Electric Power Systems Research 184 (2020) 106304

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Assessment of aggregation strategies for machine-learning based short-term load forecasting

Cong Feng, Jie Zhang*

Department of Mechanical Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA

ARTICLE INFO

Keywords:
Load forecasting
Aggregate forecasting
Hierarchical forecasting
Ensemble forecasting
Machine learning

ABSTRACT

Effective short-term load forecasting (STLF) plays an important role in power system operations. It is challenging to identify an ML model that has the outperformance in all scenarios. Therefore, there are a number of aggregation strategies developed to improve STLF. However, the superiority of these aggregation strategies has not been assessed. In this paper, STLF with three aggregation strategies are developed, which are information aggregation (IA), model aggregation (MA), and hierarchy aggregation (HA). The IA, MA, and HA strategies aggregate inputs, models, and forecasts at the pre-forecasting, model-building, or post-forecasting stage, respectively. To verify the effectiveness of the three aggregation strategies, a set of 10 models based on 4 machine-learning algorithms are developed in each aggregation category to predict 1-hour-ahead load. Case studies show that: (i) STLF-IA presents superior performance than STLF with weather data and STLF with individual load data consistently, and the performance can be further enhanced by the recursive feature elemination (RFE) feature selection method; (ii) MA improves the STLF robustness by reducing the risk of unsatisfactory single-algorithm STLF models; and (iii) STLF-HA produces the most accurate forecasts with a 0.83% normalized mean absolute error and a 1.35% mean absolute percentage error, while keeping hierarchical aggregate consistency.

1. Introduction

Accurate short-term load forecasting (STLF) plays an important role in power system operations. The most widely used STLF methods are machine learning (ML) methods, including artificial neural networks (ANNs), support vector regression models (SVR), and decision tree-based models. It is a consensus that simply applying an ML model to a forecasting task is not proper due to that: (i) the performance of an ML model is significantly affected by the data, (ii) there is no single model that can always be the best among all models, (iii) a single model can not meet the forecasting requirements, such as the aggregate consistency.[1] In order to achieve better forecasting performance, a number of methodologies[2] have been developed in the literature, which could be generally divided into three categories: the information aggregation (IA), the model aggregation (MA), and the hierarchy aggregation (HA).

Research in the first category focuses on integrating more informative and better-organized data to enhance the forecasting accuracy, which is defined as *information aggregation (IA)* in this paper. For example, endogenous load feature is the most important input to statistical models [2]. Additionally, meteorological variables, such as temperature [3] and humidity [4], have been commonly adopted to generate load forecasts. However, Refs. [3,4] only considered temperature or humidity as independent variables, where their interactions with other parameters were neglected. Moreover, residents' life patterns were used to improve the customer-level STLF in Ref. [5] by considering the most relevant six operated appliances. Nevertheless, the selection was conducted empirically, which might be suboptimal. Jiang *et al.* [6] developed an accurate and robust STLF model based on the date information. In addition, feature selection techniques, such as the mutual information-based filter method that selects features based on dependencies of the forecasting target variable [7], were also used to optimize the inputs to forecasting models. Nonetheless, filter methods depend on statistical relationships, which do not involve the forecasting process. Therefore, there is a need to use wrapper methods, which evaluate feature importance based on their impacts on the forecasting accuracy, for automatic feature selection.

The second category contains methodologies combining forecasts from multiple individual models, which is defined as *model aggregation*

* Corresponding author.
E-mail address: jiezhang@utdallas.edu (J. Zhang).

[1] Aggregate consistency is defined as the equality between the sum of forecasts and the forecast of the sum.

[2] Four terms are repeatedly used in this paper, which are *methodology, algorithm, method*, and *model*. A methodology refers to a general solution framework that can be implemented with different models [1]. Several models can be built based on one method or algorithm.

https://doi.org/10.1016/j.epsr.2020.106304
Received 21 November 2019; Received in revised form 12 February 2020; Accepted 4 March 2020
Available online 21 March 2020
0378-7796/ © 2020 Elsevier B.V. All rights reserved.

Nomenclature

Acronyms

ANN, SVR Artificial neural network, support vector regression
B*i*, UTD The *i*th building, the summation of the 13 buildings in the University of Texas at Dallas
GBM, RF Gradient boosting machine, random forest
GLS, OLS, MinT Generalized least square, ordinary least square, minimum trace
IA, MA, HA Information aggregation, model aggregation, hierarchy aggregation
LF, STLF Load forecasting, short-term load forecasting
Mi_a, Mi_{ab} The *i*th model in group *a*, the comparison of *i*th model in group a and b
RFE Recursive feature elimination
STLF-B Group of STLF with the bottom-up strategy in the hierarchy aggregation category
STLF-F Group of STLF with features selected from both weather information and individual building load in the information aggregation category
STLF-HA STLF with hierarchy aggregation
STLF-I Group of STLF with both weather data and individual building load in the information aggregation category
STLF-IA STLF with information aggregation
STLF-L Group of STLF with individual building load in the information aggregation category
STLF-M Group of STLF with model aggregation in the model aggregation category
STLF-MA STLF with model aggregation
STLF-O Group of STLF with the ordinary least square reconciliation in the hierarchy aggregation category
STLF-T Group of STLF with the minimum trace reconciliation in the hierarchy aggregation category
STLF-W Group of STLF with weather information in the information aggregation category

Variables, vectors, and matrices

$\boldsymbol{\beta}, \boldsymbol{\beta}^{OLS}$ Unknown mean vector of the bottom-level entries and the ordinary least square unbiased estimates of $\boldsymbol{\beta}$
$\boldsymbol{\beta}^{GLS}, \boldsymbol{\beta}^{MinT}$ Minimum variance and minimum trace unbiased estimates of $\boldsymbol{\beta}$
$\hat{\boldsymbol{y}}_{IA}, \hat{\boldsymbol{y}}_{MA}, \hat{\boldsymbol{Y}}_H$ Vectors of forecasts in information aggregation and model aggregation, forecasts of all entries in hierarchy aggregation categories
$\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\dagger$ Unknown variance of bottom-level entries and the Moore-Penrose generalized inverse of $\boldsymbol{\Sigma}$
$\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{Y}}$ Forecast vector provided by the first-layer model f_i, combination of the first-layer forecast vectors
$\boldsymbol{b}_{i/j,H}, \boldsymbol{B}$ Base forecast vector of an entry at the bottom-level and base forecasts of entries at all levels
$\boldsymbol{D}, \boldsymbol{F}, \boldsymbol{X}, \boldsymbol{X}^D$ Decision matrix, decision matrix constructed by feature selection, matrix with all variables, and input matrix selected by the decision matrix
$\boldsymbol{W}, k_h$ Sample covariance matrix and a positive scaling factor
$\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{X}^l, \boldsymbol{x}^s$ Matrices of weather data, calendar data, individual building load data, and target variable vector
$\hat{\boldsymbol{S}}, \boldsymbol{\varepsilon}$ Summing matrix and error vector with zero mean and unknown variance
$\hat{\boldsymbol{Y}}_{i/j,H}, \hat{\boldsymbol{Y}}_{i,H}, \hat{\boldsymbol{Y}}_H$ Forecast vectors of entries in the bottom-level, mid-level, and top level

Functions and metrics

$\hat{y}, y, y_{\max}$ Forecast value, actual value, and maximum actual value
f_i, Φ Forecasting model in the first-layer and blending model in the second-layer.
nMAE, MAPE Forecasting normalized mean absolute error and mean absolute percentage error
Imp_{ab}^A, Imp_{ab}^P Forecasting *nMAE* and *MAPE* improvements of model a over model b
BE, nBE Forecasting bias error and normalized bias error

(MA). For example, Zhang et al. [8] ensembled a set of extreme learning machines and took the median value of their outputs as forecasts, which showed both superior training efficiency and forecasting accuracy over benchmark models. An STLF model that integrated individual models using weight-coefficient optimization was developed in Ref. [9], which showed better performance than six benchmark single-algorithm models. Hassan et al. [10] ensembled 100 ANN models with simple average, trimmed mean, and Bayesian model averaging, and found that the Bayesian model averaging approach performed better than other ensemble models. A collection of ANN models were built based on a two-stage diversity controlled resampling procedure and then ensembled by a linear combiner in Ref. [11]. The ensemble model was found to improve the reliability of individual household energy consumption forecasts. Bagged-boosted ANNs were sequentially trained and to reduce the bias and averaged to reduce the variance [12]. This advanced MA strategy outperformed averaging, bagging, and boosting strategies. However, the above MA methods assign fixed weights to model members, which fail to consider the dynamic characteristics in the forecasting [13,14].

