Explainable district heat load forecasting with active deep learning

Huang, Yaohui; Zhao, Yuan; Wang, Zhijin; Liu, Xiufeng; Liu, Hanjing; Fu, Yonggang

Published in:
Applied Energy

Link to article, DOI:
10.1016/j.apenergy.2023.121753

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Explainable district heat load forecasting with active deep learning
Yaohui Huang\textsuperscript{a, d, 1}, Yuan Zhao\textsuperscript{c, 1}, Zhijin Wang\textsuperscript{a, 1}, Xiufeng Liu\textsuperscript{b, 4}, Hanjing Liu\textsuperscript{a}, Yonggang Fu\textsuperscript{a}

\textsuperscript{a} College of Computer Engineering, Jimei University, Yinjiang Road 185, 361021 Xiamen, China
\textsuperscript{b} College of Electronic Information, Guangxi Minzu University, Daxue East Road 188, 530006 Nanning, China
\textsuperscript{c} School of Business Administration, South China University of Technology, Wushan Road 381, 510630, Guangzhou, China
\textsuperscript{d} Department of Technology, Management and Economics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

\begin{abstract}
District heat load forecasting is a challenging task that involves predicting future heat demand based on historical data and various influencing factors. Accurate forecasting is essential for optimizing energy production and distribution in district heating systems. However, most existing forecasting models lack transparency and interpretability and fail to capture the spatial–temporal dependencies in the data. Moreover, they often require a large amount of annotated data for training, which can be costly and time-consuming to obtain. In this paper, we present a novel approach to district heat load forecasting, which involves predicting future heat demand based on historical data and various influencing factors. The proposed approach is based on an Active Graph Recurrent Network (Ac-GRN), which leverages the strengths of active deep learning and graph neural networks to capture the complex spatial–temporal dependencies in the data. The approach also provides explainability for its predictions by using correlation-based attribution methods. The active deep learning component can effectively select the most informative and representative samples from a large pool of data, reducing the frequency and cost of data collection and human effort. The graph neural network component can model both linear and nonlinear relationships among heat meters using bidirectional recurrent connections, enhancing the accuracy and robustness of the predictions. We conduct extensive experiments and compare our approach with eleven state-of-the-art models on a real-world dataset of district heating consumption data from Danish residential buildings. Our results show that our approach outperforms other models in terms of accuracy, robustness, reliability, and computational efficiency for multi-horizon multi-step district heat load forecasting. Our approach also provides meaningful explanations for its predictions by highlighting the most influential factors and heat meters for each prediction. This paper makes a novel contribution to district heat load forecasting with explainability.
\end{abstract}

1. Introduction

District heating is a system for distributing heat generated in a centralized location, such as a power plant or a waste incinerator, through a pipeline network to residential, commercial, and industrial buildings for space heating and water heating [1]. District heating can provide energy-efficient, cost-effective, and environmentally friendly heating solutions for urban areas, especially in cold climates [2]. However, to optimize the operation and planning of district heating systems, accurate forecasting of the district heat load is essential. District heat load is the amount of heat demand that needs to be met by the heat supplier at a given time [3]. The forecasting of the district heat load is a challenging task due to the non-linear and non-stationary nature of heat demand, which is influenced by various factors such as weather conditions, building characteristics, human behavior, and calendar effects [4–6] (see Table 1).

District heat load forecasting is a challenging task that involves predicting future heat demand based on historical data and various influencing factors such as weather conditions and building characteristics [3, 7]. The non-linear and non-stationary nature of heat demand, coupled with the multitude of influencing factors, makes accurate forecasting difficult [6]. Therefore, researchers have explored advanced machine learning (ML) techniques, particularly deep learning, for their ability to learn complex patterns and dependencies from data [8, 9]. Among these techniques, active deep learning is a promising approach that allows the model to actively query the most informative data points

\* Corresponding authors.
\textsuperscript{1} Equal contributor

\begin{flushleft}
https://doi.org/10.1016/j.apenergy.2023.121753

Received 4 June 2023; Received in revised form 25 July 2023; Accepted 10 August 2023

Available online 21 August 2023

0306-2619/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
from a large and unlabeled dataset, thus improving the forecasting accuracy while diminishing the costs associated with data collection [10]. This approach is well suited for district heat load forecasting, where data can be highly variable and complex [8].

District heating systems need accurate and transparent forecasting models to optimize their heat production and supply operations. However, most existing models that use active deep learning techniques for district heat load forecasting are not very interpretable. These models are like "black boxes" that do not explain how they make their predictions or what features are important [11]. This can be a major drawback for critical applications like district heating, where explainability is important for trust, reliability, and decision support [12]. Moreover, many models do not capture the spatial-temporal dependencies in the data, which are essential for modeling the interrelationships between different heat meters and their temporal dynamics [13]. Therefore, there is a need for a novel approach that can achieve both high accuracy and high explainability for district heat load forecasting. One possible way to address these challenges is to leverage the strengths of active deep learning and graph neural networks. Active deep learning is an approach where the model actively queries the most informative data points using an uncertainty-based sampling strategy and its use of graph neural networks to model the interdependencies between different heat meters. Moreover, our approach provides explainability by using correlation-based methods that highlight the spatial features that influence the predictions. Our approach has several advantages over existing methods that use active deep learning or graph neural networks for district heat load forecasting. First, it enhances forecasting accuracy and diminishes data collection costs by implementing a strategy based on uncertainty sampling. Second, it captures the complex spatial-temporal patterns in the data by using graph neural networks. Third, it provides explainability by using correlation-based methods.

The key contributions of this paper are threefold:

• We propose a novel approach to district heat load forecasting using an Active Graph Recurrent Network, which leverages the strengths of active deep learning and graph neural networks to capture the complex spatial-temporal dependencies in the data.
• We provide explainability for our model's predictions by employing relationship-inspired methods that utilize correlations among heat meters. This approach highlights the spatial features in the input that contribute to pattern discovery and prediction.
• We evaluate our approach on real-world district heating data and compare it with 11 state-of-the-art models in terms of accuracy and robustness. We also conduct ablation studies to analyze the impact of different components of our model, as well as computational efficiency compared with other models.

The rest of the study is structured as follows. Section 2 reviews the literature on heat load forecasting and AL. Section 3 describes the proposed Ac-GRN model. Section 4 conducts the evaluation. Section 6 discusses the contributions and limitations of the proposed model. Section 7 concludes the paper and presents the future research directions.

2. Related work

In this section, we review the existing literature on district heat load forecasting, and highlight the main challenges and limitations of current methods. We also review the literature on active deep learning and graph neural networks, and explain how they can address some of the challenges and limitations of existing methods. Finally, we summarize the main contributions of our paper and how it differs from the existing literature.

2.1. District heat load forecasting

District heat load forecasting is a challenging task that involves predicting future heat demand based on historical data and various influencing factors such as weather conditions and building characteristics [3,19]. Accurate forecasting is essential for optimizing energy production and distribution in district heating systems [2]. Existing methods for district heat load forecasting can be broadly categorized into physical models, black-box models [20], and gray-box models [21]. Physical models use equations that describe the physical behavior of a system to predict an output, while black-box models use supervised ML methods where measurements of input and output variables of a system are collected, and then used to mathematically describe the system [22]. Gray-box models provide a balance between high accuracy and good generalization capabilities, by extracting the mathematical model/structure from the system's physics, and estimation of model parameters from measured data [23,24]. Physical models have the advantage of being easily interpretable, but they require a large number of parameters and detailed information about the system as input.
which can be costly and time-consuming to obtain [22]. Moreover, physical models suffer from poor generalization capabilities, as they may not capture the non-linear and non-stationary nature of heat demand [25]. Black-box models have the advantage of being easy to build and capable of learning complex patterns and dependencies from data [12]. However, black-box models require a large amount of data for training purposes, which may not be available or reliable in some cases [26]. Moreover, black-box models lack transparency and interpretability, as they provide little insight into how they make their predictions [27]. Gray-box models have the advantage of combining the strengths of physical models and black-box models, by exploiting the domain knowledge and data-driven learning [28]. However, gray-box models also face some challenges, including determining the optimal model structure and parameters, and dealing with incomplete or noisy data [29].

