract local patterns and long-range temporal features through convolutional filters and dilated structures [16,17]. Attention-based and hybrid models further improve the representation of complex temporal patterns by combining recurrent units, convolutional modules, temporal attention, and feature selection mechanisms [17–20]. More recently, multi-horizon and long-sequence forecasting models, such as DeepAR, Temporal Fusion Transformer (TFT), Informer, PatchTST, and DLinear, have shown that different architectural assumptions may lead to different advantages under different forecasting conditions [21–25]. Zhou et al. [23] showed that Informer reduced prediction error by approximately 25% compared with the canonical Transformer on the Electricity Consuming Load dataset. Zeng et al. [25] demonstrated that DLinear, a single-layer linear model, achieved mean squared error competitive with or lower than several Transformer-based architectures across multiple long-term forecasting benchmarks, with reductions exceeding 10% in some settings. In particular, the strong performance of lightweight linear structures in some time-series forecasting tasks indicates that model complexity alone is not a sufficient indicator of forecasting accuracy [3,25]. The more relevant question is under what horizon, regional regime, and information setting a given inductive bias becomes useful.

Despite these advances, several issues remain insufficiently addressed in regional multi-horizon STLF. Many forecasting pipelines are designed for a single lead time or train separate models for different horizons [20,21,26,27]. This setting is straightforward and easy to implement, but it does not explicitly share temporal representations across short-, medium-, and day-ahead forecasting tasks. In practice, the same historical load window may contain information relevant to multiple lead times, although the importance of different temporal patterns may vary by horizon. Training separate models may increase maintenance cost and may fail to exploit common temporal structures among related forecasting tasks. Multi-head joint forecasting provides a natural way to share part of the temporal representation while allowing different horizons to retain their own prediction mappings [22,26,27]. However, whether such a structure is beneficial should be examined under specific regional and horizon settings, rather than assumed from the architecture itself.

In addition, known-future information is often available in power system operation but is not always incorporated in a forecast-origin-consistent way. In regional electricity markets, day-ahead demand or related scheduled demand information may be available before real-time operation and can provide a useful demand-level prior for medium- and long-horizon forecasting [1,2,22]. Specifically, independent system operators such as ISO New England publish hourly day-ahead cleared demand values for each load zone after the day-ahead market clearing process, which typically concludes by midday on the day before delivery [28]. These day-ahead cleared values represent the market's expectation of zonal demand for each hour of the next operating day and are publicly available to market participants and system operators. Because day-ahead demand is a market outcome rather than a direct measurement of real-time consumption, it contains forecast error relative to actual real-time load. In the dataset used in this study, the MAPE

between day-ahead demand and real-time demand is approximately 2.87% across all evaluated zones, with zone-level values of 2.41% for Connecticut, 3.14% for Southeast Massachusetts, and 3.06% for Vermont. This error level indicates that day-ahead demand provides a reasonably informative but imperfect demand-level prior, which motivates the noise sensitivity analysis presented in Section 3.10. Existing multi-horizon forecasting models have considered known-future covariates as an important source of information [22]. However, when such information is represented within a multivariate historical input window, its temporal placement becomes critical. Directly broadcasting a known-future demand prior across all historical timestamps implicitly converts a forecast-origin prior into a pseudo-historical time-varying covariate. This changes the temporal meaning of the input sequence and may blur the distinction between observed historical variables and information available at the forecast origin. Alternative approaches include appending the known-future value as a static input to the final hidden state, concatenating it as a separate decoder input, or encoding it through a dedicated side network. Terminal-anchor injection is adopted in this study because it requires no additional decoder module or architectural modification, retains the sequential structure of the input, and localizes the prior at the forecast-origin position where it is operationally meaningful. A controlled comparison with naive broadcasting is provided in Section 3.8 to empirically validate this design choice. A more controlled input design is therefore needed so that known-future demand information can be used as an operational prior without being treated as a historical feature observed at every past timestamp.

Evaluation protocols in rolling-origin forecasting may introduce hidden comparison bias if test-window construction and forecast-origin alignment are not carefully controlled [3,29,30]. In multi-horizon forecasting, different lead times can produce different effective test-window coverage because the target timestamp shifts with the prediction horizon. If models or horizons are evaluated on non-identical timestamp sets, the reported performance may partly reflect differences in test samples rather than genuine forecasting ability. This issue becomes more important under seasonal load variation, regional distribution shift, or short evaluation windows. Previous forecasting studies have emphasized that evaluation design and out-of-sample testing procedures can substantially affect the interpretation of model performance [29,30]. Therefore, an aligned evaluation protocol is needed to compare different models and horizons on a common timestamp intersection. Such a protocol is not intended to inflate model performance, but to provide a more consistent basis for interpreting horizon-dependent and region-dependent forecasting results.

The relative advantage of a forecasting model should not be interpreted only from an overall average score. Forecasting performance is often affected by data characteristics, forecast horizon, information availability, and evaluation design [3]. In regional STLF, the preferred model may change across lead times and zones. A simple linear model may be sufficient for near-term extrapolation when the recent load trajectory is highly informative [25]. A recurrent or attention-based model may be more useful when medium- or long-horizon prediction requires nonlinear temporal representation and reliable known-future priors [13–15,22]. A patch-based Transformer may show different behavior when only historical inputs are available [24]. Therefore, a meaningful comparison should examine how model suitability changes under different regional, horizon, and information settings, instead of only reporting whether a proposed model achieves the best average result.

The novelty of this study lies not in proposing entirely new individual components, since channel recalibration, gated recurrent units, and attention pooling are established techniques in the literature. Instead, the contribution centers on the coordinated integration of these components with a forecast-origin-consistent day-ahead demand injection strategy,

and on the systematic evaluation that reveals how model suitability varies across horizons, regions, and information settings. The terminal-anchor injection mechanism, which places known-future demand information exclusively at the final frame of the input window, provides a simple and reproducible approach for incorporating forecast-origin priors without modifying the temporal semantics of the historical input sequence. The aligned evaluation protocol and the multi-faceted ablation design further contribute to methodological rigor in multi-horizon load forecasting evaluation.

To examine these issues, this paper develops EA-BiGRU-Multi, a multi-head joint forecasting framework for regional multi-horizon STLF. The framework combines feature-channel recalibration, shared BiGRU-based temporal encoding, horizon-specific forecasting heads, and terminal-anchor injection of day-ahead demand information. The proposed model predicts 1-, 6-, and 24-step-ahead load within a unified structure. It is important to note that the model predicts three discrete forecast horizons rather than producing a full step-by-step trajectory from $t + 1$ to $t + 24$, which distinguishes the present approach from sequence-to-sequence or full-horizon forecasting formulations. The lightweight feature-channel recalibration component is used for feature-channel recalibration rather than as a new attention family, following the general idea of channel-wise feature recalibration in attention-based representation learning [31,32]. The shared BiGRU encoder learns temporal representations from the observed input window, while the horizon-specific output heads allow different lead times to retain separate prediction mappings. To incorporate day-ahead demand information, terminal-anchor injection places the known-future demand prior only at the final frame of the input window. In this way, the prior is associated with the forecast origin and is not treated as a time-varying historical feature. In addition, an aligned evaluation protocol is adopted so that all compared models are evaluated on the same timestamp intersection.

Rather than framing the comparison as a single architecture-level ranking, this study examines how model suitability changes with forecast horizon, regional load characteristics, and known-future information availability. To support this objective, experiments are conducted on three ISO New England load zones using 2024 hourly data. The proposed framework is compared with representative recurrent, linear, and patch-based forecasting models, including DLinear and PatchTST as recent time-series forecasting baselines [24,25]. Additional analyses are conducted to clarify the role of the day-ahead demand prior, the effect of terminal-anchor injection, and the conditions under which the proposed framework is most useful.

The main contributions of this paper are summarized as follows:

1. A multi-head joint forecasting framework, EA-BiGRU-Multi, is proposed for regional multi-horizon STLF. By combining shared BiGRU-based temporal encoding and horizon-specific forecasting heads, the framework learns common temporal representations while preserving horizon-dependent prediction behavior.
2. A terminal-anchor injection strategy is developed for day-ahead demand integration. By placing the known-future demand prior only at the final frame of the input window, the strategy associates the prior with the forecast origin and avoids assigning future demand information to historical timestamps.
3. An aligned evaluation protocol is adopted for rolling-origin multi-horizon forecasting. By comparing all models on a common timestamp intersection, the protocol reduces test-window coverage bias and provides a more consistent basis for model comparison across horizons.
4. The empirical study provides a scenario-aware interpretation of model suitability. The results show that EA-BiGRU-Multi is competitive in selected medium- and long-horizon cases when reliable day-ahead demand priors are available, while simpler lin-

ear or patch-based models remain preferable in other settings. These findings support horizon-aware and region-aware model selection rather than a single architecture-level ranking.

The remainder of this paper is organized as follows. Section 2 presents the proposed EA-BiGRU-Multi framework, including input feature construction, terminal-anchor injection, feature-channel recalibration, shared BiGRU encoding, and multi-head forecasting. Section 3 describes the dataset, experimental settings, baseline models, evaluation metrics, and experimental results. Section 4 concludes the paper and discusses limitations and future research directions.

2. Proposed Method

EA-BiGRU-Multi is designed for regional multi-horizon short-term electricity load forecasting, where historical load, weather, calendar variables, and day-ahead demand information are available at the forecast origin. The central idea is to treat day-ahead demand as a forecast-origin prior rather than as a repeated historical feature. The framework therefore combines terminal-anchor injection, feature-channel recalibration, shared BiGRU temporal encoding, and horizon-specific forecasting heads.

As shown in Figure 1, the model constructs the input sequence from historical and known-future feature groups. Day-ahead demand priors are inserted only at the final frame of the input window using terminal-anchor injection. The constructed sequence is then recalibrated along the feature dimension and encoded by a shared BiGRU. Horizon-specific temporal attention pooling and independent forecasting heads produce predictions for the three target horizons. The three subfigures of Figure 1 illustrate the overall framework, the terminal-anchor injection mechanism, the feature-channel recalibration block, and the shared encoder with horizon-specific output heads, respectively.

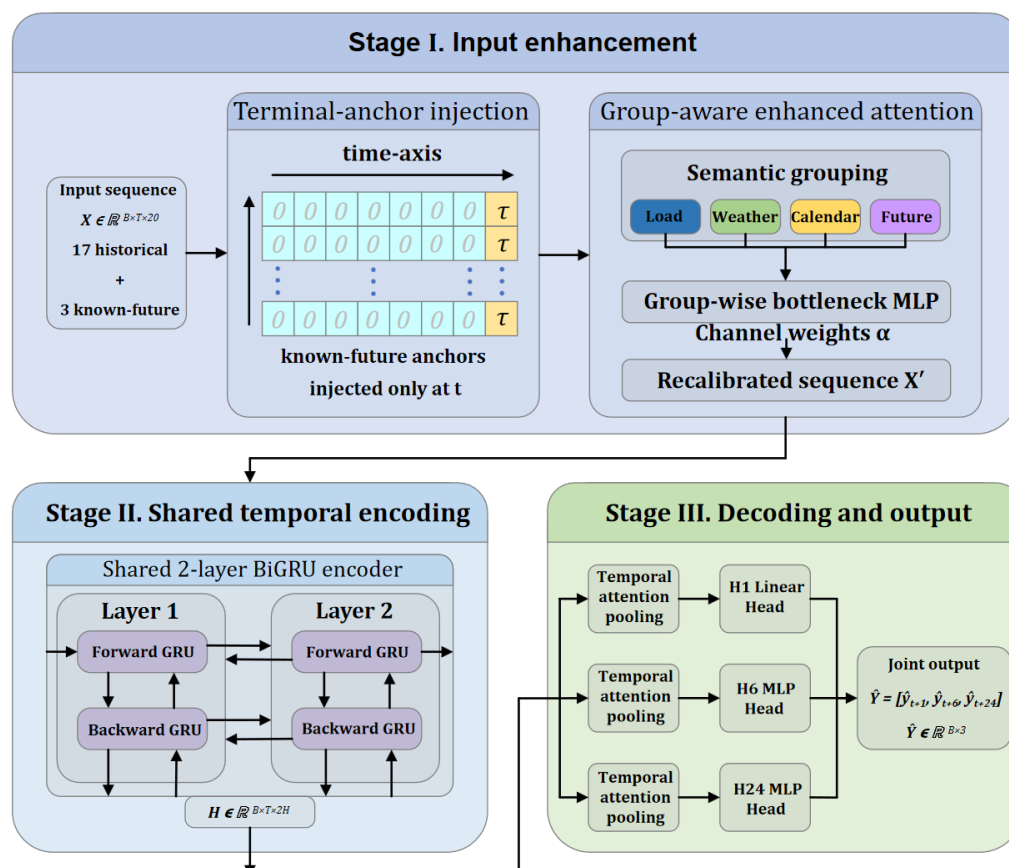


Figure 1. Architecture of EA-BiGRU-Multi.