The third category of STLF is called hierarchical forecasting, which we define as *hierarchy aggregation (HA)*. In this category, individual forecasts are aggregated to improve the top-level individual forecasts in the power system hierarchy while keeping the aggregate consistency. Compared to the other two types of aggregate strategies, research of the HA is limited. For example, Sevlian and Rajagopal [15] investigated the relationship between forecasting accuracy and aggregation size, and found the forecasting accuracy scales with load aggregation size, which follows the Law of Large Numbers, up to a point of diminishing returns. While the most common HA strategies in STLF is bottom-up (BU) summation [16–18], several strategies have been developed in other areas to reconcile the base forecasts (i.e., forecasts without reconciliations) in multiple levels so that the aggregate consistency in the hierarchy can be satisfied. For example, a reconciliation process was performed by solving a linear regression with an ordinary least squares (OLS) estimator, which improved base forecasts for Australian tourism forecasting [19]. A minimum trace (MinT) estimator and its variants were developed in Ref. [20] for the same Australian tourism forecasting and were applied to solar forecasting by Yang et al. [21].

All the three categories of aggregate strategies have been reported to enhance the STLF accuracy. However, the superiority (i.e., which methodology has better accuracy) of the three aggregation forecasting strategies has not been studied in the literature. In an attempt to comprehensively compare the aggregation strategies at different stages in the forecasting process, STLF models with IA, MA, and HA are developed in this paper to aggregate inputs, models, and forecasts, respectively. A set of 10 models based on 4 ML algorithms are built to ensure the generality of this study. The performance of models in different groups is compared to show the pros and cons of the three aggregation strategies. The main contributions and innovations of this paper include:

1) Comparing STLF with different IA strategies, including: STLF with weather information (STLF-W), with individual load (STLF-L), with the integration of weather and individual load (STLF-I), and with

STLF-I combined with feature selection (STLF-F);

2) Assessing STLF with MA by using different blending methods, including simple averaging, linear regression, and ML algorithms;
3) Introducing aggregate consistency into hierarchical STLF and comparing STLF-HA with BU (STLF-B), OLS (STLF-O), and MinT (STLF-T);
4) Comparing STLF with different aggregation strategies, which are STLF-IA (including the STLF-I and STLF-F groups), STLF-MA (including the STLF-M group),[3] and STLF-HA (including the STLF-B, STLF-O, and STLF-T groups);
5) Ensuring the generality of the assessment by using 10 ML models.

The remainder of this paper is organized as follows. STLF models with IA, MA, and HA are developed in Section 2. Section 3 describes the data for case studies, benchmarks, and evaluation metrics. Results of case studies are analyzed and compared in Section 4. Section 5 concludes the paper.

2. Short-term load forecasting methodologies with different aggregation strategies

Three types of aggregation strategies (i.e., IA, MA, and HA) are described and formularized in this section. The three aggregation strategies aggregate distinct objects at different stages (enclosed by dashed boxes in Fig. 1), which are pre-forecasting stage, model-building stage, and post-forecasting stage.

2.1. Information aggregation (IA)

The first generation of STLF only depends on the load time series itself, which is called the time series approach [22]. External information, such as meteorological data and calendar features, is integrated into the second generation of STLF [23]. With the development of advanced metering infrastructure, smart meter data provides an opportunity to further improve STLF accuracy. With increasing data dimension, feature selection methods are also used to optimally determine the input combination to forecasting models.

In this paper, STLF with four sets of inputs is studied and compared, which are: (i) weather data ($\boldsymbol{X}^w$), calendar data ($\boldsymbol{X}^c$), and target variable data (i.e., load at the top level, which is denoted by $\boldsymbol{x}^s$), (ii) individual load data (i.e., load data at the bottom level, which is denoted by $\boldsymbol{X}^l$), calendar data ($\boldsymbol{X}^c$), and $\boldsymbol{x}^s$, (iii) $\boldsymbol{X}^w$, $\boldsymbol{X}^l$, $\boldsymbol{X}^c$, and $\boldsymbol{x}^s$, and (iv) inputs selected from $\boldsymbol{X}^w$, $\boldsymbol{X}^l$, $\boldsymbol{X}^c$, and $\boldsymbol{x}^s$ using recursive feature elimination (RFE). Please note that only the latest lagged features are included in the inputs for all the models. The STLF with IA (STLF-IA) conducts aggregation at the first step in the forecasting process, as illustrated in Fig. 1(a). STLF-IA is formularized as follows:

$$\hat{\boldsymbol{y}}_{i,IA} = f_i(\boldsymbol{X}^D) \tag{1}$$

$$\boldsymbol{X}^D = \boldsymbol{X}\boldsymbol{D} = \left[\boldsymbol{x}^s_{n\times 1}\quad \boldsymbol{X}^w_{n\times d_w}\quad \boldsymbol{X}^c_{n\times d_c}\quad \boldsymbol{X}^l_{n\times d_l}\right]\boldsymbol{D}_{(1+d_w+d_c+d_l)\times k} \tag{2}$$

where n is the data length, d_w, d_c, and d_l are dimensions of weather data, calendar data, and individual load data, respectively, $f_i(*)$ is the ith model, $\hat{\boldsymbol{y}}_{IA}$ is a forecasting vector in the IA category, $\boldsymbol{X}$ is a input matrix with all variables, $\boldsymbol{X}^D$ is a selected input matrix, and $\boldsymbol{D}$ is a decision matrix. k is the decision matrix dimension, which equals to the number of selected inputs. The matrix $\boldsymbol{D}$ has four forms in terms of two benchmark scenarios (STLF-W and STLF-L) and two IA scenarios (STLF-I and STLF-F).

$$\boldsymbol{D}_{(1+d_w+d_l)\times(1+d_w+d_c)} = \left[\boldsymbol{I}_{(1+d_w+d_c)} \mid \vec{\boldsymbol{0}}_{(1+d_w+d_c)\times d_l}\right]^T \tag{3a}$$

$$\boldsymbol{D}_{(1+d_w+d_l)\times(1+d_l+d_c)} = \begin{bmatrix} 1 & \vec{\boldsymbol{0}}_{(1+d_w)\times(d_l+d_c)} \\ \vec{\boldsymbol{0}}_{d_w\times 1} & \\ \vec{\boldsymbol{0}}_{d_l\times 1} & \boldsymbol{I}_{(d_l+d_c)} \end{bmatrix} \tag{3b}$$

$$\boldsymbol{D}_{(1+d_w+d_l+d_c)\times(1+d_w+d_l+d_c)} = \boldsymbol{I}_{(1+d_w+d_l+d_c)\times(1+d_w+d_l+d_c)} \tag{3c}$$

