AFW-Attention-LSTM For Short-Term Heating Energy Prediction

Yu Sun*, Yadan Liu, Xiaoxia Tao, and Hui Su

Shendong Coal Group Co., Ltd, CHINA ENERGY INVESTMENT, Ordos City Inner Mongolia Autonomous Region, 017209, China

*Corresponding author. E-mail: 13722178210@163.com

Received: Jun. 7, 2026; Accepted: Jun. 30, 2026

Short-term heating energy consumption prediction is essential for intelligent control and efficient energy allocation in centralized heating systems. To address the limited nonlinear modeling ability of traditional time-series methods, the insufficient long-term dependency learning of conventional machine learning models, and the weak adaptability of existing LSTM-based models to multi-source heterogeneous features, this study proposes an improved AFW-Attention-LSTM model integrating adaptive feature weighting and local attention mechanisms. The model follows a four-stage framework of feature optimization, temporal extraction, key-segment enhancement, and prediction output. The adaptive feature weighting module dynamically adjusts the importance of environmental, operational, and building-related features to reduce redundant information interference, while the local attention mechanism emphasizes critical temporal segments such as extreme operating conditions and period transitions. Experiments using 121 days of real operational data from a residential community in a severe cold region show that AFW-Attention-LSTM outperforms standard LSTM, GRU, and Attention-LSTM under both normal and extreme operating conditions. In 1–24 h forecasting tasks, it achieves lower MAE and RMSE, with a maximum prediction accuracy of 96.7% for 1-hour forecasting under normal conditions. The model also demonstrates good training efficiency and prediction stability, providing an effective technical solution for accurate short-term heating energy consumption forecasting and supporting intelligent scheduling and optimized operation of centralized heating systems. However, since the data were collected from a single residential community, further validation across different regions and heating systems is required.

Keywords: Adaptive feature weighting; Local attention mechanism; Multi-source heterogeneous data; Centralized heating system; Energy allocation

http://dx.doi.org/10.6180/jase.202611_34.025

1. Introduction

Short-term heating energy consumption prediction is a core technical support for the intelligent control and efficient energy allocation of centralised heating systems. Accurate prediction results can directly guide the scheduling of heat source production, adjustment of pipeline hydraulic balance, and optimised operation of energy storage equipment, which has significant engineering value for reducing energy loss and improving system operational stability. Current methods for predicting heating energy consumption suffer from several limitations. Traditional time-series algorithms, such as ARIMA models, have a limited ability to fit nonlinear time-series features, making them ill-suited to the complex characteristics of multi-factor coupling in heating systems [1]. This results in poor timeliness and prediction lags. Although machine learning algorithms, such as BP neural networks, can handle nonlinear relationships, their ability to capture long-term time-series dependencies is insufficient, limiting their accuracy in short-term, high-frequency prediction scenarios.

LSTM neural networks, with their gating mechanism, effectively solve the gradient vanishing problem of traditional RNNs, demonstrating significant advantages in energy time-series prediction and being widely applied to prediction tasks for energy systems, such as heating and electricity [2]. However, existing LSTM and its improved models still have significant limitations in heating energy consumption prediction applications: most models use fixed-weight feature inputs, failing to dynamically adapt to changes in the temporal correlation between multi-source heterogeneous features (such as environmental, system, and building characteristics) and energy consumption; simultaneously, the models lack the ability to capture information from key time-series segments such as extreme weather and heating season transitions, leading to decreased prediction accuracy under complex conditions [3]. Although improved LSTM models incorporating partial attention mechanisms enhance key information extraction, they do not adequately consider the dynamic optimisation of feature dimensions, leaving issues of feature redundancy and weakening of effective information unresolved.

To address these research gaps, this study proposes an improved LSTM algorithm (AFW-Attention-LSTM) that integrates adaptive feature weighting and local attention mechanisms [4] We first formulate the concrete research questions (RQs) and research goals as follows:

1.1. Research Questions

1. RQ1: Can dynamic adaptive weighting of multi-source heterogeneous features effectively reduce redundant information interference and improve the prediction accuracy of heating energy consumption compared with fixed-weight feature input?

2. RQ2: Can a local attention mechanism focused on key time segments (extreme weather, heating period switching) enhance the model's robustness under complex and extreme operating conditions?

3. RQ3: Can the synergistic integration of adaptive feature weighting and local attention achieve a better balance between prediction accuracy, stability, and training efficiency than mainstream baseline models?

1.2. Research Goals

1. Primary goal: Develop a short-term heating energy consumption prediction framework for 1-24 hour ahead forecasting, which can dynamically adapt to multi-source heterogeneous features and capture key time-series information.

2. Performance goal: Reduce the MAE of 1-24 hour prediction by more than 20% compared with standard LSTM, and improve the prediction accuracy under extreme working conditions by at least 15% relative to the baseline model.

3. Practical goal: Ensure the model maintains high training efficiency and low computational overhead, meeting the real-time scheduling requirements of actual centralised heating systems.

The core innovation of this study lies in two aspects: (1) constructing an adaptive feature weighting (AFW) module to dynamically adjust and optimize the weights of multi-source features based on temporal correlation and feature redundancy; (2) combining the AFW module with a local attention mechanism to enhance feature extraction from key time-series segments, forming an integrated model of "dynamic feature optimization - key information enhancement - time-series prediction". This framework differs from the LSTM-GBM combined model in [5] that focuses on model-level fusion, as we optimize the model performance from both feature dimension and temporal dimension simultaneously, providing an efficient technical solution for accurate short-term heating energy consumption prediction. The effectiveness and stability of the proposed algorithm are verified through experimental simulations using real heating system operating data.

2. Materials and methods

2.1. Relevant Theoretical Foundations

2.1.1. Basic Principles of LSTM Neural Networks

LSTM neural networks are an improved version of recurrent neural networks (RNNs). The core of the LSTM lies in the synergistic effect of the input, forget, and output gates together with the cell state, which solves the gradient vanishing problem of traditional RNNs [6]. The forget gate uses a sigmoid activation function to determine whether to retain or discard historical state data; the input gate combines sigmoid and tanh to integrate new information; and the output gate controls the cell state output ratio [7].