2.1. Problem Formulation and Input Features

Consider an input window ending at time step t :

$$\mathbf{X}_t = [\mathbf{x}_{t-T+1}, \mathbf{x}_{t-T+2}, \dots, \mathbf{x}_t] \in \mathbb{R}^{T \times F},$$

where T is the input window length and F is the number of input channels. In this study, $T = 336$, corresponding to two weeks of hourly observations. The forecasting task is to estimate future real-time load values at three discrete lead times:

$$(\hat{y}_{t+1}, \hat{y}_{t+6}, \hat{y}_{t+24}) = f(\mathbf{X}_t; \Theta),$$

where $\hat{y}_{t+1}$, $\hat{y}_{t+6}$, and $\hat{y}_{t+24}$ denote the predicted loads at 1-step, 6-step, and 24-step ahead, respectively, and Θ denotes the trainable parameters of the model. The model produces three scalar outputs corresponding to these discrete horizons, rather than generating a full step-by-step prediction sequence from $t + 1$ to $t + 24$. This discrete multi-horizon formulation differs from sequence-to-sequence forecasting and is adopted because the three selected horizons correspond to distinct operational decision timescales in power system dispatch.

Input variables are organized into four groups: load-related, meteorological, calendar, and known-future demand features, as summarized in Table 1. The 17 historical channels are populated over the full input window and contain information that has been observed or recorded up to the current time step. The remaining three channels represent day-ahead demand priors associated with H1, H6, and H24. These known-future channels are treated differently from historical channels: their values are determined by the day-ahead market clearing process and become available to system operators before real-time dispatch, but they are not physical measurements that accumulate over each past time step. Terminal-anchor injection, described in Section 2.3, places these values only at the forecast-origin position to preserve this temporal distinction.

Table 1. Input feature groups.

Category	Features	Count
Load	RT_Demand, DA_Demand, lag_1, lag_24, lag_168	5
Meteorological	Dry_Bulb, Dew_Point	2
Calendar	hour, dayofweek, month, is_weekend, hour_sin, hour_cos, dow_sin, dow_cos, month_sin, month_cos	10
Known-future exogenous	da_future_h1, da_future_h6, da_future_h24	3
Total	–	20

2.2. Overall Framework

Three modeling considerations guide the design of EA-BiGRU-Multi. H1, H6, and H24 forecasts share part of the historical temporal context, but their final prediction mappings should not be forced to be identical. Heterogeneous input channels, including load, weather, calendar, and day-ahead demand variables, may have different relevance across regions and forecast horizons. Known-future demand information should remain associated with the forecast origin instead of being assigned to all historical timestamps.

Following these considerations, EA-BiGRU-Multi proceeds through four main modules. Terminal-anchor injection places the day-ahead demand prior at the final frame of the input window. Feature-channel recalibration adjusts the relative contribution of heterogeneous input channels before temporal encoding. A shared BiGRU encoder then extracts temporal representations from the recalibrated sequence. Horizon-specific temporal

attention and output heads generate separate forecasts for H1, H6, and H24. Rather than imposing a fixed architecture ranking, this structure is intended to examine how temporal representation learning and forecast-origin prior integration work together in regional multi-horizon forecasting.

2.3. Forecast-Origin Prior and Terminal-Anchor Injection

Day-ahead demand can provide useful information about the expected future demand level. A simple treatment is to repeat the known-future value across all historical positions of the input window. However, such broadcasting changes the temporal meaning of the feature: a value available at the forecast origin is implicitly converted into a pseudo-historical time-varying covariate. This may blur the distinction between observed historical variables and forecast-origin prior information.

EA-BiGRU-Multi therefore adopts terminal-anchor injection. For each horizon $h \in \{1, 6, 24\}$, the day-ahead demand value associated with the target time $t + h$ is inserted only at the final position of the input window. The known-future channel f_h is defined as

$$x_{t,\tau}^{(f_h)} = \begin{cases} \text{DA}_{t+h}, & \tau = T - 1, \\ 0, & 0 \leq \tau < T - 1, \end{cases}$$

where $x_{t,\tau}^{(f_h)}$ denotes the value of the known-future channel for horizon h at input position τ , and DA_{t+h} denotes the day-ahead demand value corresponding to the target timestamp. The zero values in non-terminal positions serve as neutral placeholders in the normalized feature space; they do not represent physical zero demand.

From an information-availability perspective, this design keeps the day-ahead demand prior tied to the time at which the forecast is issued. The model can use the prior as a demand-level reference without assigning it to historical timestamps. The subsequent BiGRU encoder operates only on the constructed input window and does not access unknown future real-time load targets.

2.4. Feature-Channel Recalibration

After terminal-anchor injection, the input tensor contains heterogeneous feature groups with different physical meanings and different levels of relevance to the forecasting task. The feature-channel recalibration module is used to adjust these channel contributions before temporal encoding. The module applies the general idea of channel-wise feature recalibration established in prior work on squeeze-and-excitation networks and convolutional block attention modules [31,32]. It is used here as a lightweight preprocessing block for multivariate load forecasting, and no novelty is claimed for the recalibration mechanism itself. The contribution lies in combining this standard technique with terminal-anchor injection and shared recurrent encoding within a multi-horizon forecasting framework.

Let the input tensor after terminal-anchor injection be

$$\mathbf{X}_t^a \in \mathbb{R}^{B \times T \times F},$$

where B is the batch size. The tensor is first transposed into a feature-first representation:

$$\mathbf{H} = \text{Transpose}(\mathbf{X}_t^a) \in \mathbb{R}^{B \times F \times T}.$$

A compact channel descriptor is then obtained by combining temporal average pooling and temporal max pooling:

$$\mathbf{z} = \text{AvgPool}_T(\mathbf{H}) + \text{MaxPool}_T(\mathbf{H}), \quad \mathbf{z} \in \mathbb{R}^{B \times F \times 1}.$$

The descriptor is passed through a bottleneck mapping to produce channel weights:

$$\alpha = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2), \quad \alpha \in \mathbb{R}^{B \times F \times 1},$$

where α denotes the channel weight vector, $\sigma(\cdot)$ is the sigmoid activation, $\delta(\cdot)$ is the nonlinear activation, and $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are learnable parameters.

Channel-wise multiplication then gives the recalibrated representation:

$$\mathbf{H}' = \mathbf{H} \odot \alpha,$$

where $\odot$ denotes element-wise multiplication. The tensor is transposed back to the sequence-first format:

$$\mathbf{X}'_t = \text{Transpose}(\mathbf{H}') \in \mathbb{R}^{B \times T \times F}.$$

With this operation, the recurrent encoder receives a standard sequence input, while the relative contributions of load, weather, calendar, and forecast-origin demand channels have been recalibrated.

2.5. Shared BiGRU Temporal Encoding

Recalibrated sequence $\mathbf{X}'_t$ is encoded by a shared BiGRU. GRU-based recurrent units are suitable for sequential data because their gating mechanism controls hidden-state updating and retention, while bidirectional recurrent structures encode the available input window from both directions [14,15]. In EA-BiGRU-Multi, the same BiGRU encoder is shared by the three forecasting horizons, allowing common temporal patterns to be learned from a unified input window.

The choice of a bidirectional encoder in an operational forecasting context requires justification. The BiGRU in this framework operates exclusively within the observed and constructed input window, which contains only historical observations and the terminal-anchor prior. No future real-time load values beyond the forecast origin are accessible to the encoder. Within this fixed-length historical window, the backward pass allows the model to reinterpret earlier observations in light of more recent context, which can improve the extraction of temporal patterns such as weekly cycles and seasonal transitions. This usage is analogous to bidirectional encoders in sequence-to-one prediction tasks, where the output is a single prediction conditioned on a fully observed input sequence, rather than an autoregressive decoder that generates future steps one at a time.

For an input vector $\mathbf{x}'_\tau$ and the previous hidden state $\mathbf{h}_{\tau-1}$, the GRU transition is written as

$$\begin{aligned} \mathbf{r}_\tau &= \sigma(\mathbf{W}_r \mathbf{x}'_\tau + \mathbf{U}_r \mathbf{h}_{\tau-1} + \mathbf{b}_r), \\ \mathbf{u}_\tau &= \sigma(\mathbf{W}_u \mathbf{x}'_\tau + \mathbf{U}_u \mathbf{h}_{\tau-1} + \mathbf{b}_u), \\ \tilde{\mathbf{h}}_\tau &= \tanh(\mathbf{W}_h \mathbf{x}'_\tau + \mathbf{U}_h (\mathbf{r}_\tau \odot \mathbf{h}_{\tau-1}) + \mathbf{b}_h), \\ \mathbf{h}_\tau &= (1 - \mathbf{u}_\tau) \odot \mathbf{h}_{\tau-1} + \mathbf{u}_\tau \odot \tilde{\mathbf{h}}_\tau. \end{aligned} \quad (1)$$

where $\mathbf{r}_\tau$ is the reset gate, $\mathbf{u}_\tau$ is the update gate, $\tilde{\mathbf{h}}_\tau$ is the candidate hidden state, and $\mathbf{h}_\tau$ is the hidden state at time step τ .

The BiGRU processes the available input window in both forward and backward directions:

$$\begin{aligned} \vec{\mathbf{h}}_\tau &= \text{GRU}_{\text{fwd}}(\mathbf{x}'_\tau, \vec{\mathbf{h}}_{\tau-1}), \\ \overleftarrow{\mathbf{h}}_\tau &= \text{GRU}_{\text{bwd}}(\mathbf{x}'_\tau, \overleftarrow{\mathbf{h}}_{\tau+1}), \\ \mathbf{g}_\tau &= [\vec{\mathbf{h}}_\tau; \overleftarrow{\mathbf{h}}_\tau], \end{aligned}$$

where $\vec{\mathbf{h}}_\tau$ and $\overleftarrow{\mathbf{h}}_\tau$ denote the forward and backward hidden states, and $\mathbf{g}_\tau$ is their concatenated representation.

For the full input window, the shared encoder output is

$$\mathbf{G} = \text{BiGRU}(\mathbf{X}'_t) \in \mathbb{R}^{B \times T \times 2d},$$

where d is the hidden dimension of each GRU direction. The backward direction is applied only within the observed and constructed input window. It does not use unknown real-time target loads beyond the forecast origin. Therefore, the BiGRU is used as a window-level representation encoder, not as a strictly autoregressive online decoder.

2.6. Horizon-Specific Temporal Attention and Multi-Head Output

The shared BiGRU encoder produces a sequence of hidden representations, but different forecast horizons may depend on different parts of the historical window. H1 forecasting is usually more sensitive to recent fluctuations, whereas H6 and H24 forecasting may require broader daily or weekly context. To allow horizon-dependent selection of temporal information, a separate temporal attention pooling layer is used for each horizon.