Among these three categories of methods, black-box models have received increasing attention lately, owing to their efficacy in learning complex patterns and dependencies from data [3]. Black-box models are ML techniques that do not reveal the internal logic or structure of the model, but only provide the input-output relationship [30]. In particular, black-box models such as artificial neural network (ANN) [31], convolutional neural network (CNN) [32], long short-term memory network (LSTM) [8], attention mechanism [8], extreme gradient boosting (XGBoost) [33], and among others, have been applied to district heat load forecasting with promising results. For example, Xue et al. [17] used SVR, DNN, and XGBoost to validate the accuracy and stability of black-box models for multi-step ahead heat load forecasting in district heating system (DHS). Dang et al. [18] employed XAI methods, such as SHAP and PDP, to explain the significance of exogenous features for district heating prediction using XGBoost. Li et al. [34] proposed a heat load forecasting model based on gated recurrent units with time-varying features using LSTM and attention mechanism. However, these methods require large storage space and high costs of data collection, which can be prohibitive in many cases and affect the efficiency in DHS. In addition, collecting and transferring excessive data for DHS may lead to expensive labor and communication overhead. To address these challenges, we propose an AI-based framework for district heat load prediction, which can utilize less heat meter data to achieve accurate prediction for all smart heat meters in a certain time period. With the AI framework, we select representative heat meters for training the prediction model based on an uncertainty-based sampling strategy, while excluding heat meters with similar patterns.

2.2. Active deep learning and graph neural networks

High accuracy, cost-effectiveness, and timeliness are important goals in the field of energy consumption monitoring [27]. Active deep learning offers a potential approach to enhancing cost-efficiency in data acquisition and storage by selecting the most informative or uncertain samples for training, as per specific criteria or strategies [14]. This method optimizes the learning process by focusing on samples that the model finds most challenging to classify or predict [35]. Specifically, the most uncertain samples are those that the model has the most difficulty classifying or predicting. In the context of classification tasks, these typically are the samples where the predicted class probabilities closely resemble a uniform distribution, implying that the model struggles to assign a clear class to these samples. For regression tasks, these are often the samples that yield the widest confidence interval for the model’s prediction. This strategic selection of high-uncertainty samples empowers the model to maximize information extraction from a minimal number of samples, thereby fostering accelerated learning and enhanced accuracy. For example, Settles et al. [36] proposed several query strategies, e.g., uncertainty sampling, query-by-committee, expected model change, and expected error reduction. Zhang et al. [10] utilized active learning to counter data bias and achieve high performance in building energy forecasting. Aryandoust et al. [37] proposed a spatio-temporal, multi-modal method for electric load forecasting, which exhibits improved accuracy. Wang et al. [38] designed a selector–predictor framework for electric load forecasting. Active deep learning can be enhanced through the combination with various techniques, such as transfer learning [39], semi-supervised learning [40], and reinforcement learning [41]. These approaches can further improve its performance and expand its applicability.

Graph neural networks have been widely used in various domains such as social networks [42], knowledge graphs [43], computer vision [44], natural language processing [45], etc., where data has rich structural or relational information. Graph neural networks can capture the complex spatial-temporal dependencies in the data by modeling the interdependencies between different nodes using graph structures [15]. For example, Kipf et al. [46] proposed a graph convolutional network (GCN) that can learn node representations by aggregating information from their neighbors using spectral graph theory. Li et al. [47] proposed a gated graph neural network (GGNN) that can learn node representations by propagating information through recurrent neural networks using spatial graph theory.

In this paper, we propose to use active deep learning and graph neural networks for district heat load forecasting. Our approach can address some of the challenges and limitations of existing methods by actively learning from the most informative data samples and modeling the inter-dependencies between different heat meters using graph structures. To the best of our knowledge, this is the first attempt to predict district heat load utilizing a graph neural network and AI framework.

3. Materials

This study uses the public dataset of district heating consumption from residential buildings in Aalborg, Denmark [48]. The dataset covers 3,127 smart heat meters with hourly readings for three years (2018–2020). It also includes contextual information such as dwelling type, construction year, and energy efficiency level. This study primarily targets the prediction of heat load for single-family houses, terraced houses, and apartments, which account for 3,021 smart meters in the dataset. Smart heat meters lacking dwelling type and those that are unoccupied (a total of 105 samples) are excluded from the focus of this study. Additionally, meteorological data from a nearby weather station was collected. This data has nine variables, such as outdoor temperature, solar radiation Intensity, wind speed, and others, that may affect heat consumption patterns.

Due to the unavoidable interference for the transmitting process, there are about 0.3% smart heat meter data missing. To ensure compatibility with the model and adhere to the original data processing method described in [48], missing data from the smart meters is filled using a weighted moving average imputation. This method utilizes a symmetric window size of 48 (including 48 preceding and following values) and applies a linear weighting scheme. The meteorological data is complete and devoid of missing values. Then, we conducted a correlation analysis to evaluate the relationship between various meteorological variables and the mean value of heat load. We visualized the results of the correlation analysis using a heatmap (see Fig. 1), which shows the Spearman’s rank correlation coefficient (SCC) between each pair of variables. The SCC measures the monotonic relationship between two variables, regardless of whether the relationship is linear or nonlinear. Additionally, the SCC is robust to outliers and apt for handling data that may not follow the same distribution. The coefficient values of the SCC range from −1 to 1. A value of −1 denotes a perfect negative correlation, a value of 0 signifies no correlation, and a value of 1 corresponds to a perfect positive correlation. The heatmap displays the correlation between variables, with darker colors indicating stronger correlations and lighter colors indicating weaker correlations. The diagonal cells display the correlation of each variable with itself, invariably equal to 1.
As illustrated in Fig. 1, outdoor temperature and grass temperature emerge as the most significant meteorological variables for heat load forecasting. The strong correlation between these two factors highlights their closely aligned trend. To reduce the redundancy of temperature information, we choose the outdoor temperature as an exogenous feature in our study. Outdoor temperature directly influences the energy required for heating and cooling in residential and commercial buildings [49]. Solar radiation intensity also significantly correlated with heat load. The solar radiation might contribute to reducing the heating load by naturally warming the building. This effect is particularly significant in structures designed for passive solar heating, where the building’s design and orientation synergize to maximize the capture and retention of solar heat [50]. Hence, solar radiation intensity is included as an exogenous meteorological factor in our study. Relative humidity and cloud cover both exhibit a relatively minor impact on heat load. Given the established correlation between relative humidity and building heat transfer, as confirmed by prior research [51], we have included relative humidity as the third meteorological factor in our analysis. The presence of moisture in the air can significantly impact the thermal characteristics of a building, consequently affecting the overall heat load. The variables of wind speed, air pressure, leaf moisture, and precipitation exhibit weak correlations with heat load data. To enhance the diversity of meteorological features without excessive inclusion, we have additionally included wind speed as a factor, as it demonstrates relatively lower correlation with the other three selected variables. Moreover, as wind speed increases, the rate of convective heat transfer from the building surfaces to the outside environment tends to increase. This prompts a need for more heating to keep the inside temperature steady, influencing the overall district heat load.

Above all, the factors of outdoor temperature, solar radiation intensity, relative humidity, and wind speed are selected as additional feature combination for our experiment. We summarized the statistics of heat load and these meteorological factors in Table 2, which shows the minimum, maximum, median, mean, standard deviation, Pearson correlation coefficient (PCC), and SCC. The PCC is another measure that quantifies the strength and direction of a linear relationship between two continuous variables. It also ranges from −1 to 1 and has a similar interpretation as the SCC. Finally, we normalized the data using min-max scaling, a technique that adjusts the range of variable values to a common scale, promoting model stability and learning efficiency.