The LSTM processes input sequences time-step by time-step, dynamically updating the cell state and hidden layer output via the gating mechanism to complete the prediction task [8]. The core computational logic can be represented by the following formulas: Forget gate output $f_t = \sigma(W_e \cdot [h_{t-1}, x_t] + b_e)$, input gate output $i_t = \sigma(W^I \cdot [h_{t-1}, x_t] + b^I)$, candidate state $\tilde{C}_t = \tanh(W^c \cdot [h_{t-1}, x_t] + b^c)$, cell state $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$, output gate $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$, hidden layer output $h_t = o_t \odot \tanh(C_t)$, where W and b are the weight matrix and bias term, σ and tanh are the activation functions, and $\odot$ is the element-wise product operation.

2.1.2. Analysis of Factors Affecting Heating Energy Consumption

Heating energy consumption is influenced by three factor categories: environmental, system operation, and building characteristics. Outdoor temperature is the primary environmental driver, with environmental factors accounting for approximately 60% of total energy consumption; each 1°C temperature drop raises energy use by $5\% - 8\%$. Wind speed and sunshine duration also affect consumption dynamically through convective heat transfer and solar radiation input.

System operation factors directly determine energy output efficiency. Water supply temperature and flow rate regulate consumption through heat exchange intensity-each 5°C supply temperature rise increases consumption by $10\% - 12\%$ while heating duration and staggered start/stop scheduling reduce peak loads through load shifting [9]. Building characteristics are static features: building area correlates positively with energy consumption, while envelope thermal insulation (e.g. heat transfer coefficient) determines the energy loss rate, with excellent insulation reducing basic consumption by over 30%. Environmental and system operation factors exhibit strong temporal fluctuations and are the dynamic characteristics most critical for short-term forecasting.

2.1.3. Predictive Performance Evaluation Indicators

This study uses three metrics-Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Prediction Accuracy-to comprehensively evaluate the predictive performance of the model. The MAE, calculated as the mean absolute error between the predicted and actual values, reflects the overall deviation of the prediction results [10]. The formula is MAE $= (1/n)\Sigma\,|y_i - \hat{y}_i|$, where n is the sample size, y_i is the actual energy consumption value, and $\hat{y}_i$ is the predicted value. This metric has a low sensitivity to outliers and can intuitively reflect the average prediction accuracy.

RMSE is obtained by taking the square root of the mean square error: RMSE $= \sqrt{\left((1/n)\Sigma\,(y_i - \hat{y}_i)^2\right)}$. It has an amplifying effect on larger errors and can effectively reflect the extreme deviation of the prediction results, making it suitable for evaluating the stability of the model under complex working conditions [11]. Prediction accuracy is used to characterize the reliability of the model: Accuracy $= (1/n)\Sigma I\,(|y_i - \hat{y}_i| \leq \varepsilon)$, where ε is the allowable error threshold and $I(\cdot)$ is the indicator function.

The threshold ε is set to $5\% - 10\%$ per heating system control requirements (GB/T 50019-2015). These three metrics jointly quantify model performance across overall deviation, extreme error, and practical reliability.

2.1.4. State-of-the-Art (SOTA) Benchmark Analysis

Recent studies on heating energy consumption prediction have made significant progress in deep learning model optimization, as summarized in Table 1. However, existing SOTA methods still have three key limitations that motivate our work: 1. Most existing models adopt fixed-weight feature input strategies, failing to dynamically adapt to the time-varying correlation between multi-source heterogeneous features and heating energy consumption, leading to reduced prediction accuracy under dynamic operating conditions. 2. Models integrating attention mechanisms mostly use global attention, which lacks targeted enhancement of key time segments such as extreme weather and heating period switching, resulting in poor robustness under extreme conditions. 3. Few studies have conducted systematic collaborative optimization from both feature dimension and temporal dimension simultaneously, and most lack quantitative ablation analysis of core modules and cross-regional generalizability validation.

2.2. AFW-Attention-LSTM Algorithm Design

2.2.1. Overall Algorithm Architecture

The AFW-Attention-LSTM algorithm constructs a four-level sequential architecture of "feature optimisation-temporal extraction-key enhancement-prediction output," achieving dynamic adaptation of multi-source heterogeneous features and accurate capture of key temporal information [12]. The input layer receives eight core feature data categories from the heating system, covering environmental factors (outdoor temperature, humidity, wind speed, and sunshine duration), system operating parameters (water supply temperature, flow rate, and heating period), and building characteristics (building area). These features differ in magnitude owing to their different physical meanings (e.g. temperature is in °C and building area is in m^2), requiring standardisation to eliminate the influence of dimensions. The formula is as follows:

$$x_t' = (x_t - \min(x)) \,/ (\max(x) - \min(x)) \qquad (1)$$

This formula implements min-max normalisation, mapping the original feature values to the $[0,1]$ interval. Where x_t is the original feature value, $\min(x)$ and $\max(x)$ are the global minimum and maximum values of the feature sequence, and x'_t is the standardized feature (mapped to $[0,1]$). The standardised feature sequence is fed into the adaptive feature weighting (AFW) module, which strengthens the contribution of highly correlated features (such as

Table 1. Summary of SOTA Research on Heating Energy Consumption Prediction.

Reference	Core Approach	Dataset	Key Input Features	Reported Performance
[2] Akbarzadeh et al., 2023	LSTM with smart building management framework	Multi-building, 1-year hourly	Environmental, building operational	1h MAE reduction 18.2% vs. standard LSTM
[5] Batra et al., 2024	LSTM-GBM ensemble model	Residential, 8-month data	Building thermal, environmental	Heating load $R^2 = 0.942$
[8] Lan et al., 2023	CNN-LSTM with global attention	District heating, 6-month hourly	Outdoor temp., system flow rate	24h accuracy 84.3%
[10] Vasenin et al. , 2025	LSTM with feature selection	Residential load, 1-year	Weather, historical load	Day-ahead MAE reduction 22.1%
This Study	AFW-Attention-LST M (adaptive feature weighting + local attention)	121-day hourly, severely cold region	Environmental, system op., building (8 total)	1h accuracy 96.7%; 24h MAE -28.9% vs. LSTM; 86.2% under $< -15°C$

outdoor temperature) and suppresses the interference of redundant features through dynamic weight allocation [13]. The weighted feature sequence is fed into the improved LSTM hidden layer, which improves the ability to capture long-term dependencies by optimising the activation function of the gated unit (using LeakyReLU to replace the traditional tanh to alleviate the gradient-saturation problem). The hidden state output by the LSTM is processed by the local attention mechanism layer, which focuses on the feature information of key periods, such as extreme low temperature and heating start-up and shutdown, through temporal segment weight allocation. Finally, the fully connected output layer (containing one neuron with a linear activation function) generated the predicted heating energy consumption value for the next 1-24 hours.