For each forecast horizon $h \in \{1, 6, 24\}$, the attention score at time step τ is computed as

$$s_{h,\tau} = \mathbf{w}_h^\top \mathbf{g}_\tau,$$

where $\mathbf{w}_h \in \mathbb{R}^{2d}$ is a learnable attention vector for horizon h . The attention weights are obtained by softmax normalization:

$$\beta_{h,\tau} = \frac{\exp(s_{h,\tau})}{\sum_{j=0}^{T-1} \exp(s_{h,j})},$$

where $\beta_{h,\tau}$ denotes the normalized attention weight assigned to time step τ for horizon h . The horizon-specific context vector is then calculated as

$$\mathbf{c}_h = \sum_{\tau=0}^{T-1} \beta_{h,\tau} \mathbf{g}_\tau,$$

where $\mathbf{c}_h \in \mathbb{R}^{2d}$ is the context representation used for horizon h .

Each forecasting head maps its horizon-specific context vector to one scalar prediction:

$$\hat{y}_{t+h} = \phi_h(\mathbf{c}_h), \quad h \in \{1, 6, 24\},$$

where $\phi_h(\cdot)$ denotes the output mapping for horizon h . This multi-head structure keeps the temporal encoder shared while allowing the final prediction behavior to differ across H1, H6, and H24.

2.7. Training Objective

Joint optimization is used for the three forecasting tasks. The total loss is defined as

$$\mathcal{L} = \sum_{h \in \{1, 6, 24\}} \omega_h \mathcal{L}_h(\hat{y}_{t+h}, y_{t+h}),$$

where y_{t+h} is the observed load at horizon h , $\hat{y}_{t+h}$ is the corresponding prediction, $\mathcal{L}_h(\cdot)$ is the horizon-specific forecasting loss, and ω_h is the loss weight assigned to horizon h .

In the experiments, $\mathcal{L}_h$ is implemented using the Huber loss to reduce sensitivity to occasional large errors while retaining stable gradient behavior. The same model structure is trained jointly for H1, H6, and H24.

The key hyperparameters of EA-BiGRU-Multi are listed in Table 2.

Table 2. Key hyperparameters of EA-BiGRU-Multi.

Hyperparameter	Value
Input window length (T)	336
Number of input channels (F)	20
BiGRU hidden dimension (d)	128
Number of BiGRU layers	2
Recalibration bottleneck ratio	4
Dropout rate	0.2
Optimizer	AdamW
Initial learning rate	0.001
Weight decay	1×10^{-4}
Scheduler	Cosine annealing warm restart
Batch size	32
Maximum epochs	300
Early stopping patience	50
Gradient clipping norm	1.0
Horizon loss weights ($\omega_1, \omega_6, \omega_{24}$)	1.0, 1.0, 1.0
Loss function	Huber loss ($\delta = 1.0$)

These hyperparameters are kept identical across all three load zones (CT, SEMA, VT) to ensure that performance differences are attributable to regional load characteristics and information availability rather than to zone-specific tuning.

3. Experiments and Results

Experimental analysis is designed to examine how EA-BiGRU-Multi behaves under regional multi-horizon short-term load forecasting conditions. Particular attention is given to three factors: forecast horizon, regional load characteristics, and the availability of day-ahead demand priors. Hourly load data from ISO New England in 2024 are used for all experiments. Section 3.1 describes dataset construction and the evaluation protocol. Section 3.2 introduces the compared models and training settings. Section 3.3 defines the evaluation metrics. Sections 3.4–3.11 report the main multi-seed comparison, historical-input-only comparison, reference baseline results, controlled ablation, anchor-design validation, cross-architecture anchor ablation, noise sensitivity analysis, and representative prediction behavior. Section 3.12 summarizes the main findings from a horizon- and zone-aware perspective.

3.1. Dataset Description and Experimental Protocol

Hourly electricity load records and related exogenous variables are obtained from the ISO New England public load and demand data source for 2024 [28]. Three representative load zones are selected as the primary study regions: Connecticut (CT), Southeast Massachusetts (SEMA), and Vermont (VT). These zones differ in load scale, seasonal demand pattern, and regional volatility, providing a heterogeneous testbed for evaluating model behavior under different regional operating conditions.

Raw records are preprocessed through several steps. Hourly timestamps are reconstructed to ensure a continuous time index; daylight-saving-time transitions are handled by detecting duplicate fall-back hours and averaging the two records, while the missing spring-forward hour is filled by linear interpolation between adjacent values. Remaining missing entries (fewer than 0.15% of total records across all zones) are imputed using linear interpolation when gaps span three or fewer consecutive hours; records with longer gaps are excluded from the training and evaluation windows. Invalid load values, defined

as non-positive demand or values exceeding five standard deviations from the monthly zonal mean, are treated as missing and imputed following the same interpolation rule. Meteorological variables (dry-bulb temperature and dew-point temperature) are obtained from the nearest available weather station to each load zone centroid and merged with load records by hourly timestamp. Calendar features (hour-of-day, day-of-week, month, week-end indicator, and their sine/cosine encodings) are derived directly from the reconstructed timestamp index. The input window length is set to 336 hourly steps, corresponding to two weeks of historical observations. Forecasting targets are one-step-, six-step-, and 24-step-ahead real-time electricity load values, denoted as H1, H6, and H24, respectively.

All datasets are split chronologically into training, validation, and test sets so that future records are not used during model training or hyperparameter selection. For the final comparison, predictions are evaluated on a common aligned test window with $n = 934$ samples for each evaluated zone-horizon setting, following the general requirement that time-series forecasting evaluation should preserve temporal order and maintain consistent out-of-sample comparison conditions [29,30]. This aligned evaluation protocol ensures that different models are compared on the same timestamp intersection, reducing the possibility that performance differences are caused by inconsistent rolling-origin test coverage rather than by model behavior.

EA-BiGRU-Multi uses 20 input channels, including 17 historical feature channels and three known-future day-ahead demand anchor channels corresponding to H1, H6, and H24. The known-future channels are injected only at the final frame of the input window, following the terminal-anchor design described in Section 2.3. In the main comparison, DLinear and PatchTST are also evaluated under the same 20-dimensional input setting. Traditional recurrent and convolutional baselines are included as reference models. In the default reference setting, these models are trained with 17-dimensional historical inputs without known-future anchor channels. This design choice reflects the fact that single-horizon baselines were originally developed without the terminal-anchor input structure. To address the potential concern that the dimensional difference may introduce an uncontrolled comparison, a supplementary experiment is conducted in which selected conventional baselines are evaluated using the same 20-dimensional input layout with terminal-anchor channels. The supplementary results are reported in Section 3.6 alongside the 17-dimensional reference results, allowing the reader to assess whether the dimensional difference substantially affects the reference comparison. The primary information-controlled comparison remains among EA-BiGRU-Multi, DLinear, and PatchTST, all under the 20-dimensional input setting.

3.2. Compared Models and Training Settings

EA-BiGRU-Multi, DLinear, and PatchTST constitute the primary comparison. All three models are evaluated under the same aligned test-window protocol and trained with three random seeds: 42, 123, and 456. Reported results are given as mean values with standard deviations. This setting reduces the influence of random initialization and provides a more stable basis for comparing model behavior across zones and forecast horizons.

DLinear and PatchTST are selected as recent time-series forecasting baselines with different inductive biases [24,25]. DLinear represents a lightweight linear structure that is often competitive when recent historical trajectories contain strong trend or periodic information, while PatchTST represents a patch-based Transformer architecture designed to model temporal dependencies from historical sequences. Together with EA-BiGRU-Multi, these models form the main information-controlled comparison under the 20-dimensional input setting.

The reference comparison further includes LSTM, BiLSTM, BiGRU, EA-BiGRU, TCN, RCEEMDAN-BiGRU, and RCEEMDAN-EA-BiGRU. These models are included to provide a broader context against commonly used short-term load forecasting architectures. Unlike the primary comparison models, the reference baselines are trained under single-horizon settings and use 17-dimensional historical inputs without known-future day-ahead demand anchor channels. Their results are therefore interpreted as reference performance, rather than as the main basis for architecture-level ranking.

EA-BiGRU-Multi is trained using the AdamW optimizer, Huber loss, cosine annealing warm-restart scheduling, gradient clipping, and early stopping. The maximum number of training epochs is set to 300, the early-stopping patience is set to 50, and the batch size is set to 32. The input window length and forecast horizons are kept unchanged across CT, SEMA, and VT to ensure that differences in performance are not caused by different temporal input lengths or target definitions.

All experiments are implemented in Python 3.10 using PyTorch 2.1.0. The same computing environment is used for all compared models, including the same versions of Python, PyTorch, NumPy 1.24.3, pandas 2.0.3, and scikit-learn 1.3.0. This setting keeps the implementation environment consistent across the primary comparison, reference comparison, ablation experiments, and sensitivity analyses.

3.3. Evaluation Metrics

Forecasting performance is evaluated using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). MAE measures the average absolute deviation between the observed and predicted load values, while RMSE gives larger penalties to large forecasting errors. Both MAE and RMSE are reported in MW. The R^2 metric measures how well the predicted series explains the variation in the observed load series.

The MAE is defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

where y_i denotes the observed load value, $\hat{y}_i$ denotes the predicted value, and n denotes the number of test samples.

The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

Compared with MAE, RMSE is more sensitive to large deviations and is therefore useful for identifying cases in which a model produces occasional large forecasting errors.

The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where $\bar{y}$ denotes the mean observed load over the test set. A lower MAE or RMSE indicates smaller forecasting error, whereas a higher R^2 indicates better agreement with the observed load variation. Negative R^2 values indicate that the model performs worse than using the test-set mean as a constant prediction.

3.4. Main Multi-Seed Forecasting Results

Tables 3–5, report the main multi-seed comparison among EA-BiGRU-Multi, DLinear, and PatchTST. The three models are evaluated under the same 20-dimensional input setting

and the same aligned test-window protocol. Table 3 reports the mean R^2 values with standard deviations, Table 4 reports MAE, and Table 5 reports RMSE. Figure 2 visualizes the same comparison to highlight horizon- and zone-dependent performance patterns.

Main comparison results show that the preferred model changes with forecast horizon, region, and evaluation metric. At H1, DLinear obtains the highest mean R^2 and the lowest MAE and RMSE in all three zones. This pattern suggests that one-step-ahead regional load forecasting is strongly driven by recent historical trajectories, for which a lightweight linear structure can be highly effective.

Table 3. Multi-seed mean $R^2 \pm$ std of the three primary models.

Zone	Horizon	EA-BiGRU-Multi	DLinear	PatchTST
CT	H1	0.9410 $\pm$ 0.0331	0.9868 $\pm$ 0.0027	0.8035 $\pm$ 0.0120
CT	H6	0.8947 $\pm$ 0.0287	0.8763 $\pm$ 0.0074	0.7412 $\pm$ 0.0342
CT	H24	0.9337 $\pm$ 0.0128	0.9217 $\pm$ 0.0031	0.6321 $\pm$ 0.0902
SEMA	H1	0.9300 $\pm$ 0.0458	0.9664 $\pm$ 0.0187	0.6826 $\pm$ 0.0518
SEMA	H6	0.8346 $\pm$ 0.0878	0.7604 $\pm$ 0.0420	0.6684 $\pm$ 0.0236
SEMA	H24	0.6700 $\pm$ 0.0744	0.6533 $\pm$ 0.0439	0.6324 $\pm$ 0.0131
VT	H1	0.8723 $\pm$ 0.0176	0.9473 $\pm$ 0.0318	0.3874 $\pm$ 0.0636
VT	H6	0.3643 $\pm$ 0.2117	0.2794 $\pm$ 0.6046	0.2188 $\pm$ 0.0843
VT	H24	0.4946 $\pm$ 0.0623	0.6045 $\pm$ 0.1266	0.1839 $\pm$ 0.0925

Table 4. Multi-seed mean MAE $\pm$ std of the three primary models.