4. Methodology

4.1. Overview

In this section, we provide a comprehensive description of the proposed Active Graph Recurrent Network (Ac-GRN) model. The Ac-GRN model combines active learning, graph convolutional components, and recurrent learning units to effectively capture both spatial and temporal dependencies in district heating data. An overview of our proposed model is illustrated in Fig. 2. The proposed Ac-GRN is composed of several key modules. First, the training and validation sets of heat load data are pre-processed and randomly segregated into annotated and unannotated groups for subsequent analysis. Second, the active learning module (ALM) is initiated using the annotated meters, while the unannotated meters are stored in the data center. The ALM consists of two components: the knowledge extractor and the feature reconstructor. The knowledge extractor is responsible for selecting the most informative samples using an uncertainty-based sampling strategy. The observations from the selected representative meters are then passed to the feature reconstructor component to reconstruct the unknown trends of the unselected heat load meters for prediction. Third, a graph recurrent network (GRN) is designed to capture the latent heat load patterns across multiple heat meters and make predictions. This network comprises three elements: (1) The temporal convolutional module (TCM) employs multi-layer graph convolutional networks in the temporal dimension, extracting temporal dependencies from the heat load data. (2) The spatio-temporal memory enhanced module (STME) augments the model’s capacity to capture dynamic spatial and temporal correlations by implementing a spatial attention mechanism and a recurrent component. (3) The multi-component fusion module (MCFM) fuses the linear representation of the input data with the output from the recurrent component to generate the final prediction. The linear representation is a simplified projection of the input data that can alleviate the vanishing gradient problem and enhances the expressiveness
of the model. With this network structure, the proposed AC-GRN model will be able to capture the complex spatial–temporal dependencies in district heating data, and actively learn from the most informative data samples, thereby improving the forecasting accuracy and reducing the costs of data acquisition.

4.2. Knowledge extractor by active learning

The knowledge extractor leverages active learning to efficiently identify the most informative samples among numerous heat meters. This method significantly curtails costs related to data collection and storage, while enhancing the efficiency of the learning process. As illustrated in Fig. 3, the module consists of two phases: a query phase and an update phase. Within the query phase, we develop a novel classifier known as the Temporal–Spatial Fully Connected Network (TSFCN). The TSFCN is combined with a temporal convolutional network (Temp-FCN), a spatial convolutional network (Spat-FCN), a fully connected network for output (Out-FCN), and two activation layers. The Temp-FCN and Spat-FCN are employed to capture the temporal and spatial patterns among annotated heat load observations, respectively. The Out-FCN and the softmax function are employed to compute the probabilities corresponding to each class. Specifically, this classifier learns patterns from annotated samples and subsequently classifies the category of unannotated meters. The detail processing of TSFCN classifier \( f_\theta(x) \) can be formulated as:

\[
f_\theta(x) = \text{softmax}(m_0 \odot (s((W_{a,1}^T x)W_{a,2})))W_{a,3} + b_a),
\]

where \( x \) is an unannotated sample from data center \( U \), which consists of unannotated heat meters, \( W_{a,1}, W_{a,2}, \) and \( W_{a,3} \) are the learnable parameters of Temp-FCN, Spat-FCN, and Out-FCN, respectively. \( \odot \) denotes the element-wise multiplication. \( m_0 \) is the dropout vector to alleviate overfitting, \( b_a \) is the bias term, softmax is the activation function that normalizes the output to a probability distribution over the possible classes. \( s \) is the transpose operation that changes the orientation of a matrix or a vector.

After obtaining the probability map, we incorporate an uncertainty sampling method to select the most uncertain samples, designated as representative meters for annotation. The uncertainty of the sample can be measured by different criteria, including entropy, least confidence, or margin. In this study, we employ the least confidence criterion, which selects the sample where the model exhibits the lowest confidence. The least confidence criterion can be formulated as follows:

\[
x^* = \arg \min_{x \in U} f_\theta(x),
\]

where \( x^* \) are samples from the unannotated data pool \( U \). \( f_\theta(x) \) is the predicted label of sample \( x \) by the model with parameters \( \theta \), which is trained on a subset of annotated meters.

In the update phase, the identified representative meters are incorporated into the annotated data subset and concurrently removed from the unannotated data center \( U \). The classifier is subsequently retrained to update the model parameters. This iterative process, involving both the query strategy and update process, continues until the proportion of annotated to unannotated meters achieves a specified target ratio or a predetermined number of iterations is reached. The targeted annotated data ratio is a hyperparameter that controls the trade-off between the data collection cost and the model performance.

4.3. Feature reconstructor for incomplete data

As illustrated in Fig. 4, traditional district heat load prediction utilizes historical records (including recent trends) to forecast the target period for each heating meter. However, when using the active learning framework, the unrepresentative meters inevitably lack information regarding recent short-term trends. This absence can significantly impact the accuracy and efficiency of predictions. Incorporating recent short-term trends can effectively help capture rapid changes, thus allowing for more agile and accurate responses to dynamic circumstances in real-time [49]. To address this issue, our feature reconstructor module leverages observations from representative heat meters to supplement the incomplete data acquired through the Active Learning framework, as shown in Fig. 5. This module consists of two components: a relation discovery module and a representation generation module.

The relation discovery module establishes associations between unannotated heat meters and selected representative heat meters based on their correlation. Specifically, we first calculate the PCC for each pair of unannotated and annotated heat meter data. The trend among

---

Table 2
The descriptive statistics and correlation coefficients of heat load and meteorological factors. The standard deviation (STD) measures the variability of each variable. The \( p \)-value indicates the significance level of the correlation coefficients. A \( p \)-value less than 0.001 (***) means that the correlation is very unlikely to occur by chance. The Pearson correlation coefficient (PCC) and Spearman’s rank correlation coefficient (SCC) measure the strength and direction of the linear relationship between each meteorological factor and the median heat load.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Number of time series</th>
<th>Min</th>
<th>Max</th>
<th>Medium</th>
<th>Mean</th>
<th>STD</th>
<th>PCC</th>
<th>SCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdoor temperature</td>
<td>1</td>
<td>-7.883</td>
<td>23.712</td>
<td>8.450</td>
<td>9.159</td>
<td>6.138</td>
<td>-0.953***</td>
<td>-0.966***</td>
</tr>
<tr>
<td>Solar radiation</td>
<td>1</td>
<td>1.804</td>
<td>391.483</td>
<td>94.046</td>
<td>126.140</td>
<td>106.669</td>
<td>-0.709***</td>
<td>-0.738***</td>
</tr>
<tr>
<td>Relative humidity</td>
<td>1</td>
<td>45.758</td>
<td>99.892</td>
<td>84.252</td>
<td>82.422</td>
<td>10.930</td>
<td>0.370***</td>
<td>0.424***</td>
</tr>
<tr>
<td>Wind speed</td>
<td>1</td>
<td>1.150</td>
<td>12.829</td>
<td>4.685</td>
<td>4.996</td>
<td>2.059</td>
<td>0.102***</td>
<td>0.101***</td>
</tr>
<tr>
<td>Heat load</td>
<td>3021</td>
<td>0</td>
<td>462.328</td>
<td>37.110</td>
<td>45.610</td>
<td>36.526</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

---

Fig. 2. The overview of the proposed Ac-GRN method.

---

5
heat load meters may exhibit some common characteristics or usage patterns [52], PCC is more suitable to measure these linear relationships. Then, for each unselected heat load meter, we choose the top-$K$ most correlated meters to construct a sparse correlation coefficient matrix $\mathbf{C}$, where $c_{i,j}$ denotes the correlation value between heat meter $i$ and $j$. This method can reduce the information redundancy inherent in these relationships, and reconstruct the sequence by exploiting the low-rank relationship structure [53]. The learning process of the sparse relationship can be formulated as follows:

$$\arg \min_{W \in \mathcal{W}} f_p(W_{i_1}^\beta \cdot x_{i_1}^*, \ldots, W_{i_K}^\beta \cdot x_{i_K}^*).$$

(3)

where $x_i^*$ represents the correlated representative samples for the $i$th unannotated meter, and $f_p$ is the function for relation learning. $W_j^\beta$ denotes the learnable weighting parameters for the $j$th unannotated meter.

The representation generation module operates as a linear weighting process based on the learned weighting matrix $W_j^\beta$. This process adaptively emphasizes certain relevant heat meters in the sparse matrix by multiplying them with their corresponding weights. The aim of this is to improve the prediction of the heat load for the target, specifically for the unannotated meters. The generation process can be described as follows:

$$M_i = W_j^\beta \cdot [x_{i_1}^*, \ldots, x_{i_K}^*].$$

(4)

$M_i$ represents the generation for the $i$th unannotated meter, and $W_j^\beta$ is the obtained relational weighting matrix. $M_i$ encompasses trend information that can provide linear characteristics for the prediction model, thereby helping to reduce the effects of missing data in the heat load predictions for unannotated heat meters.

4.4. Graph recurrent network

The overview of graph recurrent network (GRN) is shown in Fig. 6, which consists of three main parts: temporal convolutional module, spatio-temporal memory enhanced module, and multi-component fusion module. They are described in the following.