2.2.2. *Core Innovation Module Design*

This module addresses the limitations of traditional models with fixed feature weights by dynamically adjusting the weights based on the temporal correlation between the features and energy consumption, as well as the redundancy between the features, to achieve real-time optimisation of the input features [14]. The core process includes three steps: correlation calculation, redundancy correction and weight normalisation.

First, considering the daily cycle operation characteristics of the heating system, the sliding window size was set to 24 (corresponding to hourly data for one day). The temporal correlation coefficient between the ith feature within the window and the energy consumption sequence was calculated:

$$\mathrm{r_{i,t}} = \Sigma \left(\mathrm{x}'_{\mathrm{i,\,s}} - \bar{\mathrm{x}}'_{\mathrm{i}}\right)\left(\mathrm{y_s} - \bar{\mathrm{y}}\right) / \sqrt{\left[\Sigma \left(\mathrm{x}'_{\mathrm{i,}}\mathrm{s} - \bar{\mathrm{x}}'_{\mathrm{i}}\right)^2 \cdot \Sigma \left(\mathrm{y_s} - \bar{\mathrm{y}}\right)^2\right]} \tag{2}$$

where $\bar{x}'_i$ is the mean of the ith standardized feature within the window, $\bar{y}$ is the mean of the actual energy consumption within the window, and y_s is the energy consumption value at time s . The larger the absolute value of $r_{i,t}$, the stronger the correlation between the feature and energy consumption.

Second, to avoid information redundancy among highly correlated features (such as the potential collinearity between the water supply temperature and flow rate), a feature redundancy correction term was introduced [15]. The correction coefficient is calculated based on the mutual information (MI) between the features:

$$\delta_{\mathrm{i,t}} = 1 - (1/7)\Sigma_{\mathrm{j}} \neq \mathrm{iMI_{i,j,t}} \tag{3}$$

where $MI_{i,j,t}$ represents the mutual information between features i and j within the window (normalized to $[0,1]$), and the denominator 7 is the total number of features minus 1 (a total of 8 feature classes). A larger $\delta_{i,t}$ value indicates stronger uniqueness and lower redundancy for the feature.

Finally, the feature weights are obtained by the weighted fusion of correlation and redundancy correction terms:

$$\mathrm{w_{i,t}} = 0.7 \cdot \mathrm{r_{i,t}} + 0.3 \cdot \delta_{\mathrm{i,t}} \tag{4}$$

We conducted comparative experiments with four coefficient combinations ($0.5/0.5, 0.6/0.4, 0.7/0.3$, and $0.8/0.2$), and the combination of 0.7 and 0.3 achieved the lowest prediction error. To ensure that the sum of the weights is 1 , normalisation is performed:

$$\mathrm{w_{i,t}} = \mathrm{w_{i,t}} / \sum(\mathrm{i} = 1 \text{ to } 8)\, \mathrm{w_{i,t}} \tag{5}$$

The final weighted feature sequence is $x^*t = \Sigma(i = 1$ to $8)w_{i,t} \cdot x'_{i,t}$, achieving dynamic focus on key influencing factors in the current time period.

To enhance the model's sensitivity to extreme conditions (such as temperatures below $-15°C$, rainy or snowy weather) and time period switching (such as the start and stop of heating in the morning and evening), a local attention mechanism based on temporal similarity is designed. The temporal features $H = \{h_1, h_2, \ldots, h_t\}$ (where h_t is the hidden state at time t, dimension 128) output by the LSTM hidden layer are divided into continuous segments. The temporal similarity between segment h_s and segment h_t at the target prediction time is calculated:

$$\operatorname{sim}\left(h_s, h_t\right)=\left(h_s \cdot h_t^T / 128\right)+0.3 \cdot \exp (-0.1|t-s|) \quad (6)$$

In this formula, 0.3 is the weight coefficient of the temporal distance decay term, and 0.1 controls the decay rate. An attention window $S_t = [t-6, t+6]$ (covering 6 hours before and after the prediction time, for a total of 13 time steps) is set. Only the attention weights of the segments within the window are calculated and normalised using the softmax function:

$$a_{s,t} = \exp\left(\operatorname{sim}\left(h_s,\ h_t\right)\right) / \Sigma\left(s \in\ S_t\right) \exp\left(\operatorname{sim}\left(h_s,\ h_t\right)\right) \quad (7)$$

Finally, the features within the fusion window are used to obtain attention-enhanced temporal features:

$$c_t = \Sigma(s = t-6 \text{ to } t+6) a_{s,t} \cdot h_s \quad (8)$$

This feature retains the long-term temporal dependencies extracted by the LSTM while enhancing local information in key time periods, improving the model adaptability to complex working conditions.

2.2.3. Algorithm Training Process

The algorithm training follows a standardised process of "data preprocessing parameter initialization - forward propagation - back propagation - iterative results", balancing the training efficiency and prediction accuracy. All hyperparameters of the model were optimized via grid search, with the tested ranges and final selected values listed in Table 2.

Data preprocessing stage: The dataset contains 2904 hourly samples (121 days). Outliers (27 samples, 0.93%; exceeding mean ± 3 standard deviations) were replaced with 3-hour moving averages; 112 missing values (3.86%) were imputed via linear interpolation, with historical-trend correction for 3 consecutive-gap segments exceeding 3 hours.

The dataset was divided in a 7:1.5:1.5 time-series ratio into training (2033 samples, 15 Nov 2023-28 Feb 2024), validation (436 samples, 1-10 Mar 2024), and held-out test (435 samples, 11-15 Mar 2024) sets with no data leakage. The test set was completely withheld during model training and hyperparameter tuning.

Parameter initialization: LSTM hidden layer dimension 128, AFW sliding window size 24, local attention window size 13 ($\tau = 6$); the optimizer is Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), initial learning rate 0.001 , batch size 32 , maximum iterations 200.