$$\boldsymbol{D}_{(1+d_w+d_l+d_c)\times k} = \boldsymbol{F} \tag{3d}$$

where $\boldsymbol{I}$ and $\vec{\boldsymbol{0}}$ are an identity matrix and a zero matrix, respectively, $\boldsymbol{F}$ is a decision matrix constructed by feature selection. RFE is adopted as the feature selection method in this paper, since it has been widely used in renewable energy or load data analyses [24–26]. RFE is a wrapper feature selection method that selects features by recursively evaluating the forecasting behaviour with smaller and smaller sets of features. RF is used as the forecasting engine, where the impurity (i.e., variance for regression trees) decrease of each feature is averaged and used to indicate the feature importance. The least important features are pruned from the current set of features in each iteration, and the set of features that generates the best forecasting is determined as the selected optimal feature set. More details about the RFE method can be found in Refs. [24–26]. After determining the selected inputs using RFE, the $\boldsymbol{F}$ matrix is constructed by replacing the element in the ith row and jth column with 1 in the initial $\vec{\boldsymbol{0}}_{(1+d_w+d_l+d_c)\times k}$ matrix, where i is the index of the selected feature in the $\boldsymbol{X}$ space and j is the index of the same feature in the newly constructed $\boldsymbol{X}^D$ space. An example of the decision matrix construction after RFE feature selection process is expressed as:

$$\begin{array}{rc} & \begin{array}{cc} j_1=1 & \\ \downarrow & \end{array} \\ \begin{array}{r} \\ i_1=2\rightarrow \\ \\ i_2=2+d_w+d_c\rightarrow \\ \\ \end{array} & \begin{bmatrix} 0 & 0 \\ \boxed{0} & 0 \\ \vdots & \vdots \\ 0 & \boxed{0} \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \\ & \begin{array}{cc} & \uparrow \\ & j_2=2 \end{array} \end{array} \xrightarrow{\text{RFE}} \begin{bmatrix} 0 & 0 \\ \boxed{1} & 0 \\ \vdots & \vdots \\ 0 & \boxed{1} \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}_{(1+d_w+d_l+d_c)\times 2}$$

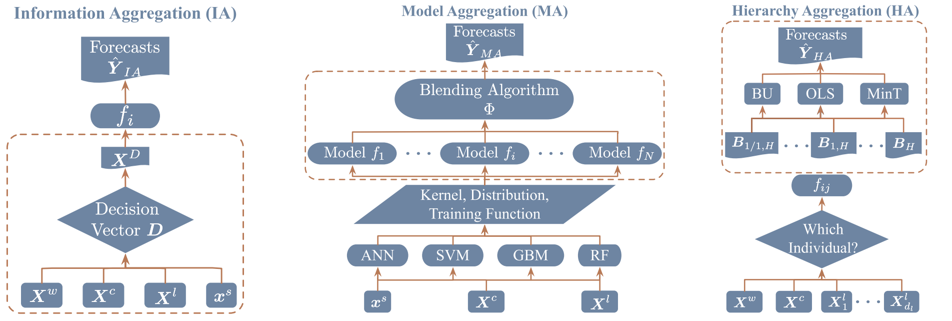
where two elements in the initial $\vec{\boldsymbol{0}}_{(1+d_w+d_l+d_c)\times 2}$ matrix are replaced by 1. Therefore, two features are selected, which are $\boldsymbol{x}_1^w$ and $\boldsymbol{x}_1^l$ as boxed in Eq. 2.1:

$$\begin{array}{c} \begin{array}{ccccccccc} & i_1=2 & & & & i_2=2+d_w+d_c & & & \\ & \downarrow & & & & \downarrow & & & \end{array} \\ \begin{bmatrix} x^s_{11} & \boxed{x^w_{11}} & \cdots & x^w_{1d_w} & \cdots & \boxed{x^l_{11}} & \cdots & x^l_{1d_l} \\ x^s_{21} & \boxed{x^w_{21}} & \cdots & x^w_{2d_w} & \cdots & \boxed{x^l_{21}} & \cdots & x^l_{2d_l} \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ x^s_{n1} & \boxed{x^w_{n1}} & \cdots & x^w_{nd_w} & \cdots & \boxed{x^l_{n1}} & \cdots & x^l_{nd_l} \end{bmatrix} \end{array} \xrightarrow{\text{RFE}} \begin{array}{c} \begin{array}{cc} j_1=1 & j_2=2 \\ \downarrow & \downarrow \end{array} \\ \begin{bmatrix} x^w_{11} & x^l_{11} \\ x^w_{21} & x^l_{21} \\ \vdots & \vdots \\ x^w_{n1} & x^l_{n1} \end{bmatrix} \end{array}$$

2.2. Model aggregation (MA)

MA carries out aggregation at the model-building stage, which is expected to take advantage of the learning power from different models. In the literature, averaging forecasts generated by multiple models is the first MA strategy, followed and advanced by a linear combination of models (i.e., weighted averaging). The latest MA strategies seek to combine individual models with dynamic weights. The

[3] STLF-MA is short for the category of STLF methods with MA, while STLF-M specifically refers to a group of models built based on the STLF-MA.



(a) Information aggregation (b) Model aggregation (c) Hierarchy aggregation

Fig. 1. Frameworks of STLF with three different aggregation strategies.

dynamic weights are adaptively assigned to base learners accordingly. For example, the gradient boosting machine (GBM) adds subsequent weak learners weights based on gradients of a loss function. In this paper, the ML-based Multi-Model forecasting framework (as shown in Fig. 1(b)) is adopted to aggregate individual forecasting models [27,28]. The used forecasting framework contains two layers (different from NN layers), the first of which consists of multiple ML models while the second of which has another blending model. The forecasting process of this method is expressed as [26]:

$$\tilde{\boldsymbol{y}}_i = f_i([\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^s]) \tag{5}$$

$$\hat{\boldsymbol{y}}_{MA} = \Phi(\tilde{\boldsymbol{Y}}) \tag{6}$$

where $\tilde{\boldsymbol{y}}_i$ is a forecast vector provided by the first-layer model f_i, $\tilde{\boldsymbol{Y}}$ is a combination of the first-layer forecast vectors, and $\hat{\boldsymbol{y}}_{MA}$ is the final forecast vector by a blending model Φ(*) in the second layer. Four ML algorithms with multiple training strategies, kernels, or distribution functions are adopted, which are ANN, SVR, GBM, and random forest (RF). Please note that all the models are used to construct the first layer. To compare STLF-M with different blending algorithms, simple averaging, linear regression, or one of the ML methods is adopted in the second-layer as a blending model in the MA framework.

2.3. Hierarchy aggregation (HA)

Load data is hierarchically aggregated based on the power grid network and geographical distributions. STLF-HA forecasts entries at one or multiple hierarchical level(s), which ensures the accuracy of every entry and the aggregate consistency among different levels. Aggregate consistency is defined as the equality between the sum of forecasts and the forecast of the sum. For example, in a three-level hierarchy shown in Fig. 2, the aggregate consistency requires $\hat{\boldsymbol{y}}_{i,HA} = \sum_j \hat{\boldsymbol{y}}_{i/j,HA}$ and $\hat{\boldsymbol{y}}_{HA} = \sum_i \hat{\boldsymbol{y}}_{i,HA}$, where i indicates the upper-level entry to which the lower-level individuals belong and j is used to identify entries within the same aggregation group. To improve the forecasting accuracy of the top-level entry ($\hat{\boldsymbol{y}}_{HA}$ in Fig. 2) while keeping the aggregate consistency, the most commonly used STLF-HA approach is BU. Other methods, such as reconciliation forecasting [19,20], are widely used in other areas. In this paper, STLF-HA with BU (STLF-B) and reconciled STLF-HA with OLS (STLF-O) and MinT (STLF-T) estimators are developed and compared.