4.4.1. Temporal convolutional module

The temporal convolution component comprises two temporal graph convolution networks (TGCNs) along the temporal dimension and a single-layer convolution network (SCN) in the spatial dimension. The
TGCN module extracts temporal dependencies from adjacent time steps of each smart heat meter, while the SCN module captures spatial dependencies from all smart meters.

To apply the TGCN module, we need to transform the graph structure into an algebraic form that can capture its connectivity and other patterns. One way to do this is to use spectral graph theory and the Laplacian matrix, which are tools for analyzing the properties of graphs in the frequency domain. We define the multi-meters network as a directed graph $G = (V, E, A)$, where $V$ represents a finite set of $N$ nodes, corresponding to the smart meters; $E$ is the set of edges, representing the relationships between each heat meter and other meters; and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix of graph $G$. The normalized Laplacian matrix can be formulated as $L = I_{N \times N} - \frac{1}{2}AD^{-\frac{1}{2}}$, where $L \in \mathbb{R}^{T \times T}$ is the normalized Laplacian matrix; the diagonal matrix $D \in \mathbb{R}^{T \times T}$ is a degree matrix, $D^{ii} = \sum_i A^{ii}$; $I_{N \times N}$ denotes a unit matrix.

The embedding process of TGCN can be described as follows: First, we multiply the normalized Laplacian matrix with the input time series. This operation produces the temporal dependencies from the data. Third, we multiply the output of the previous operation by another learnable weight matrix. This operation enhances the model's ability to learn complex and robust features. Finally, we multiply the normalized Laplacian matrix with the input time series after applying a ReLU activation function and a dropout mask to introduce non-linearity and regularization. This operation yields the spatial attention score that dynamically adjusts the weights used in the spatial attention mechanism.

The recurrent component is built upon a gated recurrent unit (GRU) capable of learning long-term dependencies in sequential data. The GRU possesses two gates: a reset gate and an update gate. The reset gate determines the degree to which the previous hidden state is forgotten, while the update gate governs how much of the previous hidden state is updated with the new input.

Mathematically, the equations for these operations are:

$$W_{g,o} * g = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}MW_{g,o} + b_g,$$

$$R_{g,o} = W_{g,o} * g(m_o \odot \sigma(W_{g,0} * gM))),$$

where $g \odot$ denotes the graph convolution operation. $A = A + I_{N \times N}$ is an adjacent matrix with inserted self-loops. $M \in \mathbb{R}^{N \times N}$ is an adjacent matrix with inserted self-loops. $R_{g,o}$ is the final output of TGCN component. $\odot$ indicates element-wise multiplication, $\sigma(\cdot)$ is rectified linear unit (ReLU) activation, and $m_o$ is a binary mask denotes dropout processing. $W_{g,0}$ and $W_{g,o}$ are both the learnable weighting matrix, $b_g$ denotes the biases.

After capturing information from adjacent time steps for each node in the graph's temporal dimension, we proceed to integrate an SCN layer in the spatial dimension to emphasize important pattern information among all smart meters. The SCN module is a single-layer convolutional network that applies a one-dimensional convolutional operation to the output of the TGCN along the spatial dimension. This operation involves sliding a kernel or filter over each node's features and computing a dot product. The size of the kernel dictates the number of heat meters considered in each convolution step. Capable of extracting spatial features, the convolutional module effectively discerns similarities or differences between nodes based on their heat meter observations. The spatial convolution can be formulated as follows:

$$R_{c,o} = m_c \odot \sigma(W_{c,o} * R_{g,o}),$$

where $R_{g,o}$ denotes the generated representation of convolution component, $W_{c,o}$ is the learnable parameters of convolution kernel, $m_c$ is a binary mask matrix that prevents the networks from overfitting.

4.4.2. Spatio-temporal memory enhanced module

The spatio-temporal memory enhanced module is designed to further enhance the model's ability to capture dynamic spatial and temporal correlations by applying a spatial attention mechanism and a recurrent component. The spatial attention mechanism assigns different weights to different parts of the input based on their relevance, while the recurrent component stores and updates information over time. This spatial attention is defined as follows:

$$S_a = \text{softmax}(W^{(1)}_a \cdot \sigma(W^{(2)}_a R_{c,o}) + b_a),$$

$$R_{s,a} = R_{s,a} \odot S_a.$$

The recurrent component is built upon a gated recurrent unit (GRU) capable of learning long-term dependencies in sequential data. The GRU possesses two gates: a reset gate and an update gate. The reset gate determines the degree to which the previous hidden state is forgotten, while the update gate governs how much of the previous hidden state is updated with the new input.

The equations governing these gates are:

$$e_t = \sigma(W_{e,r}[h_{t-1}; R_{s,a}] + b_{e,r}),$$

$$u_t = \sigma(W_{u,r}[h_{t-1}; R_{s,a}] + b_{u,r}),$$

$$z_t = \tanh(W_{z,r}[e_t \odot h_{t-1}; R_{s,a}] + b_{z,r}),$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot z_t,$$

where $e_t$, $u_t$, and $z_t$ are the internal state of the update gate, reset gate, and candidate hidden state at time step $t$, respectively. $W_{e,r}$, $W_{u,r}$, and $W_{z,r}$ are the learnable weighting matrices. $b_{e,r}$, $b_{u,r}$, and $b_{z,r}$ are the correspondent biases. $\sigma(\cdot)$ denotes sigmoid activation function, and $[;]$ indicates the concatenation operation. In these equations, $h_{t-1}$ is the previous hidden state, and $R_{s,a}$ is the output of the attention component. The recurrent component takes $R_{s,a}$ as the input and updates $h_{t-1}$ using the three gates to produce $h_t$, which is the current hidden state. The integration of exogenous meteorological features necessitates an update to the GRU’s input, as detailed in the following equation:

$$Y_{i,s} = [R_{s,a}; W_e \cdot E].$$
where $\mathcal{L}$ represents the exogenous data. $Y_{t,o}$ signifies the input, combining observations from both the heat meters and meteorological features.

The embedding process maps the output of the recurrent component into an embedding space that matches the desired output shape. The embedding process consists of two steps: first, it applies a temporal embedding function $\text{Emb}_{\text{temp}}(\cdot)$ on $h_t$ to reduce its temporal dimension; second, it applies a spatial embedding function $\text{Emb}_{\text{sp}}(\cdot)$ on $\text{Emb}_{\text{temp}}(h_t)$ to reduce its spatial dimension. The equation for this process is:

$$
R_{t,o} = \text{Emb}_{\text{temp}}(\text{Emb}_{\text{sp}}(m_t \odot h_t)),
$$

where $R_{t,o} \in \mathbb{R}^{B \times L \times N}$ is the final output of STME module. $m_t$ denotes a binary mask matrix for the dropout layer.

4.4.3. Multi-component fusion module

The nonlinear modules can effectively capture potential dependencies in either the temporal or spatial dimension. However, excessive nonlinearity may result in the vanishing or exploding gradients problem, which in turn can adversely affect the prediction accuracy. Therefore, to balance the nonlinearity and robustness of the proposed Ac-GRN, we introduce a linear representation unit (LEU) that adds a linear combination of the input features $\mathcal{M}$ to the recurrent component output $R_{t,o}$ as the final prediction:

$$
\hat{Y}_{t+h; t+h+b-L} = R_{t,o} \oplus (W_{f,o}M + b_f),
$$

where $\hat{Y}_{t+h; t+h+b-L} \in \mathbb{R}^{B \times L \times N}$ is the predictive values, $\oplus$ indicates the addition operation, $W_{f,o}$ is learnable weighting parameters, and $b_f$ is correspondent bias.

We use the mean square error (MSE) as the loss function to measure the prediction accuracy and add an $L_2$ regularization term to prevent overfitting. The loss function can be defined as:

$$
\mathcal{L}_{\text{Loss}}(Y, \hat{Y}) = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{N} \sum_{l=1}^{L} (Y_{i,j,o} - \hat{Y}_{i,l,o})^2 + \frac{\lambda}{2} \sum_{i=1}^{L} w_i^2,
$$

where $K$ is the length of training samples, $N$ denotes the dimension of target data, and $L$ is the prediction steps. Specifically, $\lambda$ is a constant that denotes the regularization coefficient, $\Theta$ is the number of weights in the model. $\omega$ indicates the weight parameter of the model.