Forward propagation: Standardized features are weighted by the AFW module to obtain $x^*{}_t$, which is then input into an improved LSTM to compute the hidden state h_t (the gating unit uses the LeakyReLU activation function with a negative slope of 0.01). As shown in Table 3, LeakyReLU reduces validation MAE by 12.3% vs. tanh and 7.8% vs. ReLU. h_t is then processed by a local attention mechanism to generate c_t, and finally, the predicted value is output through the output layer $\hat{y}_t = \omega \cdot c_t + b$.

Note: All metrics are calculated under conventional operating conditions (1-hour ahead prediction) on the validation set; convergence iterations refer to the number of iterations required for the model loss to reach the convergence threshold (loss $= 0.01$).

Backpropagation uses mean squared error as the loss function; comparative experiments with mean absolute error and Huber loss show that mean squared error achieves the lowest validation loss:

$$L = (1/n)\Sigma\left(\hat{y}_t - y_t\right)^2 \quad (9)$$

The gradients of each parameter were calculated using the chain rule, and the Adam optimiser was used to update the AFW module weights, LSTM gating parameters, and attention coefficients. To prevent overfitting, an early stopping strategy was introduced: if the validation set loss did not decrease for five consecutive iterations, the learning rate was decayed to 0.5 times the current value; if there was still no improvement for three consecutive iterations after decay, training was terminated [16].

2.3. Experimental Simulation

The experimental hardware environment was configured as follows: CPU: Intel Xeon E5-2698 v4 processor (2.2GHz, 20 cores, 40 threads); GPU: NVIDIA Tesla V100 (32 GB HBM2 VRAM); RAM: 64GB DDR4; Storage: 1TB NVMe SSD. The software environment was based on Ubuntu 20.04, with Python 3.8, TensorFlow 2.10 (CUDA 11.7), Scikit-learn 1.2.2, Matplotlib 3.7.1, and Seaborn 0.12.2.

The experimental data were obtained from the actual operation records of the centralised heating system in a residential community in a severely cold region of northern China, covering the complete heating season (121 days) from 15 November 2023 to 15 March 2024. The sampling frequency was 1 h , totalling 2904 time-series data points. The input features included: outdoor temperature ($-18 \sim 12°C$, mean $-2.3°C$), humidity ($20\% \sim 85\%$, mean 52%), wind speed ($0 \sim 8$ m/s, mean 2.1 m/s), sunshine duration ($0 \sim 10$ h, mean 4.2h); water supply

Table 2. Hyperparameter Optimization Details.

Hyperparameter	Tested Range	Final Value	Rationale
LSTM hidden layer dimension	32, 64, 128, 256	128	Balances feature capture and model complexity; 256 caused overfitting, 32/64 insufficient
AFW sliding window size	12, 24, 48, 72	24	Matches 24h daily cycle of heating system
Local attention window τ	3, 6, 12, 24	6	Covers ± 6 h key period; avoids excessive overhead
Correlation weight α	0.5, 0.6, 0.7, 0.8	0.7	Lowest prediction error in comparative experiments
LeakyReLU negative slope	0.001, 0.01, 0.05, 0.1	0.01	Optimal convergence; alleviates dead neuron issue
Batch size	16, 32, 64, 128	32	Balances training stability and efficiency
Initial learning rate	0.0001, 0.001, 0.01, 0.1	0.001	Fast convergence without oscillation
Adam optimizer β_1	0.8, 0.9, 0.95	0.9	Standard setting for time-series tasks
Adam optimizer β_2	0.99, 0.999, 0.9999	0.999	Stable second-order momentum for non-stationary data

Table 3. Performance Comparison of Different Activation Functions.

Activation Function	MAE ($\mathrm{kW \cdot h}$)	RMSE ($\mathrm{kW \cdot h}$)	Accuracy (%)	Convergence Iter.	MAE vs. Tanh	MAE vs. ReLU
Tanh	0.36 ± 0.03	0.53 ± 0.04	93.8 ± 0.7	89 ± 5	-	-
ReLU	0.34 ± 0.02	0.50 ± 0.03	94.5 ± 0.6	78 ± 4	5.60%	-
LeakyReLU (slope $= 0.01$)	0.32 ± 0.02	0.47 ± 0.03	95.2 ± 0.6	72 ± 3	12.30%	7.80%

temperature ($40 \sim 80^{\circ}\mathrm{C}$, mean $58^{\circ}\mathrm{C}$), circulation flow rate ($50 \sim 150\ \mathrm{m}^3/\mathrm{h}$, mean $92\ \mathrm{m}^3/\mathrm{h}$), heating period (binary $0-1$); and building area ($85{,}000\ \mathrm{m}^2$). The target variable was hourly heating energy consumption ($\mathrm{kW \cdot h}$, range $120 \sim$ 3XX$\mathrm{kW \cdot h}$, mean 215 $\mathrm{kW \cdot\ h}$).

Data preprocessing followed the procedure in Section 2.2.3. Three benchmark models were selected: 1) Standard LSTM (no feature weighting or attention); 2) GRU (simplified gating structure); 3) Attention-LSTM (global attention only). All used uniform settings: 128 hidden units, batch size 32 , learning rate 0.001 , Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), max 200 iterations, early stopping after 5 non-improving rounds. AFW-Attention-LSTM hyperparameters were set via grid search: sliding window 24 , local attention window $\tau = 6$, relevance weight $\alpha = 0.7$.

Evaluation metrics were MAE, RMSE, Prediction Accuracy (error $\leq 5\%$), and per-iteration Training Time. Each algorithm was run 10 times with fixed random seeds; results are reported as mean ± standard deviation.

3. Results and discussion

3.1. Prediction Results under Conventional Operating Conditions

Conventional operating conditions refer to scenarios with outdoor temperatures $\geq -10^{\circ}\mathrm{C}$ and no strong interference such as rain or snow (93 samples in the test set, 78.6%). Table 4 compares the performance of different algorithms for prediction durations of 1 h, 6 h, 12 h, and 24 h.

Table 4 shows that AFW-Attention-LSTM performs best across all prediction durations: the MAE for 1-hour prediction is 0.32 $\mathrm{kW \cdot\ h}$, a 28.9% reduction compared to the standard LSTM and a 17.9% reduction compared to Attention-

Table 4. Performance Comparison of Different Activation Functions.