STLF-B forecasts load of the bottom-level individuals (level 3 in Fig. 2), i.e., $\boldsymbol{b}_{i/j,HA}$, by using weather data, calendar data, and specific individual load data (i.e., $\boldsymbol{x}^l$), which are aggregated to the upper-level (level 2) until reaching the top-level (level 1). This process can be expressed by using matrix notation $\hat{\boldsymbol{Y}}_{HA} = \boldsymbol{S}\boldsymbol{b}_{i/j,HA}$, which is further expanded as [20]:

$$\hat{\boldsymbol{Y}}_{HA} = [\hat{\boldsymbol{y}}_{HA} \quad \hat{\boldsymbol{y}}_{1,HA} \quad \hat{\boldsymbol{y}}_{2,HA} \quad \hat{\boldsymbol{y}}_{1/1,HA} \quad \hat{\boldsymbol{y}}_{1/2,HA} \quad \hat{\boldsymbol{y}}_{2/1,HA} \quad \hat{\boldsymbol{y}}_{2/2,HA}]^T$$
$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_4 & & \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_{1/1,HA} \\ \boldsymbol{b}_{1/2,HA} \\ \boldsymbol{b}_{2/1,HA} \\ \boldsymbol{b}_{2/2,HA} \end{bmatrix} \tag{7}$$

where $\hat{\boldsymbol{Y}}_{HA}$ is a forecasting matrix containing all entries in the hierarchy. $\boldsymbol{S}$ is a summing matrix, which is determined by the hierarchical structure in Fig. 2. $\boldsymbol{b}_{i/j,HA}$ are base forecasts at the bottom level, shown in Fig. 2. $\boldsymbol{I}_4$ is a 4 × 4 identity matrix. Please note that the objective of this paper is to forecast the load at the top-level, $\hat{\boldsymbol{y}}_{HA}$, which might sacrifice the accuracy of forecasts at lower levels.

STLF-O and STLF-T leverage correlations and interactions between entries at different levels, which are different from STLF-B. Therefore, instead of using only the base forecasts at bottom-level ($\boldsymbol{b}_{i/j,HA}$), base forecasts of all entries in the hierarchy are optimally combined to generate the reconciled final forecasts, $\hat{\boldsymbol{Y}}_{HA}$. The base forecasting reconciliation is achieved by solving a linear regression problem [19]:

$$\boldsymbol{B}_t = \boldsymbol{S}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon} \tag{8}$$

where $\boldsymbol{B}_t$ is base forecasts of entries at all levels at time t. $\boldsymbol{\beta}_t$ is the unknown mean matrix of the bottom-level entries. $\boldsymbol{\varepsilon}$ is an error vector with zero mean and unknown variance $\boldsymbol{\Sigma}$. The minimum variance unbiased estimate of $\boldsymbol{\beta}_t$ can be obtained by using generalized least squares (GLS) estimation as [29,30]:

$$\boldsymbol{\beta}_t^{GLS} = (\boldsymbol{S}'\boldsymbol{\Sigma}^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}^{\dagger}\boldsymbol{B}_t \tag{9}$$

where $\boldsymbol{\Sigma}$† is the Moore-Penrose generalized inverse of $\boldsymbol{\Sigma}$. And the reconciled unbiased final forecasts are expressed as:

$$\hat{\boldsymbol{Y}}_t = \boldsymbol{S}\boldsymbol{\beta}_t^{GLS} \tag{10}$$

To deal with the unknown $\boldsymbol{\Sigma}$, two simplified estimators are adopted in this paper, which are OLS and MinT. The reasons to select these two estimators are that OLS is the most popular reconciliation method and MinT is the best reconciliation method as reported in the literature [20,21]. The two estimators are described as follows [20]:

$$\boldsymbol{\beta}_t^{OLS} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{B}_t \tag{11a}$$

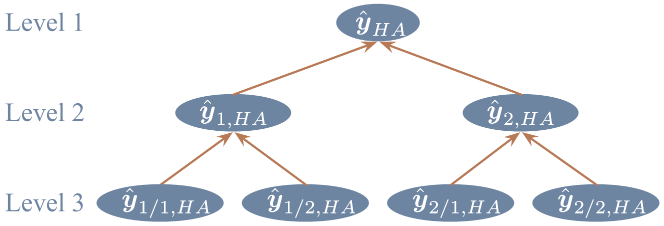


Fig. 2. A three-layer hierarchical structure.

$$\boldsymbol{\beta}_t^{MinT} = (\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{B}_t \tag{11b}$$

where $\boldsymbol{\Sigma}$ equals $k_h\boldsymbol{I}$ and $k_h\boldsymbol{W}$ in OLS and MinT, respectively. k_h is a positive scaling factor and $\boldsymbol{W}$ is a historical sample covariance matrix of base forecasting errors based on the validation dataset. Assumptions and proofs of the two simplification processes can be found in Ref. [20]. Finally, the unbiased final forecasts with the two reconciliation methods are expressed as:

$$\hat{\boldsymbol{Y}}_t^{OLS} = \boldsymbol{S}\boldsymbol{\beta}_t^{OLS} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{B}_t \tag{12a}$$

$$\hat{\boldsymbol{Y}}_t^{MinT} = \boldsymbol{S}\boldsymbol{\beta}_t^{MinT} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{B}_t \tag{12b}$$

3. Experimental setup

In this section, experimental setups for case studies are described, including the data description and pre-analysis, benchmarks and comparison settings, and evaluation metrics.

3.1. Data description and pre-analysis

In this paper, hourly load data of 13 buildings (selected based on the data availability) at the University of Texas at Dallas (UTD) is used for case studies [31]. The whole campus data is assumed to be the sum of 13 buildings' load. The reasons to research with university campus load are threefold: (i) the demand-side LF is more challenging than the upper-level LF in power system hierarchy [32], (ii) large electricity consumers, such as universities, are more critical in demand-side management, and (iii) a university campus has buildings with diverse load patterns that are interesting to explore. In addition to campus load, hourly weather information is retrieved from the National Solar Radiation Database (NSRDB).[4] The weather features in the NSRDB dataset include air temperature, relative humidity, air pressure, wind speed, wind direction, direct normal irradiance, global horizontal irradiance, and diffuse horizontal irradiance. Calendar features, i.e., the holiday indicator, hour of the day, day of the week, and month of the year, are extracted and included in all the case studies. Please note that only the latest lagged features are included as the inputs for all the models. Both UTD load and NSRDB weather data span from January 1st 2014 to December 31st 2015. The training data and validation data are randomly selected from each month, and the remaining data is used for testing. The ratio of training samples, validation samples, and testing samples is 3:1:1. We assume that by randomly partitioning days into training or testing datasets, the model generality can be better assessed. This data partitioning strategy has been widely used in power system time series forecasting, such as Global Energy Forecasting Competition (GEFCom) 2012 [33] and GEFCom 2014 [34].

Fig. 3 shows load profiles of the total 13 buildings (i.e., the top-level entry in HA) and individual buildings (i.e., the bottom-level individuals in HA). It is observed that the load profiles have evident diurnal patterns. This is also proved by a time series analysis [28,35], showing that all the load time series have a periodicity of 24 (1 day). Moreover, load patterns of the 13 buildings (B1–B13) are different, which could be further validated by load statistics as shown in Fig. 4. Among the 13 buildings, B1 is a parking structure equipped with photovoltaic panels, which may have negative netload during the daytime, as shown in Fig. 4. B2 is an administration building that has load with larger variance from 8am to 5 pm. B3 is a library that has the largest and most stable load among all buildings. B4 is a lecture hall, which has relatively small but chaotic load. B5–B9 are five classroom/lab buildings with similar patterns. B10–B13 are four student residence halls that have diverse load patterns in contrast to other buildings. Compared to individual buildings, the whole campus load (UTD) is relatively smoother.

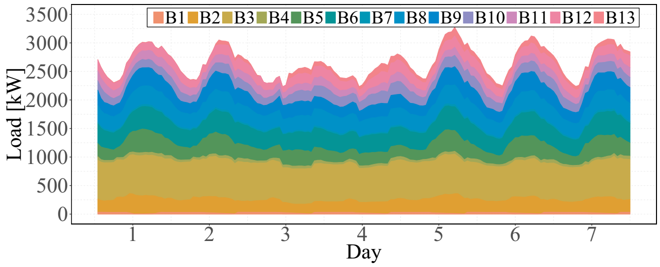


Fig. 3. UTD campus and building load profiles for seven days in spring.