5. Experiment

5.1. Experimental setup

To ensure a fair comparison, the same constant parameters related to training are applied to all methods. The Adam optimizer [54] is used to obtain the training models, and the mean squared error (MSE) is selected as the loss function. The batch size is set as 32. The number of training epochs and the learning rate are adjusted to ensure each method’s optimal performance. To increase the dependenceability and stability of the results, every experiment is performed five times using different random seeds, with the reported results being the mean values. The grid search approach is employed to search for relatively optimal hyperparameters. The detailed settings for the hyperparameters of each method are provided in Table 9. The model was implemented using the deep learning framework, Pytorch v1.12.1 [55]. All experiments were carried out on a server equipped with Intel(R) Xeon(R) Gold 6226R CPU (2.90 GHz) with 128G memory and were accelerated by two NVIDIA RTX A6000 GPUs. In this study, we use the first 70% data for training (766 data points), the following 10% data (109 data points) as the validation set, and the remaining 20% data (219 data points) for the test. For the experiment with varying sampling radio, the input to the models was manipulated in the test set. Specifically, the input from unselected heat meters was set to zero for the baseline methods, while these observations were reconstructed using a feature reconstructor for the Ac-GRN model. This processing does not affect the data in the training or validation sets, ensuring that these changes do not influence the models during their training phase. Despite these adjustments, the prediction target remained the true value.

5.2. Baselines

We compare our proposed model with several state-of-the-art methods for multivariate time-series forecasting, including:

- **Autoregressive (AR):** a linear model that predicts the current value of a time series based on its past values [56].
- **Dilinear:** a linear regression model that uses the previous values of all time series as input features [57].
- **MTNet:** a multi-task learning model that combines convolutional neural networks and recurrent neural networks to capture both local and global temporal dependencies [58].
- **TPA:** a temporal pattern attention model that uses an attention mechanism to learn the importance of different temporal patterns [59].
- **LSTMNet:** a long- and short-term time-series network that combines convolutional neural networks and recurrent neural networks to capture both short-term and long-term dependencies [60].
- **LSTM:** a type of recurrent neural network that can learn long-term dependencies in sequential data using memory cells and gates [61].
- **ED (GRU):** an encoder–decoder model that uses gated recurrent units to encode the input sequence and decode the output sequence [62].
- **CRNN:** a convolutional recurrent neural network that uses convolutional layers to extract local features and recurrent layers to capture temporal dependencies [63].
- **CRNN (Res):** a variant of CRNN that uses residual connections to improve the information flow and gradient propagation [64].
- **MLSL:** a prediction method that uses shapelet learning from multiple variables [65].
- **StemGNN:** a spectral temporal graph neural network that can capture both inter-series correlations and temporal dependencies in the spectral domain. It combines graph Fourier transform and discrete Fourier transform to model the complex patterns and dynamics in multivariate time series data [66].

5.3. Evaluation metrics

In this subsection, we present the evaluation metrics that we use to assess the accuracy and robustness of our proposed model and compare it with several state-of-the-art models for district heat load forecasting. We use three metrics to evaluate the performance of different models: root mean square error (RMSE), mean absolute error (MAE), and coefficient of variation of root mean square error (CVRMSE). RMSE measures the average magnitude of the errors between the predicted and actual values, MAE measures the average absolute errors, and CVRMSE measures the normalized RMSE by dividing it by the mean of the actual values. Lower values of these metrics indicate higher accuracy of the model. These metrics are commonly used for district heat load forecasting because they can capture the overall error distribution and account for different scales of heat demand.

- **Root Mean Square Error (RMSE):**

$$
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2},
$$

- **Mean Absolute Error (MAE):**

$$
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i - Y_i|,
$$
Table 3
The multi-horizon one-step performance comparison in three metrics and four models on district heat load observations. The best result in terms of each metric is highlighted in gray, and the second best is underlined.

<table>
<thead>
<tr>
<th>Horizon (h)</th>
<th>Metrics</th>
<th>AR</th>
<th>Dlinear</th>
<th>MTNet</th>
<th>TPA</th>
<th>LSTNet</th>
<th>LSTM ED (GRU)</th>
<th>CRNN (Res)</th>
<th>CRNN (85%)</th>
<th>MSL</th>
<th>StemGGNN</th>
<th>Ac-GRN (80%)</th>
<th>Ac-GRN (85%)</th>
<th>Ac-GRN (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CVR MSE</td>
<td>0.361</td>
<td>0.362</td>
<td>0.333</td>
<td>0.331</td>
<td>0.343</td>
<td>0.702</td>
<td>0.671</td>
<td>0.529</td>
<td>0.409</td>
<td>0.472</td>
<td>0.340</td>
<td>0.372</td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td>CVR MSE</td>
<td>0.439</td>
<td>0.435</td>
<td>0.391</td>
<td>0.357</td>
<td>0.389</td>
<td>0.722</td>
<td>0.699</td>
<td>0.550</td>
<td>0.479</td>
<td>0.511</td>
<td>0.390</td>
<td>0.371</td>
<td>0.348</td>
</tr>
<tr>
<td></td>
<td>CVR MSE</td>
<td>0.511</td>
<td>0.511</td>
<td>0.416</td>
<td>0.403</td>
<td>0.459</td>
<td>0.707</td>
<td>0.713</td>
<td>0.564</td>
<td>0.555</td>
<td>0.538</td>
<td>0.491</td>
<td>0.377</td>
<td>0.365</td>
</tr>
<tr>
<td></td>
<td>CVR MSE</td>
<td>0.616</td>
<td>0.614</td>
<td>0.514</td>
<td>0.446</td>
<td>0.512</td>
<td>0.728</td>
<td>0.747</td>
<td>0.583</td>
<td>0.605</td>
<td>0.597</td>
<td>0.537</td>
<td>0.425</td>
<td>0.382</td>
</tr>
</tbody>
</table>

- Coefficient of Variation of RMSE (CVRMSE):

\[
CVRMSE = \frac{\text{RMSE}}{\bar{Y}}.
\]

where \(\bar{Y}\) is the predicted value, \(Y\) is the actual value, and \(\bar{Y}\) is the mean of the actual values. \(n\) is the number of observations.

5.4. Comparison on multi-horizon prediction

We evaluate the performance of our proposed model Ac-GRN and compare it with 11 state-of-the-art baseline models for multi-horizon one-step district heat load forecasting using three metrics: RMSE, MAE, and CVR MSE. All comparable methods are trained with the whole samples (i.e., 100% sampling ratio). Table 3 shows the results of the comparison for four horizons (15, 30, 45, and 60 days). The results demonstrate that our model Ac-GRN with 90% sampling ratio consistently outperforms other models across all metrics and horizons, especially when the prediction horizon is longer. This indicates that our model can effectively capture the complex spatial-temporal dependencies in the data and actively learn from the most informative samples. Moreover, our model shows a remarkable improvement over other models when the prediction horizon is longer, which reflects its robustness and generalization ability. Among the baseline models, TPA and LSTNet perform better than the others, but they still lag behind the Ac-GRN models. This implies that using graph neural networks and active deep learning can improve the performance of district heat load forecasting.

We also evaluate the impact of different sampling ratios on the performance of our model. The sampling ratio is the percentage of samples that are selected by the active learning module. We use three sampling ratios: 80%, 85%, and 90%. The results show that our model can achieve high accuracy with a relatively low sampling ratio (80%), which demonstrates the efficiency and effectiveness of our active learning strategy. However, increasing the sampling ratio to 90% can further improve the accuracy of our model, especially for longer prediction horizons. This suggests that there is a trade-off between the accuracy and efficiency of our model, and that choosing an appropriate sampling ratio can balance the costs associated with data collection and the model performance. We also compare the results of different sampling ratios with each other and with the baseline models. We find that our model with an 80% sampling ratio can outperform most of the baseline models except for TPA and LSTNet, while our model with 90% sampling ratio can outperform all of them. This shows that our model can achieve competitive or superior results with fewer heat meter samples than other models.