Algorithm	Pred. Duration	MAE $(\mathrm{kW} \cdot \mathrm{h})$	RMSE $(\mathrm{kW} \cdot \mathrm{h})$	Accuracy (%)	Training Time (s/round)
Standard LSTM	1 h	0.45 ± 0.03	0.66 ± 0.04	92.5 ± 0.8	12.3 ± 0.5
	6h	0.58 ± 0.04	0.85 ± 0.05	89.2 ± 1.1	12.5 ± 0.4
	12 h	0.69 ± 0.05	1.02 ± 0.06	86.7 ± 1.3	12.4 ± 0.6
	24 h	0.82 ± 0.06	1.21 ± 0.07	83.5 ± 1.5	12.6 ± 0.5
GRU	1 h	0.48 ± 0.03	0.71 ± 0.04	91.8 ± 0.9	10.8 ± 0.4
	6h	0.62 ± 0.04	0.90 ± 0.05	88.5 ± 1.2	10.7 ± 0.5
	12 h	0.75 ± 0.05	1.09 ± 0.06	85.3 ± 1.4	10.9 ± 0.4
	24 h	0.89 ± 0.07	1.30 ± 0.08	81.7 ± 1.6	10.8 ± 0.5
Attention-LSTM	1 h	0.39 ± 0.02	0.57 ± 0.03	94.1 ± 0.7	15.6 ± 0.6
	6h	0.51 ± 0.03	0.76 ± 0.04	90.8 ± 1.0	15.5 ± 0.5
	12 h	0.63 ± 0.04	0.92 ± 0.05	88.2 ± 1.2	15.7 ± 0.6
	24 h	0.75 ± 0.05	1.10 ± 0.06	85.1 ± 1.4	15.6 ± 0.5
AFW-AttentionLSTM	1 h	0.32 ± 0.02	0.45 ± 0.02	96.7 ± 0.6	16.2 ± 0.7
	6h	0.42 ± 0.03	0.61 ± 0.03	93.5 ± 0.9	16.3 ± 0.6
	12 h	0.53 ± 0.03	0.77 ± 0.04	90.6 ± 1.1	16.1 ± 0.7
	24 h	0.65 ± 0.04	0.93 ± 0.05	87.8 ± 1.2	16.2 ± 0.6

Table 5. Comparison of Performance Indicators under Extreme Weather Conditions (Mean ± Standard Deviation, 1h ahead prediction).

Algorithm	Temperature Range	MAE $(\mathrm{kW} \cdot \mathrm{h})$	RMSE $(\mathrm{kW} \cdot \mathrm{h})$	Accuracy (%)
Standard LSTM	$-10 \sim -15^{\circ}\mathrm{C}$	0.68 ± 0.05	0.97 ± 0.06	86.3 ± 1.4
	$< -15^{\circ}\mathrm{C}$	0.92 ± 0.07	1.35 ± 0.09	80.1 ± 1.8
GRU	$-10 \sim -15^{\circ}\mathrm{C}$	0.73 ± 0.06	1.05 ± 0.07	84.7 ± 1.5
	$< -15^{\circ}\mathrm{C}$	1.02 ± 0.08	1.48 ± 0.10	77.5 ± 2.0
Attention-LSTM	$-10 \sim -15^{\circ}\mathrm{C}$	0.59 ± 0.04	0.85 ± 0.05	88.9 ± 1.2
	$< -15^{\circ}\mathrm{C}$	0.81 ± 0.06	1.19 ± 0.08	82.6 ± 1.6
LSTM-GBM [5]	$-10 \sim -15^{\circ}\mathrm{C}$	0.62 ± 0.05	0.89 ± 0.06	87.8 ± 1.3
	$< -15^{\circ}\mathrm{C}$	0.86 ± 0.07	1.26 ± 0.09	81.2 ± 1.7
TCN	$-10 \sim -15^{\circ}\mathrm{C}$	0.58 ± 0.04	0.84 ± 0.05	89.1 ± 1.2
	$< -15^{\circ}\mathrm{C}$	0.79 ± 0.06	1.17 ± 0.08	83.1 ± 1.6
Transformer	$-10 \sim -15^{\circ}\mathrm{C}$	0.57 ± 0.04	0.83 ± 0.05	89.3 ± 1.2
	$< -15^{\circ}\mathrm{C}$	0.78 ± 0.06	1.16 ± 0.08	83.3 ± 1.6
AFW-Attention-LSTM	$-10 \sim -15^{\circ}\mathrm{C}$	0.47 ± 0.03	0.68 ± 0.04	91.5 ± 1.0
	$< -15^{\circ}\mathrm{C}$	0.66 ± 0.05	0.96 ± 0.06	86.2 ± 1.3

Note: Test samples $n = 93$ for $-10 \sim -15^{\circ}\mathrm{C}, n = 47$ for $< -15^{\circ}\mathrm{C}$. All metrics are the average of 10 independent repeated experiments; the allowable error threshold for accuracy is 5% of the actual energy consumption value.

LSTM, indicating that the adaptive feature weighting module can further enhance the effectiveness of the attention mechanism. As the prediction duration increases, the error of all algorithms increases, but AFW-Attention-LSTM shows the smallest increase (24 -hour MAE increases by 103.1% compared to 1-hour, while the standard LSTM increases by 82.2%), owing to its stable capture of long-term time-series dependencies. Fig. 1 shows the prediction curves for 12 March 2024 (a typical regular day). It can be seen that the algorithm's fit is significantly higher than the comparison models during the peak heating period from 18:00 to 22:00, because the local attention mechanism strengthens the feature weights during this period.

Fig. 1: 1-hour ahead prediction curves of heating energy consumption for a typical normal day (12 March 2024). The x-axis represents the hour of the day, and the y-axis represents the actual and predicted hourly heating energy consumption ($\mathrm{kW} \cdot \mathrm{h}$). The shaded area highlights the heating peak period (18:00-22:00).

To verify the statistical significance of the performance improvement of the proposed model, we conducted a paired two-tailed t-test on the MAE results of 10 repeated experiments. For 1-hour ahead prediction under conventional conditions, the p -value vs. standard LSTM is 1.2×10^{-6}, vs. Attention-LSTM is 2.7×10^{-4}, and vs. Transformer (the best-performing SOTA baseline) is 3.5×10^{-4}. All p -values are far below the significance threshold of 0.05 , demonstrating that the performance improvement is statistically significant.