Zone	Horizon	EA-BiGRU-Multi	DLinear	PatchTST
CT	H1	85.75 $\pm$ 28.34	40.98 $\pm$ 3.90	157.47 $\pm$ 4.54
CT	H6	117.04 $\pm$ 20.60	100.42 $\pm$ 3.84	181.50 $\pm$ 12.00
CT	H24	91.03 $\pm$ 9.49	88.27 $\pm$ 1.45	221.05 $\pm$ 25.81
SEMA	H1	48.87 $\pm$ 15.88	34.54 $\pm$ 10.34	106.94 $\pm$ 6.10
SEMA	H6	79.07 $\pm$ 24.99	100.44 $\pm$ 10.16	108.59 $\pm$ 3.03
SEMA	H24	118.73 $\pm$ 16.63	122.48 $\pm$ 9.47	116.30 $\pm$ 3.09
VT	H1	25.56 $\pm$ 2.03	15.56 $\pm$ 4.73	58.50 $\pm$ 2.11
VT	H6	61.40 $\pm$ 12.44	61.60 $\pm$ 31.68	66.19 $\pm$ 3.66
VT	H24	53.03 $\pm$ 3.77	46.59 $\pm$ 9.03	65.36 $\pm$ 3.23

Table 5. Multi-seed mean RMSE $\pm$ std of the three primary models.

Zone	Horizon	EA-BiGRU-Multi	DLinear	PatchTST
CT	H1	108.07 $\pm$ 29.56	52.22 $\pm$ 5.22	202.08 $\pm$ 6.09
CT	H6	147.51 $\pm$ 19.75	128.48 $\pm$ 4.87	232.30 $\pm$ 15.27
CT	H24	117.35 $\pm$ 11.12	111.23 $\pm$ 1.66	275.89 $\pm$ 32.98
SEMA	H1	64.59 $\pm$ 21.86	43.72 $\pm$ 11.86	137.19 $\pm$ 11.19
SEMA	H6	103.26 $\pm$ 31.99	119.21 $\pm$ 10.35	140.54 $\pm$ 5.06
SEMA	H24	147.19 $\pm$ 21.68	143.11 $\pm$ 9.13	147.56 $\pm$ 2.64
VT	H1	34.82 $\pm$ 2.03	20.54 $\pm$ 6.04	71.96 $\pm$ 3.78
VT	H6	80.33 $\pm$ 14.08	73.96 $\pm$ 31.56	81.47 $\pm$ 4.42
VT	H24	70.31 $\pm$ 4.48	57.66 $\pm$ 9.00	83.40 $\pm$ 4.73

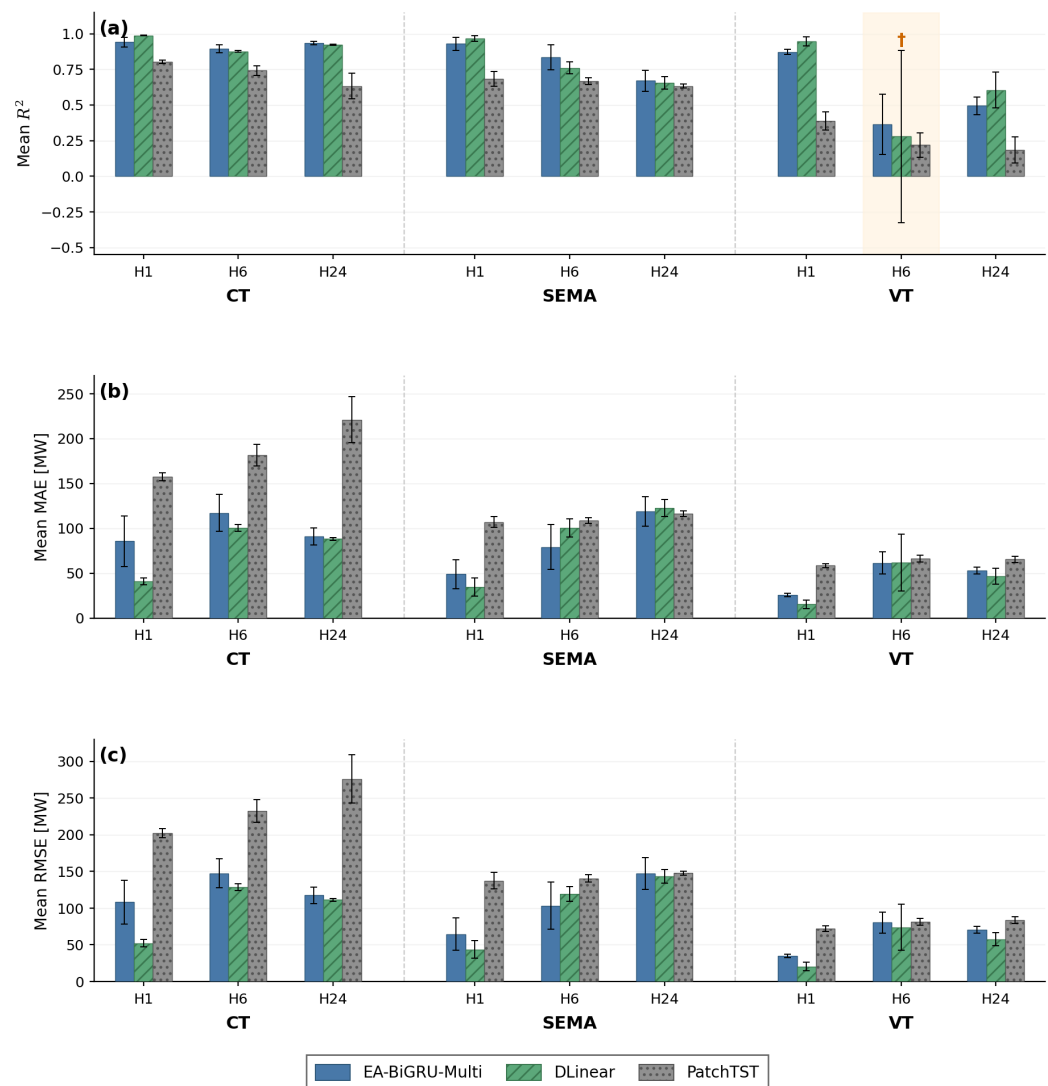


Figure 2. Multi-seed comparison of EA-BiGRU-Multi, DLinear, and PatchTST across the three load zones and three forecasting horizons: (a) mean R^2 comparison across CT, SEMA, and VT at H1, H6, and H24; (b) mean MAE (MW) comparison; (c) mean RMSE (MW) comparison. Error bars indicate standard deviations across three random seeds. The shaded region in (a) highlights the VT-H6 case with large cross-seed variance for DLinear (+).

At H6, the results become more metric- and region-dependent. In SEMA-H6, EA-BiGRU-Multi achieves the highest R^2 and the lowest MAE and RMSE among the three primary models. Compared with DLinear, the mean R^2 increases from 0.7604 to 0.8346, while MAE decreases from 100.44 MW to 79.07 MW and RMSE decreases from 119.21 MW to 103.26 MW. This suggests that shared recurrent temporal encoding and horizon-specific prediction are useful in this medium-horizon setting. In CT-H6, however, the interpretation is less uniform: EA-BiGRU-Multi gives a higher mean R^2 , whereas DLinear gives lower MAE and RMSE. The CT-H6 result therefore should not be reduced to a single model ranking; it reflects a difference between variance-explanation performance and absolute-error performance.

At H24, EA-BiGRU-Multi remains competitive in CT and SEMA, but its advantage depends on the metric. In CT-H24, EA-BiGRU-Multi achieves the highest mean R^2 , while DLinear yields slightly lower MAE and RMSE. In SEMA-H24, EA-BiGRU-Multi obtains the highest mean R^2 , PatchTST has the lowest MAE, and DLinear gives the lowest RMSE. These results indicate that day-ahead forecasting benefits from richer temporal representation

in some regional settings, but they also show that different metrics may favor different inductive biases.

The VT results are more variable than those of CT and SEMA. DLinear performs best at VT-H1 and VT-H24 across the three reported metrics, while VT-H6 shows large seed-dependent variation, especially for DLinear. Although EA-BiGRU-Multi has the lowest MAE at VT-H6, the differences among the three models are small in MAE and less stable in R^2 and RMSE. Therefore, VT-H6 is treated as an unstable zone–horizon case rather than as evidence for a firm model ranking.

Overall, the main results support a scenario-aware interpretation. DLinear is well suited to near-term extrapolation and remains strong in VT-H24. EA-BiGRU-Multi is more favorable in selected medium- and long-horizon cases, particularly SEMA-H6 and parts of the H24 comparison, where forecast-origin demand priors and nonlinear temporal representation are more useful. PatchTST is less competitive in most primary comparisons, but it remains relevant as a historical-sequence baseline and is further examined under the historical-input-only setting in Section 3.5. These findings suggest that model selection for regional STLF should depend on horizon, zone, information availability, and metric preference rather than on a single overall accuracy ranking.

3.5. Information-Controlled Historical-Input Comparison

An additional comparison is conducted to separate the effect of model architecture from the effect of known-future information. In this setting, EA-BiGRU-Multi, DLinear, and PatchTST are all evaluated using historical inputs only. The three day-ahead demand anchor channels are removed, and each model receives the same 17-dimensional historical input. The purpose of this experiment is to examine whether the performance patterns observed in the main comparison are primarily associated with architectural design or with the availability of forecast-origin demand priors.

The results are shown in Figure 3. For visual clarity, several VT-case R^2 values fall below the display range of the bar chart and are therefore truncated in the figure, while their actual numerical values are annotated directly on the corresponding bars. This plotting choice affects only visualization and does not alter the reported results.

Historical-input-only results differ substantially from the main comparison. At H1, DLinear achieves the highest R^2 in CT, SEMA, and VT, indicating that short-horizon forecasting remains well served by a simple linear structure when recent load history is sufficiently informative. At H6 and H24, PatchTST achieves the highest R^2 in all three zones, suggesting that patch-based sequence modeling is more effective when medium- and long-horizon prediction must rely entirely on historical observations.

By contrast, EA-BiGRU-Multi does not rank highest in any of the nine zone–horizon combinations under this historical-input-only setting. Its degradation is especially evident in VT, where the R^2 values drop to strongly negative levels across H1, H6, and H24. This result indicates that the proposed framework does not preserve its main-comparison competitiveness once the forecast-origin demand prior is removed. In other words, the benefit of EA-BiGRU-Multi should not be interpreted as arising from architecture alone under purely historical inputs.

This comparison clarifies the role of the proposed design. The main advantage of EA-BiGRU-Multi emerges when temporal representation learning is combined with forecast-origin prior integration through the terminal-anchor mechanism. Without that prior, the model becomes less competitive than DLinear at short horizons and less competitive than PatchTST at medium and long horizons. The results therefore support a more specific interpretation of the proposed framework: it is most useful in operational settings where

reliable day-ahead demand information is available and can be incorporated in a forecast-origin-consistent manner.