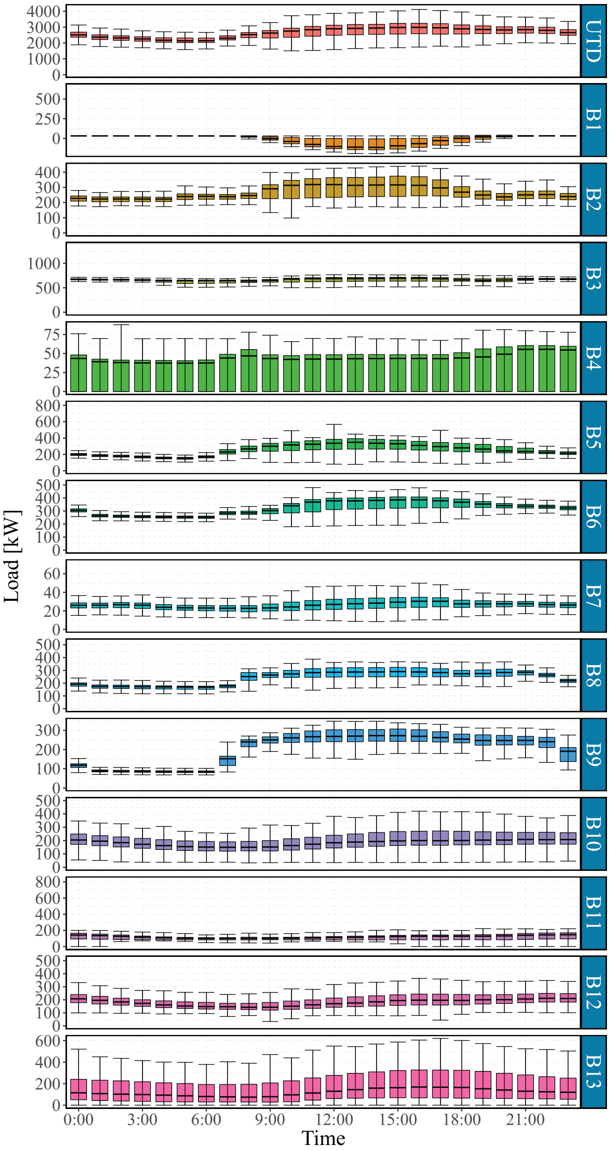


Fig. 4. Hourly statistics of buildings' and whole campus' load. Lines in the boxes are the medians. The interquartile range box represents the middle 50% of the data. The upper and lower bounds are maximum and minimum values of the data, respectively, excluding outliers (shown as points in the figure).

While the methods can be applied to different forecasting horizons, the forecasting time horizon in this paper is 1-hour-ahead. 1-hour-ahead LF plays an important role in power system operations, such as helping decision-making of real-time dispatch and energy storage charging/discharging. 1-hour-ahead LF is also flexible and scalable to

[4] https://nsrdb.nrel.gov

generate longer-term forecasts in a recursive or a parallel manner. The experiments are carried out on a laptop locally with 2.6 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 SDRAM in **R** language.

3.2. Benchmarks and comparison settings

In this paper, forecasting methods with three categories of aggregation strategies are investigated and compared, which are IA, MA, and HA. Although combining all the three aggregation strategies in a model might be able to obtain the most accurate forecasts, the three aggregation strategies are separated in different categories for better comparisons. The details of each category are summarized as follows and listed in Table 1:

- ***Category 1:*** In the IA category, STLF using weather data (STLF-W), individual buildings' load data (STLF-L), STLF-I, and STLF-F are compared. Note that the historical whole campus load data and calendar data are included in all the four groups as an input.
- ***Category 2:*** The second comparison is made between STLF-M and STLF with a single-algorithm ML model (STLF-S), both of which use weather data and calendar data (so STLF-S is the same as STLF-W). Please note that the historical whole campus load data is also a predictor in the two groups. Both simple methods (simple average and linear regression) and ML methods are adopted as blending algorithms in the second layer of the ML-based Multi-Model forecasting framework.
- ***Category 3:*** The third category contains three HA strategies, which are BU, OLS reconciliation, and MinT reconciliation. All the three HA strategies are tested by the two-level UTD load hierarchy that contains a top-level entry and 13 bottom-level entries. The inputs to the individual building load forecasting models are weather data, calendar data, and the corresponding historical individual building load.

After investigating the effectiveness of the three aggregation strategies, models in the 6 aggregation groups (i.e., STLF-I, STLF-F, STLF-M, STLF-B, STLF-O, and STLF-T) are further compared to show their superiority in STLF.

Ten state-of-the-art ML models are included in this paper for aggregation strategy implementations, which diversified by different training algorithms, kernel functions, or distribution functions. Specifically, three ANN models with standard back-propagation (BP), momentum-enhanced BP, and resilient BP training algorithms are selected based on their fast convergence and satisfactory performance [36]. The most popular kernels in SVR are used, which are linear, polynomial, and radial base function kernels. GBM models with squared, Laplace, and T-distribution loss functions are empirically selected. The last model is an RF model. The model hyperparameters are emperically determined by the validation dataset and summarized in Table 2, including the learning rate (`lr`) and the maximum number of epochs (`max_epoch`) in M1–M3; the minimum update value (`min_delta`) and the maximum update value (`max_delta`) in M1; the momentum (`momentum`) in M2; the penalty weight (C_d) and insentive parameter (ε_d) in M4–M6; the free parameter (δ_d) in M5 and M6; the degree of the polynomial (`degree`) in M5; the number of boosting iterations (`ntrees`), maximum tree depth (`max_depth`), learning rate (`lr`), out-of-bag fraction (`bag_frac`) in M7–M9; the degree of freedom (`DF`) in M9; and the number of trees (`ntrees`) and the number of variables randomly sampled as candidates at each split (`mtry`) in M10. It is important to note that all these models are used in the first-layer and only one of them is used in the second-layer in MA.

3.3. Forecasting accuracy assessment

To assess the forecasting accuracy, four evaluation metrics are used, which are normalized mean absolute error (*nMAE*), mean absolute percentage error (*MAPE*), *nMAE* improvement (Imp^A), and *MAPE* improvement (Imp^P). The mathematical expressions of the four metrics are respectively shown as [37,38]:

$$nMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_{\max}}\right| \times 100\% \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{14}$$

$$Imp_{ab}^{A} = \frac{nMAE_{M_b} - nMAE_{M_a}}{nMAE_{M_b}} \tag{15}$$

$$Imp_{ab}^{P} = \frac{MAPE_{M_b} - MAPE_{M_a}}{MAPE_{M_b}} \tag{16}$$

where $\hat{y}$, y, and $y_{\max}$ are the forecasting value, actual value, and maximum actual value, respectively; i is a sample index and n is the number of samples; M is the model name; a and b are group indices to which a model belongs. Specifically, a and b could be selected from L, W, S, I, F, M, B, O, and T, which represent the groups of STLF-L, STLF-W, STLF-S, STLF-I, STLF-F, STLF-M, STLF-B, STLF-O, and STLF-T, respectively. For example, Imp_{IW}^{A} means the improvement of STLF-I over STLF-W based on *nMAE*. It is important to note that both Imp^A and Imp^P are calculated based on the same model M, because the focus of this paper is to compare STLF with different aggregation strategies, instead of comparing STLF using different ML models. Quantifying the overall performance of forecasting models from different perspectives, however, the four metrics are not able to detail the local accuracy of forecasts (e.g., all the bias are positive since the absolute value calculation). Therefore, two other error metrics used for visualization in Section 4 are bias error (*BE*, which is also known as the forecasting residual) and normalized bias error (*nBE*), which are expressed as:

$$BE_i = \hat{y}_i - y_i \tag{17}$$

$$nBE_i = \frac{\hat{y}_i - y_i}{y_i} \times 100\% \tag{18}$$

4. Results and discussion

4.1. Effectiveness of IA

Four groups of STLF models in the IA category are tested and their forecasting errors and comparison results are summarized in Tables 3 and 4. It is observed from Table 3 that different ML models perform distinctively. For instance, the forecasting *nMAE* of STLF-W models ranges from 1.05% to 1.83%. In general, the last three models (two GBM models and one RF model), i.e., M8–M10, forecast more accurately than other models in all the four groups. This is due to the stronger learning power of the three ensemble models. In addition, the weather information has larger impacts on forecasting model performance than individual building load data, since all the models in the

Table 1
Different aggregate forecasting categories and groups.