Table 4 presents the prediction performance comparison considering the meteorological factors. The DLinear, MTNet, TPA, MSL, and StemGGNN models are excluded from the comparison as they do not take exogenous factors into account in their prior studies. Across all prediction horizons and metrics, Ac-GRN consistently achieves the best performance, with the lowest RMSE, MAE, and CVR MSE at 90% sample ratio. These results underscore the strong predictive performance of Ac-GRN in handling district heat load observations combined with meteorological factors, reflecting its ability to effectively incorporate and utilize exogenous data. For AR, LSTM, ED (GRU), and CRNN (Res), the introduction of weather data generally improved the accuracy, substantiating the importance of external data integration in optimizing district heat load predictions. The meteorological factors are likely to capture more complexity and variability of the district heat load, which are further leveraged by these models to yield more accurate predictions.

It is worth noting that the CRNN model's performance enhancement appears to be limited to a horizon of 15. This limitation could potentially be attributed to the nature of convolutional layers, which primarily extract local features in the data. While beneficial for certain tasks, they may fail to capture long-term dependencies or complex patterns that emerge over larger time horizons. As a result, the CRNN model demonstrates increased performance at a shorter horizon (h = 15) and performs relatively less effectively as the horizon increases. Moreover, the results suggest that the addition of residual connections in CRNN (Res) contributes to its superior stability and robustness compared to the standard CRNN. These connections help to alleviate the vanishing gradient problem and allow for more complex feature extraction, ultimately enhancing the model's performance across all horizons.

5.5. Comparison on multi-horizon multi-step prediction

Table 5 presents the results of our proposed model Ac-GRN and four state-of-the-art methods for multi-horizon multi-step district heat load forecasting. We selected these four methods based on their relatively strong performance in the multi-horizon one-step prediction task. We use three metrics: RMSE, MAE, and CVR MSE, to evaluate the accuracy and robustness of the models. We also analyze the impact of different sampling ratios on our model's performance. The results show that our model Ac-GRN consistently outperforms other models across all metrics and horizons, especially when the prediction horizon and step are longer. This indicates that our model can effectively capture the complex spatial-temporal dependencies in the data and actively learn from the most informative samples. Moreover, our model shows a remarkable improvement over other models when the prediction horizon and step are longer, which reflects its robustness and generalization ability. Our model also achieves high accuracy with a relatively low...
Table 4
The multi-horizon one-step performance comparison in three metrics and four horizons on district heat load observations with meteorological factors. The best result in terms of each metric is highlighted in gray.

<table>
<thead>
<tr>
<th>Horizon (ℎ)</th>
<th>Metrics</th>
<th>AR</th>
<th>LSTM</th>
<th>ED (GRU)</th>
<th>CRNN</th>
<th>CRNN (Res)</th>
<th>Ac-GRN (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>11.461</td>
<td>12.266</td>
<td>11.255</td>
<td>17.493</td>
<td>13.096</td>
<td>10.303</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>8.393</td>
<td>9.102</td>
<td>8.050</td>
<td>11.725</td>
<td>9.344</td>
<td>7.180</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.345</td>
<td>0.369</td>
<td>0.399</td>
<td>0.527</td>
<td>0.394</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>9.033</td>
<td>10.958</td>
<td>11.170</td>
<td>12.798</td>
<td>11.312</td>
<td>7.630</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.368</td>
<td>0.437</td>
<td>0.446</td>
<td>0.552</td>
<td>0.466</td>
<td>0.325</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.385</td>
<td>0.442</td>
<td>0.491</td>
<td>0.575</td>
<td>0.534</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.394</td>
<td>0.516</td>
<td>0.562</td>
<td>0.576</td>
<td>0.597</td>
<td>0.347</td>
</tr>
</tbody>
</table>

Table 5
The multi-horizon multi-step performance comparison in three metrics and four horizons on district heat load observations.

<table>
<thead>
<tr>
<th>Step</th>
<th>Horizon (ℎ)</th>
<th>Metrics</th>
<th>Dlinear</th>
<th>LSTNet</th>
<th>TPA</th>
<th>MTNet</th>
<th>Ac-GRN (80%)</th>
<th>Ac-GRN (85%)</th>
<th>Ac-GRN (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>9.064</td>
<td>8.586</td>
<td>8.197</td>
<td>8.207</td>
<td>7.852</td>
<td>7.872</td>
<td>7.791</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.388</td>
<td>0.371</td>
<td>0.350</td>
<td>0.365</td>
<td>0.340</td>
<td>0.339</td>
<td>0.335</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>11.292</td>
<td>9.813</td>
<td>9.067</td>
<td>10.382</td>
<td>8.004</td>
<td>8.089</td>
<td>8.005</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.473</td>
<td>0.416</td>
<td>0.382</td>
<td>0.453</td>
<td>0.343</td>
<td>0.347</td>
<td>0.345</td>
<td>0.345</td>
</tr>
<tr>
<td>45</td>
<td>RMSE</td>
<td>21.974</td>
<td>17.336</td>
<td>17.010</td>
<td>16.678</td>
<td>15.789</td>
<td>15.842</td>
<td>15.683</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.680</td>
<td>0.526</td>
<td>0.526</td>
<td>0.516</td>
<td>0.488</td>
<td>0.490</td>
<td>0.485</td>
<td>0.485</td>
</tr>
</tbody>
</table>

60    | RMSE        | 23.498 | 17.499 | 16.303 | 18.209 | 16.048 | 16.000 | 15.868 | 0.501 |
|      | MAE         | 17.269 | 12.665 | 12.024 | 13.808 | 12.272 | 12.243 | 12.100 | 0.501 |
|      | CVRMSE      | 0.741 | 0.552 | 0.514 | 0.574 | 0.506 | 0.505 | 0.501 | 0.501 |

Table 6
Ablation analysis. The best result is highlighted in gray, while a wavy line indicates the worst result.

<table>
<thead>
<tr>
<th>Horizon (ℎ)</th>
<th>Metrics</th>
<th>Ac-GRN</th>
<th>Ac-GRN w/o TCM</th>
<th>Ac-GRN w/o GRU</th>
<th>Ac-GRN w/o Attention</th>
<th>Ac-GRN w/o LEU</th>
<th>Ac-GRN w/o Generate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>7.385</td>
<td>8.484</td>
<td>7.813</td>
<td>8.185</td>
<td>9.403</td>
<td>9.871</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.327</td>
<td>0.363</td>
<td>0.351</td>
<td>0.351</td>
<td>0.386</td>
<td>0.386</td>
</tr>
<tr>
<td>30</td>
<td>RMSE</td>
<td>11.074</td>
<td>12.924</td>
<td>11.958</td>
<td>11.855</td>
<td>11.829</td>
<td>11.858</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.333</td>
<td>0.389</td>
<td>0.360</td>
<td>0.357</td>
<td>0.394</td>
<td>0.394</td>
</tr>
<tr>
<td>45</td>
<td>RMSE</td>
<td>11.864</td>
<td>13.466</td>
<td>12.159</td>
<td>12.448</td>
<td>12.611</td>
<td>12.453</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.357</td>
<td>0.405</td>
<td>0.366</td>
<td>0.375</td>
<td>0.513</td>
<td>0.393</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.377</td>
<td>0.449</td>
<td>0.380</td>
<td>0.470</td>
<td>0.509</td>
<td>0.452</td>
</tr>
</tbody>
</table>