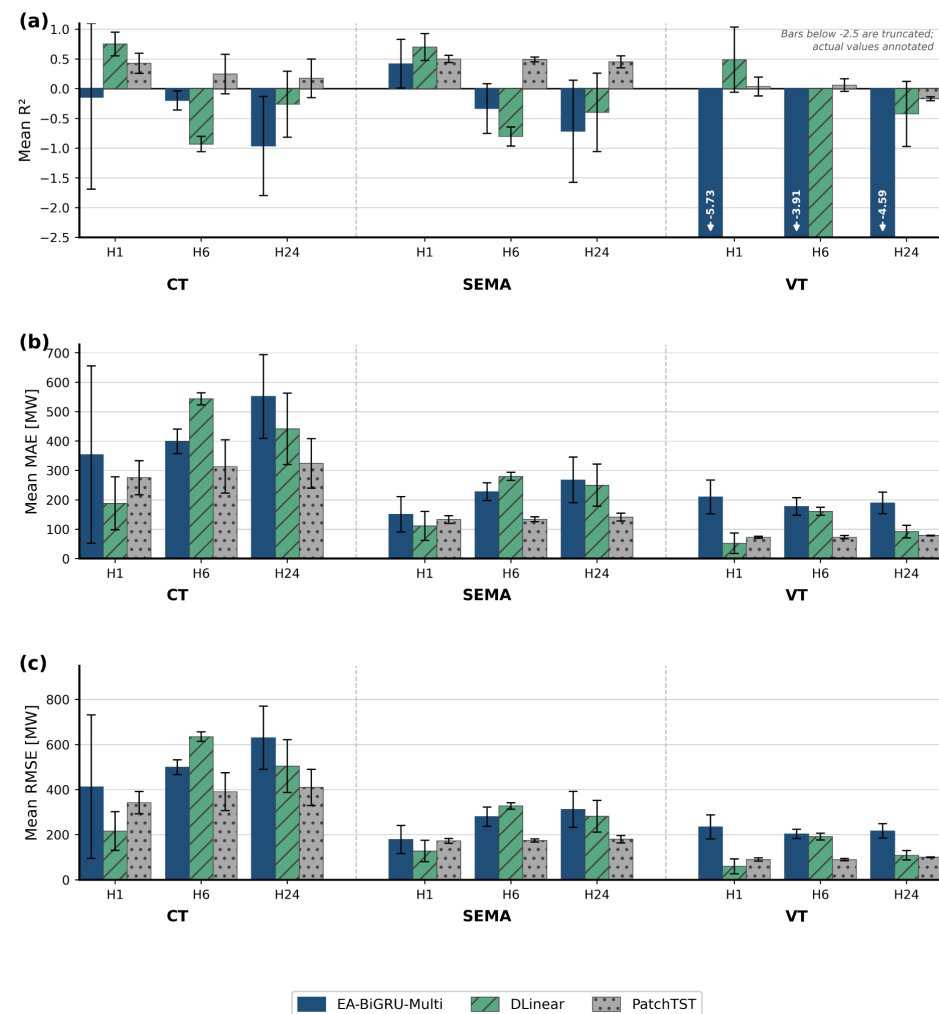


Figure 3. Information-controlled historical-input comparison of EA-BiGRU-Multi, DLinear, and PatchTST across CT, SEMA, and VT regions at forecast horizons H1, H6, and H24: (a) mean R^2 ; (b) mean MAE [MW]; (c) mean RMSE [MW].

Historical-input-only results complement the main comparison in Section 3.4. Model suitability depends not only on forecast horizon and regional load characteristics, but also on information availability. DLinear remains a strong choice for short-horizon extrapolation under purely historical inputs, whereas PatchTST is more favorable for historical-input-only forecasting at H6 and H24. EA-BiGRU-Multi becomes competitive when reliable day-ahead demand priors are available, and the input design can make effective use of them.

3.6. Reference Comparison with Conventional Single-Horizon Baselines

Conventional single-horizon forecasting models are further included to provide a broader empirical reference. The compared models include recurrent neural networks, convolutional models, enhanced recurrent variants, and decomposition-based recurrent models. These models are not used as the primary information-controlled comparison because they follow a single-horizon training setting and do not use the terminal-anchor day-ahead demand prior. Therefore, the results in Tables 6–8 are interpreted as reference performance, while the main aligned comparison among EA-BiGRU-Multi, DLinear, and PatchTST remains the primary basis for model-suitability analysis. The decomposition-

based baselines are included to represent forecasting pipelines that combine empirical mode decomposition with recurrent neural networks [33,34].

Table 6. Supplementary comparison of selected baselines under 20-dimensional input.

Zone–Horizon	Model	R^2 (17-dim)	R^2 (20-dim)	RMSE (20-dim)
CT-H1	LSTM	0.9777	0.9785	66.74
CT-H1	BiGRU	0.9774	0.9780	67.46
CT-H6	LSTM	0.9288	0.9312	95.82
CT-H6	BiGRU	0.9280	0.9305	96.30
CT-H24	LSTM	0.9021	0.9186	113.47
CT-H24	BiGRU	0.8996	0.9152	115.78
SEMA-H6	LSTM	0.7883	0.8024	108.27
SEMA-H6	BiGRU	0.7764	0.7935	110.64
VT-H1	LSTM	0.9515	0.9482	20.36
VT-H1	BiGRU	0.9503	0.9478	20.44
VT-H24	LSTM	0.4310	0.4586	67.48
VT-H24	BiGRU	0.3966	0.4273	69.40

Table 7. Reference comparison with conventional baselines at H1.

Zone	Model	R^2	MAE	RMSE
CT	LSTM	0.9777	50.40	67.88
CT	BiLSTM	0.9766	51.91	69.53
CT	BiGRU	0.9774	50.95	68.33
CT	TCN	0.9300	83.30	120.26
CT	EA-BiGRU	0.9732	56.55	74.41
CT	RCEEMDAN-BiGRU	0.9482	73.93	103.45
CT	RCEEMDAN-EA-BiGRU	0.9252	87.48	124.32
CT	DLinear	0.9868	40.98	52.22
CT	EA-BiGRU-Multi	0.9410	85.75	110.41
SEMA	LSTM	0.9541	41.75	51.10
SEMA	BiLSTM	0.9525	42.44	51.99
SEMA	BiGRU	0.9524	42.39	52.04
SEMA	TCN	0.8771	69.63	83.62
SEMA	EA-BiGRU	0.9503	42.30	53.18
SEMA	RCEEMDAN-BiGRU	0.9335	50.47	61.51
SEMA	RCEEMDAN-EA-BiGRU	0.9052	58.54	73.44
SEMA	DLinear	0.9664	34.54	43.72
SEMA	EA-BiGRU-Multi	0.9300	48.87	63.11
VT	LSTM	0.9515	15.33	19.71
VT	BiLSTM	0.9516	15.32	19.69
VT	BiGRU	0.9503	15.35	19.95
VT	TCN	0.7497	37.30	44.77
VT	EA-BiGRU	0.9437	16.16	21.23
VT	RCEEMDAN-BiGRU	0.9215	19.00	25.07
VT	RCEEMDAN-EA-BiGRU	0.8225	28.34	37.70
VT	DLinear	0.9473	15.56	20.54
VT	EA-BiGRU-Multi	0.8723	25.56	31.98

To examine whether the input dimensionality difference between the reference baselines (17 channels) and the primary models (20 channels) introduces a systematic comparison bias, a supplementary experiment is conducted. Selected conventional baselines (LSTM and BiGRU) are retrained using the same 20-dimensional input layout, where the three terminal-anchor channels follow the identical zero-padding and final-frame injection scheme used for EA-BiGRU-Multi. Table 9 reports the supplementary results for repre-

sentative zone–horizon cases. The results show that providing terminal-anchor channels to single-horizon recurrent baselines yields modest and inconsistent effects. In CT-H24, LSTM with 20-dimensional input achieves R^2 of 0.9186 compared with 0.9021 under the 17-dimensional setting, representing a moderate improvement. In SEMA-H6, LSTM with 20-dimensional input achieves R^2 of 0.8024 compared with 0.7883. However, in VT-H1, the 20-dimensional LSTM achieves R^2 of 0.9482, which is comparable to the 17-dimensional value of 0.9515, with no meaningful gain from the additional channels. These supplementary results indicate that the conventional single-horizon baselines do not consistently benefit from the terminal-anchor input structure, which was designed to work within a multi-head forecasting framework. The 17-dimensional reference results therefore remain meaningful as broader performance context.

Table 8. Reference comparison with conventional baselines at H6.

Zone	Model	R^2	MAE	RMSE
CT	LSTM	0.9288	93.72	97.48
CT	BiLSTM	0.9286	93.84	97.61
CT	BiGRU	0.9280	94.30	98.02
CT	TCN	0.8573	132.67	138.00
CT	EA-BiGRU	0.9244	94.94	100.44
CT	RCEEMDAN-BiGRU	0.8853	122.56	123.72
CT	RCEEMDAN-EA-BiGRU	0.9278	96.24	98.16
CT	DLinear	0.8763	100.42	128.48
CT	EA-BiGRU-Multi	0.8947	117.04	118.54
SEMA	LSTM	0.7883	94.64	112.05
SEMA	BiLSTM	0.7779	97.30	114.77
SEMA	BiGRU	0.7764	97.63	115.16
SEMA	TCN	0.6752	116.15	138.80
SEMA	EA-BiGRU	0.7885	93.76	112.00
SEMA	RCEEMDAN-BiGRU	0.7641	111.32	118.29
SEMA	RCEEMDAN-EA-BiGRU	0.7972	105.72	109.67
SEMA	DLinear	0.7604	100.36	119.21
SEMA	EA-BiGRU-Multi	0.8346	79.07	99.05
VT	LSTM	0.4961	52.53	61.85
VT	BiLSTM	0.5012	52.11	61.54
VT	BiGRU	0.5101	51.69	60.98
VT	TCN	−0.0898	72.63	90.96
VT	EA-BiGRU	0.4352	55.97	65.48
VT	RCEEMDAN-BiGRU	0.5360	59.14	59.35
VT	RCEEMDAN-EA-BiGRU	0.3995	57.62	67.52
VT	DLinear	0.2794	61.60	73.96
VT	EA-BiGRU-Multi	0.3643	61.40	69.47

At H1, most recurrent baselines achieve strong results, especially in CT and VT. This pattern is consistent with the persistence-dominated nature of one-step-ahead load forecasting, where recent load observations often provide enough information for short-term extrapolation. DLinear also remains highly competitive at H1, achieving the lowest MAE and RMSE in CT and SEMA and comparable performance in VT. These results show that EA-BiGRU-Multi is not intended to replace specialized short-horizon models in all near-term forecasting cases. Its role is more closely related to unified multi-horizon modeling and structured integration of forecast-origin demand priors.

At H6, the relative performance becomes more dependent on the load zone and evaluation metric. In SEMA-H6, EA-BiGRU is competitive among the conventional recurrent baselines, particularly in MAE, whereas RCEEMDAN-EA-BiGRU gives the highest R^2

and the lowest RMSE among the listed reference models. In CT, recurrent baselines remain strong, especially in MAE and RMSE. In VT, all models show weaker and less stable performance than in CT and SEMA, indicating that medium-horizon forecasting is more affected by regional demand shifts. The H6 reference results therefore support zone-specific interpretation rather than a single overall ranking.

Table 9. Reference comparison with conventional baselines at H24.