Category	Group	Group index	Input
IA	STLF-W	W	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^s]$
	STLF-L	L	$[\boldsymbol{X}^l, \boldsymbol{X}^c, \boldsymbol{x}^s]$
	STLF-I	I	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{X}^l, \boldsymbol{x}^s]$
	STLF-F	F	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{X}^l, \boldsymbol{x}^s]$
MA	STLF-S	S	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^s]$
	STLF-M	M	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^s]$
HA	STLF-B	B	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^l]$
	STLF-O	O	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^l]$
	STLF-T	T	$[\boldsymbol{X}^w, \boldsymbol{X}^c, \boldsymbol{x}^l]$

Table 2
Machine learning models.

Algorithm	Model	Function/algorithm	Hyperparameter
ANN	M1	Resilient back-propagation (BP)	lr = 0.01, max_epoch = 1,000, min_delta = 1 × 10^6, max_delta = 50
	M2	Momentum-enhanced BP	lr = 0.01, max_epoch = 1,000, momentum = 0.9
	M3	Standard BP	lr = 0.01, max_epoch = 1,000
SVR	M4	Linear kernel	$C_d = 0.1$, $\varepsilon_d = 0.001$
	M5	Polynomial kernel	$C_d = 0.1$, $\varepsilon_d = 0.001$, $\delta_d = 0.1$, degree = 3
	M6	Radial basis function kernel	$C_d = 0.1$, $\varepsilon_d = 0.001$, $\delta_d = 0.1$
GBM	M7	Squared loss	lr = 0.01, ntrees = 1,000, max_depth = 20, bag_frac = 0.5
	M8	Laplace loss	lr = 0.01, ntrees = 1,000, max_depth = 20, bag_frac = 0.5
	M9	T-distribution loss	lr = 0.01, ntrees = 1,000, max_depth = 20, bag_frac = 0.5, DF = 4
RF	M10	CART aggregation	ntrees = 1,000, mtry = 5

Table 3
Forecasting *nMAE* [%] and *MAPE* [%] in the IA category.

Model	STLF-W		STLF-L		STLF-I		STLF-F	
	nMAE	*MAPE*	*nMAE*	*MAPE*	*nMAE*	*MAPE*	*nMAE*	*MAPE*
M1	1.68	2.67	1.69	2.76	1.38	2.22	1.29	2.10
M2	1.34	2.20	1.58	2.56	1.26	2.10	1.35	2.24
M3	1.36	2.20	1.70	2.76	1.36	2.22	1.32	2.13
M4	1.83	3.01	1.72	2.81	1.53	2.50	1.57	2.55
M5	1.46	2.35	1.57	2.53	1.31	2.12	1.37	2.21
M6	1.43	2.32	1.44	2.36	1.26	2.07	1.16	1.90
M7	1.67	2.71	1.83	2.95	1.67	2.70	1.67	2.69
M8	1.08	1.74	1.36	2.23	1.00	1.64	1.03	1.68
M9	1.28	2.14	*1.29*	*2.13*	1.16	1.91	1.06	1.73
M10	*1.05*	*1.72*	1.32	2.17	***0.99***	***1.61***	*1.00*	*1.63*
Average	1.42	2.31	1.55	2.53	1.29	2.11	1.28	2.09

Note: *Italic values* indicate the best results within the same group and **bold values** indicate the smallest forecasting errors among all models.

Table 4
Forecasting improvements in the IA category according to Imp^A [%] and Imp^P [%].

	Imp^A_{IW}	Imp^P_{IW}	Imp^A_{IL}	Imp^P_{IL}	Imp^A_{FI}	Imp^P_{FI}
M1	*17.86*	16.85	18.34	19.57	6.52	5.41
M2	5.97	4.55	20.25	17.97	−7.14	−6.67
M3	0.00	−0.91	20.00	19.57	2.94	4.05
M4	16.39	*16.94*	11.05	11.03	−2.61	−2.00
M5	10.27	9.79	16.56	16.21	−4.58	−4.25
M6	11.89	10.78	12.50	12.29	7.94	8.21
M7	0.00	0.37	8.74	8.47	0.00	0.37
M8	7.41	5.75	***26.47***	***26.46***	−3.00	−2.44
M9	9.38	10.75	10.08	10.33	*8.62*	*9.42*
M10	5.71	6.40	25.00	25.81	−1.01	−1.24
Average	8.49	8.13	16.90	16.77	0.79	1.09

Note: *Italic values* indicate the most improvements within the same comparison while **bold values** identify the most improvements in all comparisons.

STLF-W group have smaller forecasting errors than those in the STLF-L group, except for M4 and M9. However, the influence of the inputs on models is varying. For example, the forecasting results could be competitive (e.g., M9) or even worse (e.g., M4) by using individual building load compared to those using weather data.

It is found from Table 4 that STLF-I models reduce forecasting errors notably and consistently, compared with STLF-W and STLF-L models. The accuracy improvements are more evident by aggregating weather information data into forecasting models. Regarding different models, M1 (an ANN model) and M8 (a GBM model) are enhanced the most by IA. A further comparison is made between STLF-I and STLF-F models, where the RFE feature selection further improves some of the models, such as M9. *It is concluded that IA improves STLF forecasting accuracy notably and consistently.*

Table 5
Forecasting *nMAE* [%], *MAPE* [%], Imp^A_{MS} [%], and Imp^P_{MS} [%] in the MA category.

	Model	STLF-S		STLF-M		Imp^A_{MS}	Imp^P_{MS}
		nMAE	*MAPE*	*nMAE*	*MAPE*		
SP	M0†	NA	NA	1.38	2.26	NA	NA
	M0*	NA	NA	*1.10*	*1.77*	NA	NA
ML	M1	1.68	2.67	1.32	2.16	27.27	19.10
	M2	1.34	2.20	1.36	2.24	−1.47	−1.82
	M3	1.36	2.20	1.45	2.35	−6.21	−6.82
	M4	1.83	3.01	1.35	2.22	*35.56*	*26.25*
	M5	1.46	2.35	1.20	1.98	21.67	15.74
	M6	1.43	2.32	1.76	2.85	−18.75	−22.84
	M7	1.67	2.71	1.39	2.32	20.14	14.39
	M8	1.08	1.74	1.24	2.06	−12.9	−18.39
	M9	1.28	2.14	1.15	1.89	11.30	11.68
	M10	***1.05***	***1.72***	1.16	1.91	−9.48	−11.05
	Average	1.42	2.31	1.34	2.20	6.71	2.62

Note: *Itaic values* indicate the best results within the same group, while **bold values** indicate the smallest forecasting errors or the most significant improvements among all models. M0† and M0* are ML-based Multi-Model forecasting frameworks with simple averaging and linear regression in the second layer.