5.6. Ablation and sensitivity analysis

To evaluate the contribution of each component of our proposed Ac-GRN model, we conduct an ablation study by removing one component at a time and comparing the performance with the full model. The components considered for this analysis include the spatial–temporal sampling ratio (80%), which demonstrates the efficiency and effectiveness of our active learning strategy. This table complements the results in Table 3, which only considers one-step prediction for different horizons. It shows that our model can also handle multi-step prediction for different horizons, which is more challenging and realistic for district heat load forecasting.
convolution (TCM) module, the GRU and Attention in the spatio-temporal memory enhancement module, the linear representation unit (LEU) in the multi-component fusion, and the generation component in the feature-enhanced module. Table 6 shows the results of the ablation analysis in terms of RMSE, MAE, and CV-RMSE metrics for four horizons (15, 30, 45, and 60 days). The results show that removing any component leads to a decrease in accuracy and an increase in error metrics for all horizons. This indicates that each component is essential for the effectiveness of our model and that they work well together. The most significant drop in performance occurs when we remove the Generate or the LEU components, depending on the horizon. The Generation component in the feature-enhanced module plays a crucial role in complementing the incomplete data, which provide by active learning. This mitigates potential accuracy loss caused by missing data from unselected heat meters. The LEU in the multi-component fusion aids in enhancing the trend features of model inputs via linear embedding, which helps improve the model’s ability to capture and predict heat load trends accurately. For short horizons (15 and 30 days), removing the Generate component causes the largest performance degradation. This observation implies that generating informative samples is vital to accurately capture the short-term dynamics of heat load. For long horizons (45 and 60 days), removing the LEU component causes the largest performance degradation, which suggests that enhancing the linear features is crucial for capturing the long-term trends of heat load. To examine how the performance of our model changes with different values of a parameter or a variable, we conduct a sensitivity analysis by varying one parameter or variable at a time and measuring different values of a parameter or a variable, we conduct a sensitivity analysis by varying one parameter or variable at a time and measuring how the outcomes. The results demonstrate how the CV-RMSE metric varies with different window sizes $T$ for four different forecasting horizons (15, 30, 45, and 60 days). These results reveal that an optimal value of $T$ exists for each horizon that minimizes the CV-RMSE metric and optimizes the accuracy of our model. The optimal values of $T$ are 88, 80, 55, and 55 for horizons 15, 30, 45, and 60 days, respectively. This suggests that our model can capture the temporal dependencies in the data better when using an appropriate window size. If the window size is too small, our model may not have enough information to make accurate predictions. If the window size is too large, our model may suffer from overfitting or noise. All comparisons are made using the optimal window size settings for each method.

The ablation analyses demonstrate the robustness and generalization capabilities of our proposed Ac-GRN model. Meanwhile, the sensitivity analyses offer valuable insights into the significance and influence of various parameters within our model for district heat load forecasting. These analyses collectively affirm the potential of the Ac-GRN model as a reliable tool for district heat load prediction.

5.7. Computational efficiency and sampling ratio analysis

In this subsection, we compare the computational efficiency and sampling ratio analysis of different methods for district heat load prediction. The computational efficiency is measured by the running time per batch for prediction processing, while the sampling ratio indicates the proportion of annotated samples used for training. The results are shown in the two tables, which report the running time and the prediction accuracy for different methods under different horizons and sampling ratios, respectively. The relatively worse result in terms of each metric is highlighted in red.

We compared the computational efficiency of the proposed Ac-GRN with other comparable methods for district heat load prediction without meteorological factors, as shown in Table 7. The prediction horizon $h$ and the prediction step $s$ were fixed at 15 and 1, respectively. The performances are evaluated based on two statistical measures: the mean and the standard deviation, computed over 20 repetitions of the experiment for each model. The mean represents the average processing time per batch, while the standard deviation measures the degree of variation in the processing times. As can be seen from Table 7, the AR model has the fastest running time, with a mean value of 0.00103 s per batch. This is likely attributable to the linear structure of the AR model, which, while beneficial for computational efficiency, could limit its performance in long-horizon predictions. The StemGNN model has the slowest running time, with a mean value of 1.74973 s per batch, which is considerably higher than any of the other models. The performance of StemGNN also has a large variation, as indicated by its high standard deviation of 0.00689. The proposed model (Ac-GRN) has a moderate running time, with a mean value of 0.01058 s per batch. Although not the fastest, our proposed method still demonstrates reasonable efficiency, especially when compared to the StemGNN model. Both methods are based on graph neural network structures, with our method being over 160 times faster than StemGNN. This suggests...
that Ac-GRN is capable of balancing the computational efficiency and prediction performance more effectively than some of the other models. The standard deviations for each method are relatively small compared to their corresponding means, indicating a relatively tight distribution of running times. The largest standard deviation is for StemGNN, which could suggest variability in its performance.

Table 8 shows the multi-horizon one-step performance comparison of different methods for district heat load prediction with different sampling ratios. The sampling ratios indicate the proportion of annotated samples used for training, ranging from 80% to 100%. The prediction accuracy is measured by three evaluation metrics: MAE, RMSE, and CVRMSE. The relatively worse result in terms of each metric is highlighted in red.

<table>
<thead>
<tr>
<th>Horizon (h)</th>
<th>Metrics</th>
<th>Dlinear (90%)</th>
<th>Dlinear (100%)</th>
<th>MTNet (80%)</th>
<th>MTNet (90%)</th>
<th>MTNet (100%)</th>
<th>LSTNet (80%)</th>
<th>LSTNet (90%)</th>
<th>LSTNet (100%)</th>
<th>TPA (80%)</th>
<th>TPA (90%)</th>
<th>TPA (100%)</th>
<th>Ac-GRN (80%)</th>
<th>Ac-GRN (90%)</th>
<th>Ac-GRN (100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.547</td>
<td>0.501</td>
<td>0.362</td>
<td>0.372</td>
<td>0.371</td>
<td>0.333</td>
<td>0.459</td>
<td>0.393</td>
<td>0.331</td>
<td>0.412</td>
<td>0.377</td>
<td>0.343</td>
<td>0.372</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.612</td>
<td>0.547</td>
<td>0.435</td>
<td>0.435</td>
<td>0.418</td>
<td>0.391</td>
<td>0.445</td>
<td>0.417</td>
<td>0.357</td>
<td>0.474</td>
<td>0.438</td>
<td>0.389</td>
<td>0.371</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.640</td>
<td>0.578</td>
<td>0.511</td>
<td>0.471</td>
<td>0.472</td>
<td>0.416</td>
<td>0.503</td>
<td>0.467</td>
<td>0.403</td>
<td>0.501</td>
<td>0.474</td>
<td>0.459</td>
<td>0.377</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>CVRMSE</td>
<td>0.707</td>
<td>0.673</td>
<td>0.614</td>
<td>0.537</td>
<td>0.541</td>
<td>0.514</td>
<td>0.535</td>
<td>0.493</td>
<td>0.446</td>
<td>0.636</td>
<td>0.577</td>
<td>0.512</td>
<td>0.425</td>
<td>0.377</td>
</tr>
</tbody>
</table>

6. Discussions

In this paper, we proposed Ac-GRN, a novel model that combines active deep learning and graph neural networks for district heat load forecasting. Our model consists of three main components: a knowledge extractor that uses active learning to select the most informative heat meters, a feature reconstructor that imputes the missing data of unselected meters using correlation analysis, and a graph recurrent network that models the spatial–temporal dependencies among heat meters. Our model can capture the complex patterns and relationships in the data and provide accurate and robust forecasts for different horizons and steps.

One of the main contributions of our paper is that we use active learning to reduce the frequency and cost of data collection and human effort. By selecting only a small subset of heat meters that are most representative and informative, we can train our model with fewer data and achieve better performance than using all the data. We also use correlation analysis to reconstruct the missing data of unselected meters, which improves the data quality and completeness. We show that our model can achieve high accuracy with only 80% of the data, which demonstrates the efficiency and effectiveness of our active learning strategy.

Another contribution of our paper is that we use graph neural networks to model the spatial–temporal dependencies among heat meters. We represent the heat meters as nodes in a graph, where the edges are defined by the heat load similarity. We use a graph convolutional network to learn the spatial features of the nodes, and a gated recurrent unit to learn the temporal features of the sequences. We also use an attention mechanism to learn how to weigh different nodes and time steps for prediction. We show that our model can capture both linear and nonlinear relationships in the data, and outperform other methods that use different types of neural networks.

A third contribution of our paper is that we provide explanations for our model and its predictions. Our proposed Ac-GRN model leverages active deep learning methods and correlation-based techniques to offer explainability. A novel framework was designed to select representative smart heat meters by active deep learning. This framework identifies both representative and non-representative meter groups using the least confidence measure. The annotated results of these meter groups can provide actionable feedback to corporations or governmental bodies, thereby aiding in heating regulation and management. Further, we introduce a feature reconstructor to interpolate incomplete data based on correlations among heat meters. This reconstruction process is both...
transparent and interpretable, effectively linking unselected heat meters with their correlated, annotated counterparts. In addition, we use attention mechanisms to explain how our model attends to different nodes in the graph and different time steps in the time series. These attention weights can be visualized, demonstrating how they adjust depending on the input data and the horizon. To provide a concrete example of our model’s explainability, consider a scenario where our model is predicting heat load for a specific time period. The active deep learning component identifies the heat meters that are most influential for the prediction. The feature reconstructor then fills in any missing data in the heat meter readings, using correlations among the meters. This reconstructed data is transparent and interpretable, offering a clear view of how the model arrived at its prediction.