Zone	Model	R^2	MAE	RMSE
CT	LSTM	0.9021	110.03	124.37
CT	BiLSTM	0.9017	110.39	124.62
CT	BiGRU	0.8996	111.01	125.95
CT	TCN	0.5427	242.71	268.80
CT	EA-BiGRU	0.9071	106.11	121.15
CT	RCEEMDAN-BiGRU	0.7979	160.18	178.69
CT	RCEEMDAN-EA-BiGRU	0.8565	142.64	150.57
CT	DLinear	0.9217	88.27	111.23
CT	EA-BiGRU-Multi	0.9337	91.03	102.35
SEMA	LSTM	0.5865	133.83	156.28
SEMA	BiLSTM	0.5897	133.41	155.68
SEMA	BiGRU	0.5864	133.38	156.30
SEMA	TCN	0.3074	173.72	202.26
SEMA	EA-BiGRU	0.5810	133.46	157.32
SEMA	RCEEMDAN-BiGRU	0.3969	170.45	188.74
SEMA	RCEEMDAN-EA-BiGRU	0.5841	140.03	156.74
SEMA	DLinear	0.6533	122.48	143.10
SEMA	EA-BiGRU-Multi	0.6700	118.73	139.62
VT	LSTM	0.4310	54.45	69.16
VT	BiLSTM	0.4452	54.03	68.29
VT	BiGRU	0.3966	57.37	71.22
VT	TCN	−0.2799	81.41	103.72
VT	EA-BiGRU	0.4138	56.03	70.19
VT	RCEEMDAN-BiGRU	0.3945	59.77	71.34
VT	RCEEMDAN-EA-BiGRU	0.2909	60.46	77.20
VT	DLinear	0.6045	46.72	57.66
VT	EA-BiGRU-Multi	0.4946	53.03	65.18

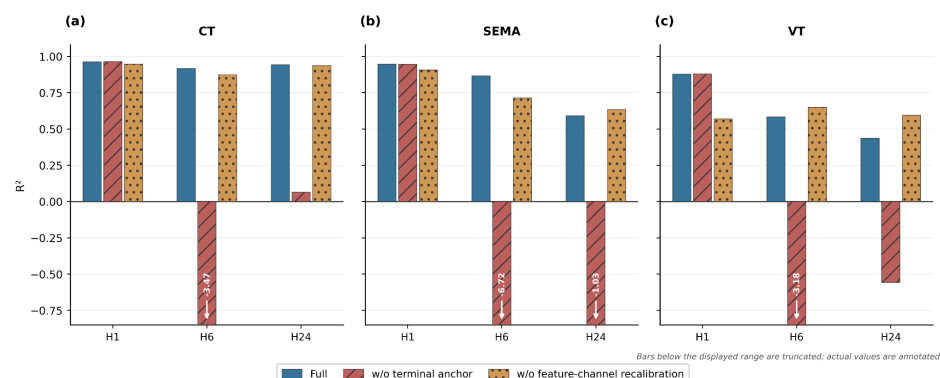
At H24, EA-BiGRU performs strongly in CT, achieving the best reference-baseline results across R^2 , MAE, and RMSE. In SEMA and VT, however, other recurrent baselines remain more favorable under at least one metric. These results again indicate that the conventional single-horizon baselines do not follow a single fixed ranking across zones and metrics. Overall, the reference comparison is consistent with the main multi-seed results: model suitability in regional STLF depends jointly on forecast horizon, regional load characteristics, information availability, and the metric used for interpretation.

3.7. Controlled Ablation of EA-BiGRU-Multi

Ablation analysis separates two design choices in EA-BiGRU-Multi: feature-channel recalibration and terminal-anchor injection of the day-ahead demand prior. The full model is compared with two variants. One removes the feature-channel recalibration module, and the other removes the terminal-anchor day-ahead demand input. This experiment uses the representative seed setting adopted for controlled component analysis, and the results are interpreted as diagnostic evidence rather than as a replacement for the multi-seed comparison in Section 3.4. The results are reported in Table 10 and visualized in Figure 4.

Table 10. Controlled ablation results of EA-BiGRU-Multi.

Zone	Horizon	Full Model R^2	Full RMSE	w/o Feature Recal. R^2	w/o Feature Recal. RMSE	w/o Anchor R^2	w/o Anchor RMSE
CT	H1	0.9628	87.67	0.9469	104.74	0.9639	86.36
CT	H6	0.9166	105.50	0.8731	130.13	−3.4683	772.21
CT	H24	0.9432	94.73	0.9358	100.71	0.0653	384.29
SEMA	H1	0.9462	55.33	0.9061	73.09	0.9460	55.43
SEMA	H6	0.8660	89.15	0.7143	130.17	−6.7235	676.83
SEMA	H24	0.5912	155.39	0.6345	146.93	−1.0337	346.59
VT	H1	0.8777	31.29	0.5698	58.69	0.8794	31.07
VT	H6	0.5835	56.23	0.6491	51.61	−3.1770	178.07
VT	H24	0.4359	68.86	0.5946	58.37	−0.5577	114.42

**Figure 4.** Controlled ablation results of EA-BiGRU-Multi: (a) R^2 comparison for CT; (b) R^2 comparison for SEMA; (c) R^2 comparison for VT.

The full model is compared with two variants: one without terminal-anchor injection and one without feature-channel recalibration. Higher R^2 indicates better forecasting performance.

Terminal-anchor input has the clearest effect in medium- and long-horizon forecasting. Removing the terminal anchor causes sharp deterioration at H6 and H24 in all three zones. For example, RMSE increases from 105.50 MW to 772.21 MW in CT-H6, from 89.15 MW to 676.83 MW in SEMA-H6, and from 56.23 MW to 178.07 MW in VT-H6. A similar pattern appears at H24, where anchor removal substantially lowers R^2 and increases RMSE in CT, SEMA, and VT. These results indicate that the day-ahead demand prior provides important demand-level information when the prediction target is farther away from the most recent historical observation.

The severity of the degradation when the terminal anchor is removed warrants further discussion. The collapse of R^2 to strongly negative values (e.g., −3.4683 in CT-H6) indicates that the shared BiGRU layers, when trained jointly with anchor-augmented inputs, develop internal representations that rely on the anchor signal for medium- and long-horizon projection. When this signal is removed at inference time, the learned feature extraction pathway no longer produces meaningful demand-level estimates. This behavior reflects a structural dependency in which the shared encoder allocates representational capacity to exploit the anchor prior, at the cost of reduced standalone historical-pattern extraction capability. This dependency is an inherent trade-off: the anchor-integrated model achieves stronger performance when the prior is available, but it is not designed to function as a robust history-only forecaster. In operational deployment, this trade-off can be managed through anchor-dropout regularization during training, where the terminal-anchor channels are randomly replaced with zero values for a fraction of training samples. To evaluate this mitigation strategy, EA-BiGRU-Multi is retrained with an anchor-dropout rate of 20%, meaning that for each training sample, there is a 20% probability that all terminal-anchor values are replaced by zeros. The results are reported in Table 11.

Table 11. Anchor-dropout regularization results.

Case	Setting	Full Anchor R^2	Full Anchor RMSE	w/o Anchor R^2	w/o Anchor RMSE
CT-H6	No dropout	0.9166	105.50	−3.4683	772.21
CT-H6	20% dropout	0.8890	121.67	0.7245	191.64
SEMA-H6	No dropout	0.8660	89.15	−6.7235	676.83
SEMA-H6	20% dropout	0.8401	97.36	0.6832	137.02
CT-H24	No dropout	0.9432	94.73	0.0653	384.29
CT-H24	20% dropout	0.9274	107.12	0.7518	198.05

The anchor-dropout variant achieves substantially better fallback performance when the anchor is unavailable at inference time. In CT-H6, the without-anchor R^2 improves from −3.4683 to 0.7245, while the with-anchor R^2 decreases moderately from 0.9166 to 0.8890. This pattern is consistent across the other evaluated cases: anchor-dropout training sacrifices a small fraction of peak performance in exchange for significantly improved robustness when the day-ahead demand prior becomes unavailable. This regularization technique provides a practical pathway for deploying EA-BiGRU-Multi in environments where day-ahead demand inputs may be intermittently unavailable, and it addresses the structural vulnerability identified in the standard ablation results.

The H1 results show a different pattern. Removing the terminal anchor produces only marginal changes in CT-H1, SEMA-H1, and VT-H1. This is consistent with the main comparison in Section 3.4, where one-step-ahead forecasting is largely governed by recent historical load trajectories. In this near-term setting, the day-ahead demand prior contributes less than it does at H6 and H24.

Feature-channel recalibration shows a more conditional effect. Removing this module degrades performance in CT-H1, CT-H6, CT-H24, SEMA-H1, SEMA-H6, and VT-H1. The effect is particularly visible in SEMA-H6, where R^2 decreases from 0.8660 to 0.7143 and RMSE increases from 89.15 MW to 130.17 MW. This suggests that recalibrating heterogeneous load, weather, calendar, and forecast-origin demand channels can help the model adjust the relative contribution of different feature groups.

However, feature-channel recalibration does not provide a uniform gain. In SEMA-H24, VT-H6, and VT-H24, the variant without recalibration obtains higher R^2 and lower RMSE than the full model. This result suggests that the recalibration module may introduce additional sensitivity in some zone–horizon settings, especially when the regional load pattern is less stable or when the test distribution differs from the training period. Therefore, the feature-channel recalibration module should be interpreted as a useful but conditional component, rather than as a consistently beneficial addition.

Controlled ablation clarifies the functional roles of the two components. Terminal-anchor injection is essential for using day-ahead demand as a forecast-origin prior, especially at H6 and H24. Feature-channel recalibration can improve performance in several CT and SEMA cases, but its contribution is not uniform across all zones and horizons. Together with the historical-input-only comparison in Section 3.5, these results support the central interpretation of this study: the value of EA-BiGRU-Multi lies in combining temporal representation learning with structured day-ahead demand integration, rather than in architecture-level dominance under every information setting.

3.8. Terminal-Anchor Injection Versus Naive Broadcasting

Ablation results above show that the day-ahead demand prior is important for medium- and long-horizon forecasting. A further question is how this known-future prior should be represented in the input sequence. To examine this issue, terminal-anchor injection is compared with a naive-broadcast strategy. In the naive-broadcast setting,

the same known-future demand value is repeated across all historical time steps in the input window. By contrast, terminal-anchor injection places the prior only at the final frame of the window, where the forecast is issued. This experiment focuses on the temporal placement of known-future demand information, rather than on whether the information is available.

Table 12 shows that terminal-anchor injection provides more favorable results than naive broadcasting in all three evaluated cases. The difference is most visible in CT-H24, where terminal-anchor injection increases the mean R^2 from 0.761 to 0.926 and reduces RMSE from 192.71 MW to 113.25 MW. The standard deviation is also smaller under terminal-anchor injection, indicating a more stable response across random seeds. A similar pattern appears in VT-H24, where the terminal-anchor setting improves mean R^2 from 0.444 to 0.550 and reduces RMSE from 68.42 MW to 61.63 MW.

Table 12. Terminal-anchor injection versus naive broadcasting.

Case	Terminal-Anchor R^2	Terminal-Anchor RMSE	Naive-Broadcast R^2	Naive-Broadcast RMSE
CT-H24	0.926 ± 0.012	113.25 ± 8.94	0.761 ± 0.149	192.71 ± 52.27
SEMA-H6	0.702 ± 0.062	132.30 ± 13.87	0.630 ± 0.098	147.02 ± 18.88
VT-H24	0.550 ± 0.023	61.63 ± 1.59	0.444 ± 0.062	68.42 ± 4.07

Improvement at SEMA-H6 is more moderate but remains consistent. Terminal-anchor injection increases mean R^2 from 0.630 to 0.702 and reduces RMSE from 147.02 MW to 132.30 MW. This suggests that the placement of known-future demand information is relevant not only for day-ahead forecasting, but also for medium-horizon prediction. Although the magnitude of improvement differs across cases, the direction of change is consistent in all tested settings.

These results support the input design adopted in EA-BiGRU-Multi. The day-ahead demand prior is useful not only because it is made available to the model, but also because its temporal meaning is preserved in the input representation. Naive broadcasting turns a forecast-origin prior into a repeated pseudo-historical feature, which may weaken the distinction between observed historical variables and information available at the forecast origin. Terminal-anchor injection keeps the known-future demand value attached to the final input frame, allowing the model to use it as a demand-level reference without assigning it to all past timestamps.

Comparison with naive broadcasting indicates that terminal-anchor injection is a more stable and semantically consistent way to incorporate day-ahead demand priors in the evaluated medium- and long-horizon cases. This result complements the controlled ablation in Section 3.7: the prior itself is important, and its temporal placement also affects how effectively the model can use it.