4.2. Effectiveness of MA

MA forecasting evaluation results are summarized in the first 4 columns of Table 5. The comparisons of MA with STLF-S are shown in the 5th and 6th columns of the same table. It is found that the performance of the relatively less-accurate STLF-S models is improved more notably by MA, such as M4 and M7. However, the best two models in STLF-S, i.e., M8 and M10, deteriorate in STLF-M, which is partially due to the unsatisfactory forecasts ($\tilde{\boldsymbol{Y}}$) from part of the first-layer models. Regarding second-layer blending models, two linear models, M0* and M4, outperform other models, possibly due to the linear relationship between the first-layer forecasts ($\tilde{\boldsymbol{Y}}_{ij}$) and the load observations. By comparing blending models with ML algorithms, all the ANN models (i.e., M1–M3) and SVR with linear and polynomial kernels (i.e., M4 and M5) perform relatively better in STLF-M. Among the four different ensemble learning algorithm models (M7–M10), two of them (i.e., M7 and M9) have increasing accuracies while the other two (i.e., M8 and M10) have decreasing accuracies using the MA strategy. Though three models (i.e., M6, M8, and M10) produce worse forecasts, their forecasting accuracies are still competitive. *Therefore, it is concluded that MA enhances STLF robustness by reducing the risk of unsatisfactory single-algorithm ML models.*

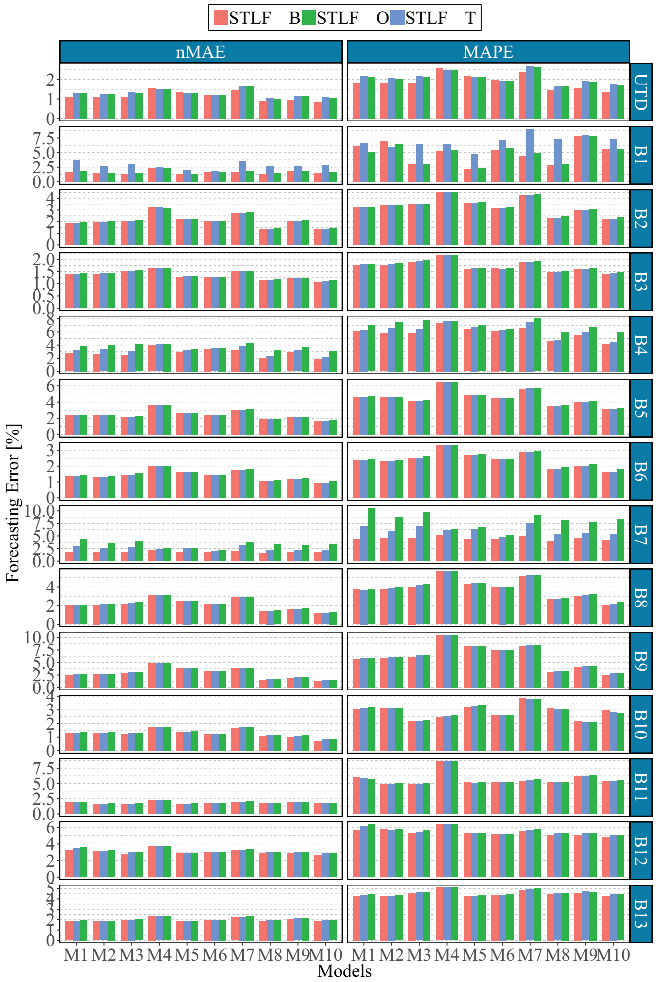


Fig. 5. Forecasting errors in the HA category. M_B, M_O, and M_T are models in STLF-B, STLF-O, and STLF-T groups, respectively.

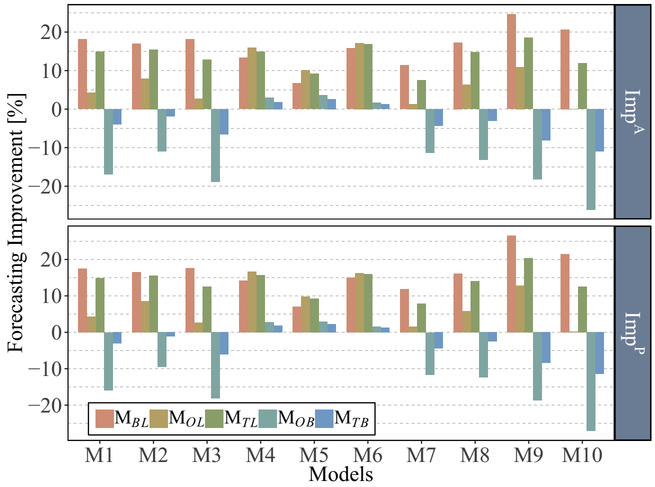


Fig. 6. Forecasting improvements regarding to the top-level entry (UTD) in the HA category.

4.3. Effectiveness of HA

The forecasting *nMAE* and *MAPE* using STLF-HA models with three different HA methods are illustrated by barplots in Fig. 5. By comparing entries in different levels of the hierarchy, it is found that the bottom-level forecasting errors are canceled out by aggregating bottom-level forecasts (B1–B13) to top-level forecasts (UTD). Taking the M1 model in STLF-B as an example, most of the individual load (B1–B13) forecasting *nMAEs* are obviously larger than 1.10%, while its forecasting error of UTD is only 1.10%. Regarding forecasting models, though no single model always outperforms others, three ensemble learning models (M8–M10) perform more accurately, especially for buildings B2, B4, and B8.

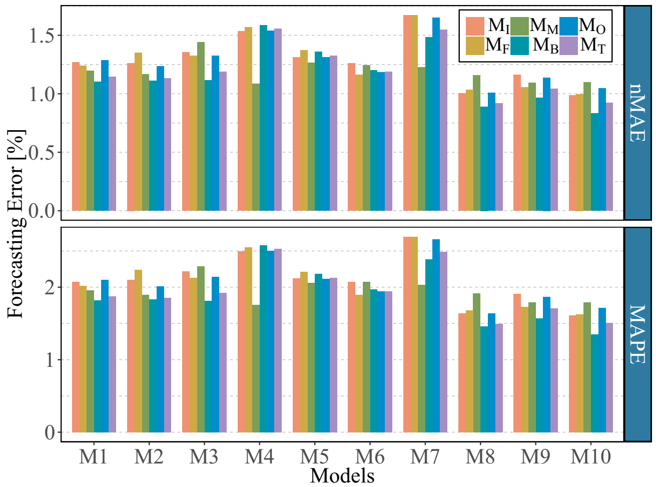
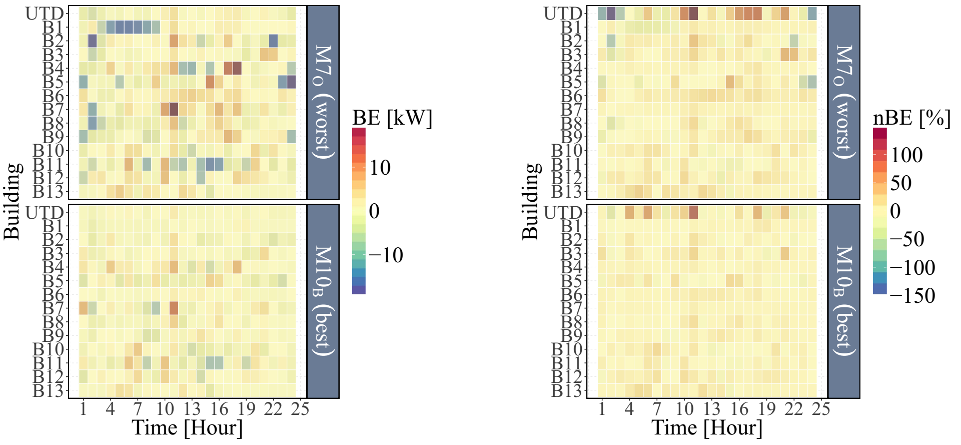


Fig. 7. Forecasting errors of models in different aggregation groups. M_I, M_F, M_M, M_B, M_O, and M_T are models in STLF-I, STLF-F, STLF-M, STLF-B, STLF-O, and STLF-T groups, respectively.