Despite the promising results of our model, we acknowledge that there are some limitations and challenges that need to be addressed in future research. First, our model relies on a fixed graph structure to represent the spatial dependencies among heat meters, which may not capture the dynamic changes in the network topology or node attributes over time. For example, there may be new nodes added or removed, or node features updated over time. Second, our model uses a single prediction horizon for all heat meters, which may not be optimal for each meter. For example, some meters may have more stable or predictable patterns than others, and may require different prediction horizons. Third, our Ac-GRN model requires a large amount of computational resources and time to train and update the model parameters, especially when dealing with large-scale and high-dimensional district heating data. This may limit the applicability and scalability of our model in real-time and resource-constrained settings, where timely and efficient forecasting is essential for optimizing energy production and distribution. Fourth, our model does not provide interpretable explanations for its predictions, which may limit its applicability and trustworthiness in real-world scenarios. For example, it may be useful to know why some meters have higher or lower heat load than others, or what factors contribute most to the prediction.

To address these limitations, we propose some possible directions for future research. First, we can use dynamic graph neural networks to model the spatial dependencies among heat meters, such as temporal graph convolutional networks [42], recurrent graph neural networks [47], or graph attention networks [46]. These networks can handle the changes in the graph structure or node attributes over time and learn more expressive representations of the data. Dynamic graph neural networks have been shown to improve the accuracy and robustness of district heat load forecasting in previous studies [67,68]. Second, we can use multi-task learning or multi-output learning to predict different horizons for different heat meters simultaneously. These learning methods can leverage the shared information and correlations among different horizons and meters and improve the accuracy and efficiency of our model. Multi-task learning or multi-output learning can also reduce the complexity and redundancy of the model by sharing parameters and features across different tasks or outputs [69]. Third, we can use explainable artificial intelligence techniques to provide interpretable explanations for our predictions, such as SHAP values [70], LIME [71], or counterfactuals [72]. These techniques can help us understand how our model works and what factors influence its predictions. Explainable artificial intelligence techniques can also enhance the trust and reliability of our model and support decision-making for optimizing energy production and distribution [73]. Last, to further validate and refine our model’s explanations, we plan to conduct user studies or expert evaluations. These studies will involve presenting the model’s explanations to users or experts and collecting their feedback on the clarity, utility, and relevance of the explanations. This feedback will provide valuable insights into how well users or experts understand the explanations and how they might use the explanations in real-world decision-making scenarios. We believe that this additional validation will not only strengthen our model’s ability to provide “meaningful explanations” but also provide directions for further improving the model’s explainability.

7. Conclusions and future work

District heat load forecasting is an important task for improving the efficiency and sustainability of district heating systems. However, existing methods for district heat load forecasting face several challenges, such as capturing the complex spatial–temporal dependencies among heat meters, reducing the frequency and cost of data collection and human effort, and providing interpretable explanations for the predictions. In this paper, we addressed these challenges by proposing Ac-GRN, a novel model that combines active deep learning and graph neural networks for district heat load forecasting. Our model can effectively select the most informative and representative samples from a large pool of data, and use them to train a graph recurrent neural network with bidirectional recurrent connections. Our model can capture both linear and nonlinear relationships in the data and provide accurate and robust forecasts for different horizons and steps. We evaluated our model on a real-world dataset of district heating consumption data from Danish residential buildings and compared it with 11 state-of-the-art methods. We conducted extensive experiments to test the performance of our model for different prediction horizons, steps, and sampling ratios. The results indicate that our model achieved superior performance in terms of accuracy, robustness, and reliability for multi-horizon multi-step district heat load forecasting. We also compared the computational efficiency of our model with other methods and found that our model has a reasonable running time per batch for prediction processing, especially when compared to other graph neural network-based methods. This shows that our model can balance the trade-off between computational efficiency and prediction performance more effectively than some of the other methods.

For future work, we plan to extend our model to handle dynamic graph structures and node attributes, to predict different horizons for different heat meters, and to provide interpretable explanations for our predictions. We also plan to apply our model to other domains that involve spatial–temporal data, such as traffic flow forecasting, air quality forecasting, and social network analysis. We hope that our work can inspire more research on active deep learning and graph neural networks for district heat load forecasting and other related tasks.

CRediT authorship contribution statement

Yaozhui Huang: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. Yuan Zhao: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. Zhijin Wang: Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. Xiufeng Liu: Writing – review & editing, Writing – original draft, Supervision, Data curation, Conceptualization. Hanjing Liu: Writing – review & editing, Writing – original draft, Methodology. Yonggang Fu: Writing – review & editing, Project administration, Conceptualization.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.
Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used the language polishing service provided by ChatGPT in order to achieve the readability and clarity. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgments

This research was supported in part by the Fujian Province Natural Science Foundation of Fujian Province (CN) (nos. 2021J01857, 2021J01859, and 2022J01335), and the EMB3Rs project (no. 84712) funded by the European Union Horizon 2020 research and innovation programme. We thank the editors and the anonymous reviewers for their valuable comments and suggestions, which greatly benefited this paper.

Appendix

A.1. Hyper-parameter setting

The parameters setting of the proposed method and benchmarks are listed in Table 9.

Table 9 Hyper-parameter settings

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Option range</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM ED (GRU)</td>
<td>Hidden size</td>
<td>($2^2, 3^2, 4^2$)</td>
</tr>
<tr>
<td>CRNN</td>
<td>Kernel size</td>
<td>3–9 (2 per step)</td>
</tr>
<tr>
<td></td>
<td>CNN out channels</td>
<td>($2^2, 3^2, 4^2, 5^2$)</td>
</tr>
<tr>
<td>CRNN (Res)</td>
<td>Hidden size</td>
<td>($2^2, 3^2, 4^2$)</td>
</tr>
<tr>
<td>CRNN (Res)</td>
<td>Residual window size</td>
<td>1–5 (1 per step)</td>
</tr>
<tr>
<td>TPA-LSTM</td>
<td>Kernel size</td>
<td>3–9 (2 per step)</td>
</tr>
<tr>
<td></td>
<td>CNN out channels</td>
<td>($2^2, 3^2, 4^2, 5^2$)</td>
</tr>
<tr>
<td>LSTMNet</td>
<td>GRU hidden size</td>
<td>($2^2, 3^2, 4^2$)</td>
</tr>
<tr>
<td></td>
<td>The number of GRU layers</td>
<td>1–3 (1 per step)</td>
</tr>
<tr>
<td></td>
<td>Highway window size</td>
<td>1–10 (1 per step)</td>
</tr>
<tr>
<td>LSTNet</td>
<td>Skip window size</td>
<td>1–3 (1 per step)</td>
</tr>
<tr>
<td></td>
<td>Skip GRU hidden size</td>
<td>($2^2, 3^2, 4^2$)</td>
</tr>
<tr>
<td>MSL</td>
<td>Block size</td>
<td>1–10 (1 per step)</td>
</tr>
<tr>
<td>StemGNN</td>
<td>Encoder layers</td>
<td>1–3 (1 per step)</td>
</tr>
<tr>
<td></td>
<td>The numbers of blocks</td>
<td>1–3 (1 per step)</td>
</tr>
<tr>
<td>Ac-GRN</td>
<td>GCN hidden size</td>
<td>($2^2, 3^2, 4^2, 5^2$)</td>
</tr>
<tr>
<td></td>
<td>GCN out channel</td>
<td>($2^3, 3^2, 4^2$)</td>
</tr>
<tr>
<td></td>
<td>GRU hidden size</td>
<td>($2^2, 3^2, 4^2$)</td>
</tr>
<tr>
<td></td>
<td>Kernel size</td>
<td>3–9 (2 per step)</td>
</tr>
<tr>
<td></td>
<td>The number of correlated meters</td>
<td>(3, 5, 10, 15, 20)</td>
</tr>
<tr>
<td></td>
<td>for reconstruction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNN out channels</td>
<td>($2^2, 3^2, 4^2, 5^2$)</td>
</tr>
</tbody>
</table>

References

[34] Li J, Gu C, Wei X, Gili IH, Xiang Y. An IoT-based thermal modelling of dwelling rooms to enable flexible energy management. IEEE Trans Smart Grid 2023;Early Access.