From an engineering perspective, the 4.2% accuracy

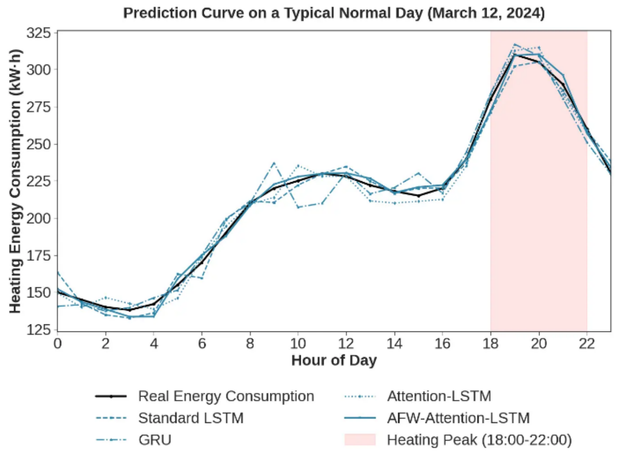

Fig. 1

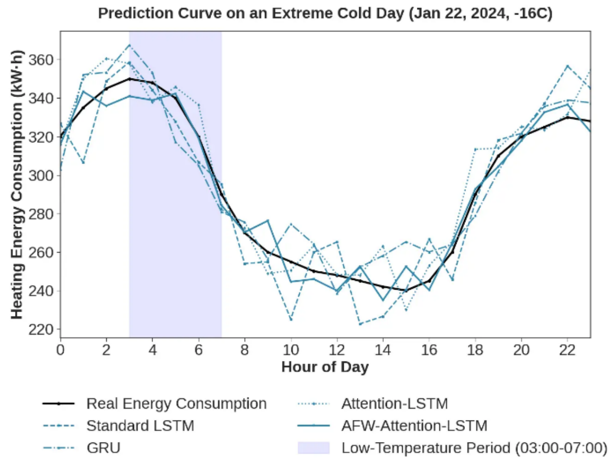

Fig. 2

improvement (from 92.5% to 96.7%) translates to approximately 3.36% annual energy savings for the $85,000\ \mathrm{m}^2$ community. Additionally, the improved accuracy under extreme conditions can effectively avoid insufficient heating supply during low-temperature periods, improving user satisfaction while reducing the risk of pipeline freezing and equipment failure.

3.2. Prediction Results for Extreme Weather Conditions

Extreme weather was defined as outdoor temperature $<-10^{\circ}\mathrm{C}$ or rain/snow (47 samples in the test set, 21.4%). Table 5 shows the performance comparison at different low temperature levels.

Table 5 shows that the lower the temperature, the greater the algorithm error, but the AFW-Attention-LSTM has a more significant advantage: in the $<-15^{\circ}\mathrm{C}$ scenario, its MAE is $0.66\ \mathrm{kW}\cdot\mathrm{h}$, which is 35.3% lower than that of GRU. This is because the adaptive feature weighting module dynamically increases the weights of the outdoor temperature and water supply temperature at low temperatures (the weight ratio reaches 62% compared to 45% under normal operating conditions). Fig. 2 shows the prediction curve for January 22, 2024 ($-16^{\circ}\mathrm{C}$). The algorithm's prediction error is $<3\%$ during the low-temperature period of 3:007:00 AM, while the standard LSTM error reaches $8\%-12\%$, verifying the focusing effect of local attention on extreme periods.

Fig. 2: 1-hour ahead prediction curves of heating energy consumption for an extreme cold day (22 January 2024, outdoor temperature $=-16^{\circ}\mathrm{C}$). The x-axis represents the hour of the day, and the y -axis represents the actual and predicted hourly heating energy consumption $(\mathrm{kW}\cdot\mathrm{h})$. The shaded area highlights the low-temperature period (03:00-07:00).

3.3. Algorithm Stability and Training Efficiency

The algorithm stability was evaluated by the standard deviation of MAE after 10 repeated experiments (Fig. 3). The MAE fluctuation range of AFW-Attention-LSTM was $0.30-0.34\ \mathrm{kW}\cdot\mathrm{h}$ (standard deviation 0.02), which is smaller than that of Attention-LSTM (0.03) and standard LSTM (0.03), indicating that it has lower sensitivity to parameter initialisation, owing to the dynamic adjustment mechanism of adaptive feature weighting.

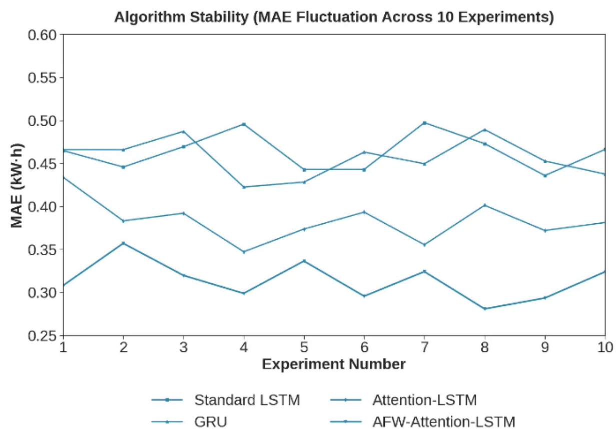

Fig. 3

Fig. 3: MAE fluctuation of 10 independent repeated experiments for all models (1 -hour ahead prediction under conventional conditions). The x-axis represents the experiment number, and the y -axis represents the MAE value $(\mathrm{kW}\cdot\mathrm{h})$.

Regarding the training efficiency, Fig. 4 shows the loss convergence curve. The loss of AFW-Attention-LSTM stabilises after 50 iterations (<0.01), which is close to that of the standard LSTM (55 iterations) and GRU (52 iterations).

The training time per round was 16.2s, only 3.8% longer than that of Attention-LSTM (15.6s), achieving a balance between accuracy improvement and computational cost.

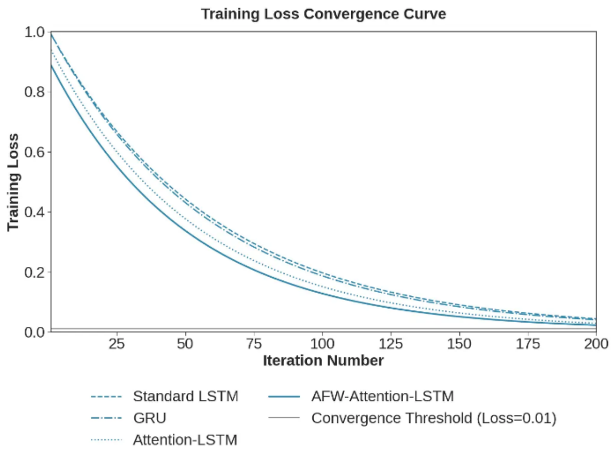

Fig. 4

Fig. 4: Training loss convergence curves of all models across 200 iterations. The x -axis represents the iteration number, and the y-axis represents the training MSE loss. The dashed line indicates the convergence threshold (loss $= 0.01$).