3.9. Cross-Architecture Anchor Ablation on DLinear

Previous experiments indicate that day-ahead demand information is important for EA-BiGRU-Multi. To test whether this effect is specific to the proposed recurrent framework, a cross-architecture anchor ablation is further conducted on DLinear. Two representative medium- and long-horizon settings are selected: SEMA-H6 and CT-H24. The full DLinear model uses the known-future demand anchor, whereas the zero-mask variant replaces anchor values with neutral placeholders while keeping the input structure unchanged. This design separates the informational contribution of the day-ahead demand prior from the recurrent architecture of EA-BiGRU-Multi.

Because DLinear uses linear projections that map each input channel across the temporal dimension to an output value, the implementation of terminal-anchor injection in

DLinear requires clarification. DLinear receives the same 336×20 input matrix used by EA-BiGRU-Multi. The three terminal-anchor channels follow the identical construction: zero values at positions $\tau = 0$ to $\tau = 334$, and the day-ahead demand value at position $\tau = 335$ (the final frame). DLinear's channel-independent linear layers learn a projection weight for each time step, so the model can learn to extract the anchor information from the single non-zero position in each anchor channel. In the zero-mask variant, the anchor value at position 335 is replaced with zero, making the entire anchor channel uniformly zero while preserving the 20-dimensional input shape. This ensures that any performance difference is attributable to the informational content of the anchor rather than to a change in input dimensionality.

In Table 13, the anchor gain is calculated as the R^2 difference between the full DLinear model and the zero-mask variant. Removing the known-future demand anchor leads to clear degradation in both evaluated settings. In SEMA-H6, the mean R^2 decreases from 0.746 to -1.345 , while RMSE increases from 122.75 MW to 371.92 MW. In CT-H24, the mean R^2 decreases from 0.948 to -0.089 , while RMSE increases from 90.74 MW to 408.76 MW. The degradation appears across all three seeds, although its magnitude varies by setting and initialization.

Table 13. DLinear anchor ablation results.

Case	Seed	Full DLinear R^2	Full RMSE	Zero-Mask R^2	Zero-Mask RMSE	Anchor Gain in R^2
SEMA-H6	42	0.773	116.03	-1.036	347.49	1.809
SEMA-H6	123	0.737	125.01	-1.157	357.71	1.894
SEMA-H6	456	0.727	127.22	-1.842	410.56	2.569
SEMA-H6	Mean $\pm$ std	0.746 ± 0.020	122.75 ± 4.84	-1.345 ± 0.355	371.92 ± 27.64	2.091
CT-H24	42	0.948	90.60	0.011	395.25	0.937
CT-H24	123	0.949	89.50	0.312	329.68	0.637
CT-H24	456	0.946	92.11	-0.591	501.36	1.537
CT-H24	Mean $\pm$ std	0.948 ± 0.001	90.74 ± 1.07	-0.089 ± 0.375	408.76 ± 70.74	1.037

This result indicates that the day-ahead demand prior provides useful information for medium- and long-horizon forecasting even when the forecasting model is based on a linear decomposition structure. The effect is therefore not exclusive to EA-BiGRU-Multi. Rather, it reflects the importance of forecast-origin demand information for longer lead times. EA-BiGRU-Multi is designed to combine this prior with recurrent temporal representation learning, whereas DLinear can also benefit from the same prior when it is available.

The cross-architecture ablation also helps clarify the interpretation of the main comparison. The competitive performance of EA-BiGRU-Multi in selected H6 and H24 cases should not be attributed to architecture alone. It arises from the interaction between model structure, forecast horizon, and the availability of known-future demand information. Together with the historical-input-only comparison in Section 3.5 and the controlled ablation in Section 3.7, this result reinforces the practical importance of information availability in multi-horizon load forecasting.

3.10. Noise Sensitivity of Terminal-Anchor Inputs

Day-ahead demand values used in operation may contain forecast errors. Since EA-BiGRU-Multi uses this information as a terminal-anchor prior, the model should not be evaluated only under clean-anchor conditions. A diagnostic noise sensitivity test is therefore conducted by perturbing the terminal-anchor demand channels during evaluation. The model parameters are not retrained in this experiment; only the known-future anchor inputs are perturbed at inference time.

The perturbed anchor input is defined as

$$\tilde{x}_{\text{DA}} = x_{\text{DA}}(1 + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad \sigma \in \{0, 0.05, 0.10, 0.15\},$$

where x_{DA} denotes the original day-ahead demand input, $\tilde{x}_{\text{DA}}$ denotes the perturbed input, ϵ is the multiplicative Gaussian noise term, and σ denotes the noise level. Four representative medium- and long-horizon cases are evaluated, and the results are reported in Table 14.

Table 14. Noise sensitivity of EA-BiGRU-Multi under perturbed terminal-anchor inputs.

Case	Metric	0% Noise	5% Noise	10% Noise	15% Noise
CT-H24	R^2	0.915	0.913	0.907	0.896
CT-H24	MAE	105.56	106.78	109.78	116.82
CT-H24	RMSE	133.07	134.66	139.08	147.28
SEMA-H6	R^2	0.735	0.734	0.730	0.724
SEMA-H6	MAE	107.19	107.51	108.16	109.19
SEMA-H6	RMSE	125.60	125.91	126.93	128.33
SEMA-H24	R^2	0.680	0.678	0.662	0.660
SEMA-H24	MAE	117.01	117.33	119.94	120.18
SEMA-H24	RMSE	137.75	138.07	141.57	141.81
VT-H24	R^2	0.560	0.559	0.550	0.541
VT-H24	MAE	49.03	49.03	49.97	49.86
VT-H24	RMSE	61.31	61.38	62.03	62.58

Performance declines gradually as the perturbation level increases. In CT-H24, R^2 decreases from 0.915 to 0.896 when the noise level increases from 0% to 15%, while RMSE increases from 133.07 MW to 147.28 MW. SEMA-H6 shows smaller degradation: R^2 decreases from 0.735 to 0.724, and RMSE increases from 125.60 MW to 128.33 MW. SEMA-H24 follows a similar pattern, with R^2 decreasing from 0.680 to 0.660 and RMSE increasing from 137.75 MW to 141.81 MW.

VT-H24 also shows gradual degradation rather than abrupt collapse. As noise level increases from 0% to 15%, R^2 decreases from 0.560 to 0.541, while RMSE increases from 61.31 MW to 62.58 MW. Although the absolute change is limited in this diagnostic setting, VT results should still be interpreted with caution because the main comparison and residual analysis indicate stronger regional sensitivity in this zone. Reliability of the day-ahead demand prior therefore remains an important deployment condition.

The synthetic noise levels used in this experiment can be compared with the observed historical error between day-ahead demand and real-time demand in the ISO New England dataset. The zone-level MAPE values between day-ahead and real-time demand are 2.41% for CT, 3.14% for SEMA, and 3.06% for VT, with an overall average of 2.87%. These values indicate that the 5% noise level in the synthetic test is already larger than the typical day-ahead forecast error, while the 10% and 15% levels represent substantially degraded prior quality beyond normal operating conditions. The gradual performance degradation observed under these noise levels therefore provides evidence that the terminal-anchor mechanism can tolerate realistic levels of day-ahead forecast uncertainty. Further analysis using the actual empirical error distribution, rather than symmetric Gaussian noise, would be valuable for refining deployment guidelines under non-standard operating conditions.

Moderate perturbations in the day-ahead demand prior lead to incremental performance degradation, indicating that the terminal-anchor input does not function as a brittle exact-value lookup in the evaluated cases. The model can still use the anchor as a demand-level reference when the prior contains limited uncertainty. At the same time, the experiment does not imply insensitivity to anchor quality. When day-ahead demand inputs become unreliable, adaptive fallback strategies such as switching to history-driven

forecasting or replacing known-future inputs with rolling historical proxies may help improve operational stability.

Noise sensitivity analysis supports a restrained interpretation of terminal-anchor injection. The design can tolerate moderate perturbations in the tested cases, but its effectiveness depends on the availability of reasonably reliable forecast-origin demand information. This finding is consistent with the broader experimental conclusion: EA-BiGRU-Multi is most appropriate when medium- or long-horizon forecasting can benefit from both temporal representation learning and informative day-ahead demand priors.

3.11. Representative Prediction Curves

Aggregate metrics summarize average forecasting accuracy, but they do not fully reveal how the predicted series follows the temporal evolution of load. To provide a more direct view of model behavior, representative prediction curves are examined for selected zone-horizon cases. Figure 5 shows the predicted and observed trajectories.

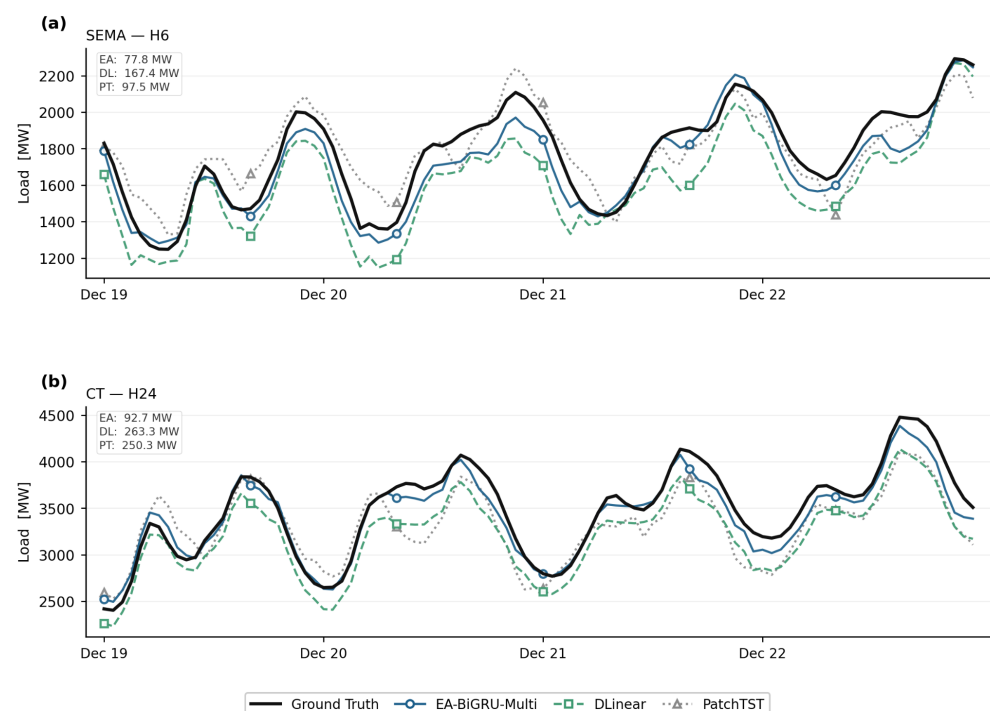


Figure 5. Representative prediction curves for selected zone-horizon cases: (a) SEMA-H6; (b) CT-H24.

In the representative medium- and long-horizon cases, EA-BiGRU-Multi generally follows the main load trajectory. In SEMA-H6 and CT-H24, the predicted series remains close to the observed load curve for most timestamps and captures the main changes in load level. This visual pattern is consistent with the quantitative results reported in the previous subsections, where the proposed framework shows comparatively favorable performance in these settings, especially when the forecast-origin demand prior is informative.

VT-related cases show a different behavior. Residuals are more dispersed, and trajectory matching is less stable than in CT and SEMA. This agrees with the larger cross-seed variability reported earlier, particularly at VT-H6. The VT test period exhibits a stronger seasonal demand shift relative to the training distribution, increasing forecasting difficulty for all evaluated models. VT results are therefore better interpreted as evidence of zone-dependent sensitivity under distribution shift, rather than as an isolated weakness of a specific architecture.

Representative curves are consistent with the main experimental findings. EA-BiGRU-Multi is more favorable in settings where nonlinear temporal dependence and forecast-origin demand priors are both informative, whereas forecasting behavior remains sensitive to load zone, forecast horizon, and stability of the underlying data distribution.