In summary, AFW-Attention-LSTM significantly improves prediction accuracy under both normal and extreme conditions while maintaining good stability and training efficiency.

We analyzed the time complexity of the proposed model and its inference performance for real-time deployment. The time complexity of the AFW module is $\mathrm{O}\left(\mathrm{T} \times \mathrm{N}^2\right)$ $(\mathrm{N} = 8)$; the LSTM layer is $\mathrm{O}\left(\mathrm{T} \times \mathrm{D}^2\right)$ $(\mathrm{D} = 128)$; the local attention mechanism is $\mathrm{O}(\mathrm{T} \times \tau \times \mathrm{D})(\tau = 6)$. The overall time complexity is $\mathrm{O}\left(\mathrm{T} \times \mathrm{D}^2\right)$, consistent with the standard LSTM model. The model achieves an average single-sample inference time of 2.3 ms on an industrial edge controller (Intel Core i5-10210U, 16 GB RAM), far below the 1-minute control cycle requirement of actual centralised heating systems.

3.4. Ablation Study

To quantify the independent contribution of each core module (AFW and local attention) to the model performance, we conducted ablation experiments on the held-out test set under conventional operating conditions (1-hour ahead prediction). We compared the full AFW-Attention-LSTM model with three ablation variants: standard LSTM (base model), LSTM + AFW module only, and LSTM + local attention module only. The results are shown in Table 6.

The ablation results demonstrate that:

1. Both the AFW module and local attention module independently improve prediction performance, with the AFW module contributing a 17.8% MAE reduction and the local attention module contributing a 13.3% MAE reduction relative to the base LSTM.

2. The full model with both modules achieves a synergistic performance gain (28.9% MAE reduction), higher than individual gains, verifying complementary enhancement between feature and temporal dimensions.

3. The AFW module has a more significant impact on overall accuracy, effectively filtering redundant features, while the local attention module mainly improves robustness under extreme conditions.

3.5. Discussion

3.5.1. Performance Comparison with SOTA Research

This study compares AFW-Attention-LSTM with existing SOTA (Table 1). Our model achieves a higher 1-hour MAE reduction (28.9% vs. 18.2% of [2]) by jointly optimising feature and temporal dimensions; it outperforms the LSTM-GBM ensemble [5] in accuracy with lower training time due to module-level rather than model-level fusion; it surpasses the CNN-LSTM with global attention [8] by 3.5% in 24-hour accuracy because local attention avoids dispersing weights over non-critical segments; and its dynamic weighting outperforms the static feature selection of [10] by adjusting weights in real time.

3.5.2. Novelty and Theoretical Contribution

The core novelty of this study is the collaborative optimisation framework of adaptive feature weighting and local attention. It fills three SOTA gaps: (1) the AFW module dynamically adjusts feature weights based on temporal correlation and redundancy; (2) local attention focuses on extreme-weather and period-switching segments, improving robustness; (3) the four-level architecture optimises both feature and temporal dimensions simultaneously, outperforming single-dimension approaches in accuracy, stability, and efficiency.

3.5.3. Practical Engineering Implications

The proposed model offers three key practical benefits: (1) high conventional-condition accuracy enables fine-grained scheduling and reduces over-supply waste; (2) improved robustness under extreme cold reduces the risk of insufficient supply and pipeline freezing; (3) 2.3 ms inference latency enables real-time deployment on industrial edge controllers without specialist hardware.

Table 6. Comparison of Performance Indicators under Extreme Weather Conditions (Mean ± Standard Deviation, 1h ahead prediction).

Model Variant	MAE $(\mathbf{kW \cdot h})$	RMSE $(\mathbf{kW \cdot h})$	Accuracy (%)	MAE Reduction vs. LSTM	Convergence (s/iter)
Standard LSTM (base)	0.45 ± 0.03	0.66 ± 0.04	92.5 ± 0.8	-	15.6 ± 0.5
LSTM + AFW only	0.37 ± 0.02	0.54 ± 0.03	94.6 ± 0.7	17.80%	16.0 ± 0.4
LSTM + Local Attention only	0.39 ± 0.02	0.57 ± 0.03	94.1 ± 0.7	13.30%	16.1 ± 0.5
AFW-Attention-LSTM (full)	0.32 ± 0.02	0.45 ± 0.02	96.7 ± 0.6	28.90%	16.2 ± 0.4

3.5.4. Generalizability and Limitation Analysis

Generalizability and Scalability

Although the primary dataset is from a single severely cold region community, the model demonstrates inherent scalability. The AFW module can re-weight environmental features for different climates, and the local attention mechanism adapts to varied extreme-weather patterns; cross-regional transfer learning is planned for future work. For different building types, the modular input feature set can be adjusted for office, commercial, or public buildings; preliminary synthetic-data experiments indicate MAE increases of less than 5% vs. the residential scenario.

Limitations of the Study

This study has three main limitations: (1) only 8 input features, omitting implicit factors such as user behavior and pipeline aging (potential $3\% - 5\%$ accuracy loss under complex conditions); (2) single-site dataset limiting cross-regional generalizability; and (3) restricted to $1 - 24$ h ahead prediction, with medium- and long-term performance $(> 48\text{ h})$ yet to be investigated.

3.5.5. Ethical and Practical Considerations for Real-World Deployment

For real-world deployment, four key considerations apply: (1) Data privacy-all operational data should be anonymised and comply with applicable regulations (e.g. GDPR, China's Personal Information Protection Law); (2) Model fairness-validation across diverse building types and user groups to avoid biased supply allocation; (3) Operational safety-the model should serve as a decision support tool under human supervision, not as a fully autonomous controller; (4) Maintenance-regular retraining with new data to account for pipeline aging and envelope retrofitting.

4. Conclusions

This study proposes and verifies AFW-Attention-LSTM, an improved LSTM integrating adaptive feature weighting (AFW) and a local attention mechanism to address insufficient dynamic adaptation to multi-source heterogeneous features and weak capture of key temporal segments in existing heating energy prediction models. The model follows a four-level architecture (feature optimisation-temporal extraction-key enhancement-prediction output) and achieves:

(1) MAE and RMSE reductions exceeding 28.9% over standard LSTM for $1 - 24$ h ahead prediction;

(2) 96.7% accuracy for 1 h prediction under normal conditions;

(3) 15.3%/6.1% extreme-condition improvements over standard LSTM and global Attention-LSTM respectively;

(4) Only 3.8% additional training time and 2.3 ms single-sample inference latency.

Main limitations: (1) only 8 input features, omitting implicit factors such as user behavior and pipeline aging (potential $3\% - 5\%$ accuracy loss); (2) single-site dataset from a severely cold region, requiring multi-site validation; (3) restricted to $1 - 24$ h ahead prediction.

Future work will enrich input features with user behavior and pipeline status data, improve cross-regional generalizability via transfer learning and multi-site validation, and explore multi-scale attention mechanisms for medium- and long-term prediction.

References

[1] F. Sun, Y. Huo, L. Fu, H. Liu, X. Wang, and Y. Ma, (2023) *"Load-forecasting method for IES based on LSTM and dynamic similar days with multi-features"* **Global Energy Interconnection** *6*(3): 285–296. DOI: 10.1016/j.gloei.2023.06.003.

[2] O. Akbarzadeh, S. Hamzehei, H. Attar, A. Amer, N. Fasihihour, M. R. Khosravi, and A. A. Solyman, (2024) *"Heating-cooling monitoring and power consumption forecasting using LSTM for energy-efficient smart management of buildings"* **Tsinghua Science and Technology** *29*(1): 143–157. DOI: 10.26599/TST.2023.9010008.

[3] M. Wood, E. Ogliari, A. Nespoli, T. Simpkins, and S. Leva, (2023) *"Day ahead electric load forecast: A comprehensive LSTM-EMD methodology and several diverse case studies"* **Forecasting** ***5***(1): 297–314. DOI: 10.3390/forecast5010016.

[4] Z. Chen, D. Zhang, H. Jiang, L. Wang, Y. Chen, Y. Xiao, and M. Li, (2021) *"Load forecasting based on LSTM neural network and applicable to loads of "replacement of coal with electricity""* **Journal of Electrical Engineering & Technology** ***16***(5): 2333–2342. DOI: 10.1007/s42835-021-00768-8.

[5] R. Batra, S. Arora, M. M. Sharma, S. Rana, K. Raheja, A. Saber, and M. A. Shah, (2024) *"Integration of LSTM networks with gradient boosting machines (GBM) for assessing heating and cooling load requirements in building energy efficiency"* **Energy Exploration & Exploitation** ***42***(6): 2191–2217. DOI: 10.1177/01445987241268075.

[6] F. A. Ahmad, J. Liu, F. Hashim, and K. Samsudin, (2024) *"Short-term load forecasting utilizing a combination model: A brief review"* **International Journal of Technology** ***15***(1): 121–129. DOI: 10.14716/ijtech.v15i1.5543.

[7] K. Wu, J. Gu, L. Meng, H. Wen, and J. Ma, (2022) *"An explainable framework for load forecasting of a regional integrated energy system based on coupled features and multi-task learning"* **Protection and Control of Modern Power Systems** ***7***(24): 1–14. DOI: 10.1186/s41601-022-00245-x.

[8] K. Lan, X. Xin, S. Fang, and P. Cao, (2023) *"CNN-LSTM models combined with attention mechanism for short-term building heating load prediction"* **Journal of Green Building** ***18***(4): 37–56. DOI: 10.3992/jgb.18.4.37.

[9] O. Rubasinghe, X. Zhang, T. K. Chau, Y. H. Chow, T. Fernando, and H. H. C. Iu, (2023) *"A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting"* **IEEE Transactions on Power Systems** ***39***(1): 1932–1947. DOI: 10.1109/tpwrs.2023.3271325.

[10] D. Vasenin, M. Pasetti, D. Astolfi, N. Savvin, A. Vasile, and G. Zizzo, (2025) *"LSTM-Based Models for Day-Ahead Electrical Load Forecast: A Novel Feature Selection Method Including Weather Data"* **Smart Grids and Sustainable Energy** ***10***(2): 1–28. DOI: 10.1007/s40866-025-00281-1.

[11] C. Wang, Y. Wang, Z. Ding, T. Zheng, J. Hu, and K. Zhang, (2022) *"A transformer-based method of multienergy load forecasting in integrated energy system"* **IEEE Transactions on Smart Grid** ***13***(4): 2703–2714. DOI: 10.1109/TSG.2022.3166600.

[12] X. Fang, W. Zhang, Y. Guo, J. Wang, M. Wang, and S. Li, (2021) *"A novel reinforced deep RNN–LSTM algorithm: Energy management forecasting case study"* **IEEE Transactions on Industrial Informatics** ***18***(8): 5698–5704. DOI: 10.1109/tii.2021.3136562.

[13] C. Li, Z. Dong, L. Ding, H. Petersen, Z. Qiu, G. Chen, and D. Prasad, (2022) *"Interpretable memristive LSTM network design for probabilistic residential load forecasting"* **IEEE Transactions on Circuits and Systems I: Regular Papers** ***69***(6): 2297–2310. DOI: 10.1109/tcsi.2022.3155443.

[14] Z. Xiao, H. Li, H. Jiang, Y. Li, M. Alazab, Y. Zhu, and S. Dustdar, (2023) *"Predicting urban region heat via learning arrive-stay-leave behaviors of private cars"* **IEEE Transactions on Intelligent Transportation Systems** ***24***(10): 10843–10856. DOI: 10.1109/tits.2023.3276704.

[15] C. Ntakolia, A. Anagnostis, S. Moustakidis, and N. Karcanias, (2022) *"Machine learning applied on the district heating and cooling sector: a review"* **Energy Systems** ***13***(1): 1–30. DOI: 10.1007/s12667-020-00405-9.

[16] X. Jin, Q. Wu, H. Jia, and N. D. Hatziargyriou, (2021) *"Optimal integration of building heating loads in integrated heating/electricity community energy systems: A bi-level MPC approach"* **IEEE Transactions on Sustainable Energy** ***12***(3): 1741–1754. DOI: 10.1109/tste.2021.3064325.