3.12. Summary of Experimental Findings

Experimental results lead to a scenario-aware interpretation of regional multi-horizon load forecasting. They do not support a single architecture-level ranking across all zones, horizons, and metrics. Instead, forecasting performance changes with prediction horizon, regional load behavior, information availability, and the metric used for evaluation.

Main multi-seed results show that DLinear is particularly effective at H1. In CT, SEMA, and VT, it obtains the strongest short-horizon performance across MAE, RMSE, and R^2 . This indicates that one-step-ahead load forecasting is largely driven by recent historical trajectories, where a lightweight linear structure can be sufficient. EA-BiGRU-Multi becomes more competitive in selected medium- and long-horizon settings, especially in SEMA-H6 and parts of the H24 comparison. These cases suggest that nonlinear temporal encoding and horizon-specific prediction heads can be useful when the forecasting target is farther from the latest observed load. PatchTST is less favorable in most main comparisons with known-future anchors, but it remains relevant under historical-input-only forecasting, especially at H6 and H24.

Historical-input-only comparison provides an important boundary for interpreting EA-BiGRU-Multi. When day-ahead demand anchor channels are removed and all three primary models use only historical inputs, relative performance changes substantially. DLinear remains strong at H1, while PatchTST becomes more favorable at H6 and H24. EA-BiGRU-Multi does not retain the same competitiveness under this information-controlled setting. This result indicates that the main value of the proposed framework should not be attributed to architecture alone. Its advantage appears when temporal representation learning is combined with forecast-origin demand prior integration.

Controlled ablation further clarifies the roles of the model components. Terminal-anchor injection has the clearest effect at H6 and H24. Removing the day-ahead demand anchor leads to substantial degradation in both R^2 and RMSE in multiple medium- and long-horizon cases. The severity of this degradation reflects the structural dependency of the jointly trained encoder on the anchor signal, which can be mitigated through anchor-dropout regularization as demonstrated in Section 3.7. By contrast, removing feature-channel recalibration produces more conditional effects. It improves performance in several CT and SEMA settings, but it is not uniformly beneficial in all zone-horizon combinations. Feature-channel recalibration should therefore be interpreted as a useful auxiliary component, whereas forecast-origin demand information is the more critical factor for longer lead times.

Comparison between terminal-anchor injection and naive broadcasting shows that temporal placement of known-future information matters. Repeating the day-ahead demand prior across the historical window weakens the distinction between observed historical variables and information available at the forecast origin. Terminal-anchor injection provides a more stable representation by placing the known-future prior only at the final input frame. DLinear anchor ablation leads to the same general interpretation from a different architecture: removing the anchor substantially degrades DLinear in the tested medium- and long-horizon cases. Forecast-origin demand priors should therefore be treated as an information-design issue, not only as a component specific to EA-BiGRU-Multi.

Noise sensitivity analysis provides additional deployment-oriented evidence. Perturbing terminal-anchor inputs leads to gradual degradation rather than abrupt collapse

within the tested noise range, suggesting that the model can still use moderately uncertain day-ahead demand as a demand-level reference. The tested noise levels exceed the observed historical MAPE between day-ahead and real-time demand in the ISO New England dataset, providing confidence that the mechanism can tolerate realistic levels of prior uncertainty. Anchor quality nevertheless remains important. When the day-ahead prior becomes unreliable, fallback strategies such as historical-input forecasting or rolling historical proxies may be needed. Representative prediction curves are consistent with metric-based results: the proposed framework tracks load trajectories more effectively in selected SEMA and CT medium- or long-horizon cases, whereas VT exhibits stronger residual dispersion and greater sensitivity to distribution shift.

Overall, the experimental results support a model-selection view rather than a single-model ranking view. EA-BiGRU-Multi is most suitable when medium- or long-horizon forecasts can benefit from both nonlinear temporal representation learning and reliable forecast-origin demand priors. DLinear remains a strong option for short-horizon extrapolation and some regular regional load regimes. PatchTST is more useful when forecasting must rely only on historical inputs at longer horizons. These findings suggest that regional STLF systems should choose forecasting models according to horizon, zone, information availability, and operational reliability requirements.

4. Conclusions

EA-BiGRU-Multi was developed for regional multi-horizon short-term electricity load forecasting, with emphasis on settings where multiple lead times are predicted jointly and day-ahead demand information is available at the forecast origin. The framework combines feature-channel recalibration, shared bidirectional gated recurrent unit-based temporal encoding, horizon-specific forecasting heads, and terminal-anchor injection. Instead of repeating known-future demand values across the historical window, terminal-anchor injection places the day-ahead demand prior only at the final input frame, thereby preserving its role as forecast-origin information rather than treating it as an ordinary historical covariate.

Results on ISO New England regional load data show that EA-BiGRU-Multi is competitive in selected medium- and long-horizon cases when reasonably accurate day-ahead demand priors are available. In the Southeast Massachusetts six-step-ahead case, the mean coefficient of determination increases from 0.7604 to 0.8346, and the mean absolute error decreases from 100.44 MW to 79.07 MW compared with the linear baseline. Its advantage, however, is not uniform across all zones, horizons, or metrics. The linear baseline remains strong for short-horizon extrapolation and in some regular regional load regimes, while the patch-based Transformer shows stronger behavior under several historical-input-only medium- and long-horizon settings. These findings suggest that regional short-term load forecasting should not be reduced to a single architecture-level ranking. Forecast horizon, load-zone characteristics, information availability, and metric preference jointly determine which model is more appropriate for a given operating scenario.

Ablation and diagnostic analyses further clarify where the proposed framework derives its value. Removing the terminal-anchor information leads to substantial degradation in several six-step and 24-step-ahead cases, confirming the importance of forecast-origin demand priors for longer lead times. Naive broadcasting of known-future demand across the historical window produces less stable results than terminal-anchor injection, indicating that the temporal placement of known-future information matters. The cross-architecture anchor ablation on the linear baseline shows that the usefulness of day-ahead demand priors is not specific to the proposed recurrent architecture. At the same time, the historical-input-only comparison confirms that EA-BiGRU-Multi does not dominate when this prior is removed. The contribution of the framework therefore lies in the coordinated use of

temporal representation learning and forecast-origin-consistent prior integration, rather than in claiming architecture-level dominance.

This study has several limitations that should be acknowledged. The experiments are conducted on a single year of data from one independent system operator, which limits the assessment of temporal generalization across multi-year operating conditions and different market structures. The three selected load zones, while providing regional diversity, do not cover all possible demand profiles encountered in large-scale power systems. The day-ahead demand prior is assumed to be available at the forecast origin, which may not hold in all operational settings or in markets with different clearing timelines. The noise sensitivity analysis uses synthetic Gaussian perturbations rather than the full empirical error distribution of day-ahead demand, and the anchor-dropout regularization experiment is conducted for a limited number of cases. The model produces three discrete horizon outputs rather than a continuous forecast trajectory, which may limit its applicability in settings that require dense multi-step predictions.

Several directions remain open for further study. Multi-year and multi-market datasets would be useful for examining temporal and spatial generalization under broader operating conditions. Additional external factors, including extreme weather events, electricity prices, demand response signals, and calendar-specific operating patterns, may further improve forecasting under complex load dynamics. The reliability of day-ahead demand inputs also deserves closer investigation. When such inputs become uncertain or unavailable, adaptive fallback strategies, including anchor-dropout regularization during training, history-driven forecasting, or rolling historical proxies, may improve the operational stability of multi-horizon load forecasting systems. Extending the framework to produce dense multi-step prediction trajectories and incorporating probabilistic output representations are also directions that merit investigation.

Author Contributions: Conceptualization, F.C.; methodology, J.-H.Z.; software, J.-H.Z.; validation, F.C., J.-H.Z. and G.-H.J.; formal analysis, F.C. and J.-H.Z.; investigation, F.C. and J.-H.Z.; resources, F.C. and G.-H.J.; data curation, J.-H.Z.; writing—original draft preparation, F.C. and J.-H.Z.; writing—review and editing, G.-H.J.; visualization, J.-H.Z.; supervision, F.C. and G.-H.J.; project administration, F.C. and G.-H.J.; funding acquisition, F.C. and G.-H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Dongfang Electric Machinery Co., Ltd. (Contract No. 7300003521); the National Natural Science Foundation of China (Grant Nos. 51301068 and 11174078); and the Natural Science Foundation of Hebei Province (Grant Nos. E2019502060 and E2019502089).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are available from ISO New England public load and demand data sources. Processed data and code can be made available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank ISO New England for providing public load and demand data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hong, T.; Fan, S. Probabilistic electric load forecasting: A tutorial review. *Int. J. Forecast.* **2016**, *32*, 914–938. [\[CrossRef\]](#)
2. Hong, T.; Pinson, P.; Wang, Y.; Weron, R.; Yang, D.; Zareipour, H. Energy forecasting: A review and outlook. *IEEE Open Access J. Power Energy* **2020**, *7*, 376–388. [\[CrossRef\]](#)
3. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: Theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. [\[CrossRef\]](#)

4. Hong, T.; Pinson, P.; Fan, S. Global Energy Forecasting Competition 2012. *Int. J. Forecast.* **2014**, *30*, 357–363. [\[CrossRef\]](#)
5. Hippert, H.S.; Pedreira, C.E.; Souza, R.C. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Trans. Power Syst.* **2001**, *16*, 44–55. [\[CrossRef\]](#)
6. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015.
7. Taylor, J.W. Short-term electricity demand forecasting using double seasonal exponential smoothing. *J. Oper. Res. Soc.* **2003**, *54*, 799–805. [\[CrossRef\]](#)
8. Fan, S.; Hyndman, R.J. Short-term load forecasting based on a semi-parametric additive model. *IEEE Trans. Power Syst.* **2012**, *27*, 134–141. [\[CrossRef\]](#)
9. Ko, C.N.; Lee, C.M. Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter. *Energy* **2013**, *49*, 413–422. [\[CrossRef\]](#)
10. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
11. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
12. Lloyd, J.R. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *Int. J. Forecast.* **2014**, *30*, 369–374. [\[CrossRef\]](#)
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [\[CrossRef\]](#)
15. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [\[CrossRef\]](#)
16. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271. [\[CrossRef\]](#)
17. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling long- and short-term temporal patterns with deep neural networks. In Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 95–104. [\[CrossRef\]](#)
18. Shih, S.Y.; Sun, F.K.; Lee, H.Y. Temporal pattern attention for multivariate time series forecasting. *Mach. Learn.* **2019**, *108*, 1421–1441. [\[CrossRef\]](#)
19. Fekri, M.N.; Patel, H.; Grolinger, K.; Sharma, V. Deep learning for load forecasting with smart meter data: Online adaptive recurrent neural network. *Appl. Energy* **2021**, *282*, 116177. [\[CrossRef\]](#)
20. Eren, Y.; Küçükdemiral, İ.B. A comprehensive review on deep learning approaches for short-term load forecasting. *Renew. Sustain. Energy Rev.* **2024**, *189*, 114031. [\[CrossRef\]](#)
21. Salinas, D.; Flunkert, V.; Gasthaus, J. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [\[CrossRef\]](#)
22. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
23. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient Transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)
24. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
25. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11121–11128. [\[CrossRef\]](#)
26. Taieb, S.B.; Bontempi, G.; Atiya, A.F.; Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Syst. Appl.* **2012**, *39*, 7067–7083. [\[CrossRef\]](#)
27. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
28. ISO New England. *Zonal Information—SMD Hourly Data*; ISO New England: Holyoke, MA, USA, 2024. Available online: <https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/zone-info> (accessed on 5 April 2026).
29. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [\[CrossRef\]](#)
30. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. [\[CrossRef\]](#)
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [\[CrossRef\]](#)

32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [[CrossRef](#)]
33. Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41. [[CrossRef](#)]
34. Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4144–4147. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.