The comparisons (only the top level) between STLF-HA models with STLF-S models (denoted as M_{BL}, M_{OL}, and M_{TL}) are shown in Fig. 6. It is observed that all the three HA strategies improve the top-level entry's forecasting accuracy using all 10 models, as indicated by the positive bars of M_{BL}, M_{OL}, and M_{ML}. As opposed to STLF-S, HA methods improve STLF by up to 24.63% and 26.59% based on Imp^A and Imp^P, respectively. By comparing three HA strategies (indicated by M_{OB} and M_{TB} in Fig. 6), it is found that only SVR models (M4–M6) are enhanced by *OLS* and *MinT* strategies, which means the more advanced *OLS* and *MinT* methods do not outperform *BU* consistently as expected in the selected case studies. *Overall, it is concluded that HA is able to provide more accurate forecasts while keeping aggregate consistency in the hierarchy.*

4.4. Superiority of different aggregation strategies

The outperformance and advantages of the three aggregation strategies over STLF-L and STLF-W have been validated in Sections 4.1–4.3. In this subsection, further comparisons are conducted among the different aggregation strategies. Results of STLF with three aggregation strategies are visualized in Fig. 7. Most STLF-B and STLF-T models outperform their counterparts in STLF-I, STLF-F, and STLF-M groups, such as ANN, GBM, and RF models. However, SVR models in the STLF-B group (M4–M6) are beaten by the same models in the STLF-I and STLF-F groups. Furthermore, two models (i.e., M4 and M7) with HA strategies produce worse forecasts than those with MA strategy. The forecasting accuracy deterioration of the STLF-HA models is due to the individual forecasting error accumulation effect, which is illustrated in Fig. 8. Two contrasts shown in Fig. 8 are STLF-O with M7 ($M7_O$) and STLF-B with M10 ($M10_B$), which are the worst and the best STLF-HA models, respectively. It is observed from Fig. 8(a) that $M10_B$ generates forecasts with smaller bias for each individual building than $M7_O$, such as B2 and B4. Moreover, the individual buildings' LF errors of $M7_O$ accumulate to larger values in contrast with those of $M10_B$, which is illustrated by the darker colors of the whole campus' forecasting errors in Fig. 8(b). Though there are some unsatisfactory models compared



(a) Bias errors (BEs)

(b) Normalized bias errors (nBEs)

Fig. 8. UTD campus and building forecasting errors using the best and worst STLF-HA.

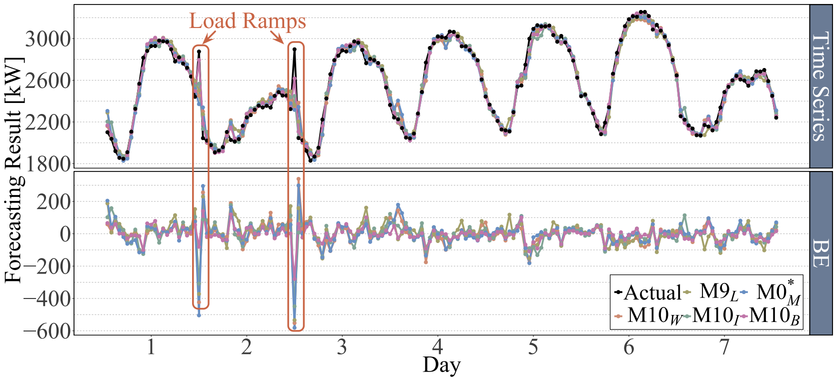


Fig. 9. Forecasting and bias error series of the best models in different groups.

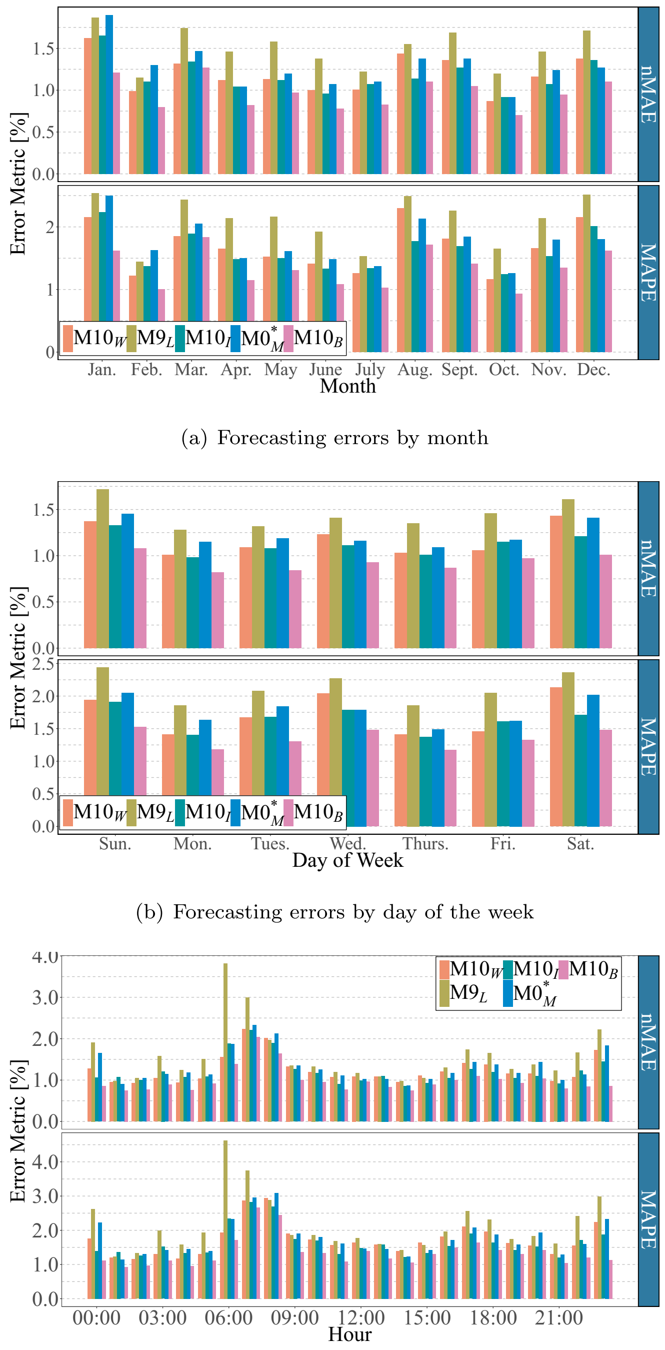
with the other two aggregation strategies, the overall improvement of STLF-HA is obvious. *Additionally, STLF-HA produces the most accurate forecasts (0.83% nMAE and 1.35% MAPE) among all models.*

The best model in each group is picked out to make further comparisons, which are M10 in the STLF-W/STLF-S group (M10$_W$), M9 in the STLF-L group (M9$_L$), M10 in the STLF-I group (M10$_F$), M0* in the STLF-M group (M0$^*_M$), and M10 in the STLF-B group (M10$_B$). Fig. 9 shows 1-week actual, forecasting, and bias error time series of the selected five models. It is observed that M10$_B$ has smaller errors than the other four models, especially during load ramps (enclosed by red boxes in Fig. 9). To characterize forecasting performance of the five models, forecasting errors with respect to calendar units (i.e., month of the year, day of the week, and hour of the day) are shown in Fig. 10(a)–(c). One interesting finding is that the calendar effect has considerable impacts on forecasting errors. For example, errors in January, August, and September are much larger than those in other months. This is possibly due to the load pattern variation by the university holidays. The calendar effect on forecasting errors is even more evident by hour of the day, as shown in Fig. 10(c). Forecasts deviate the most from 6am to 8am, during which load patterns change more considerably. However, no evident calendar effect is found on forecasting errors by day of the week, as shown in Fig. 10(b). This is possibly due to the diverse building load of the university, for example, classroom and library buildings have higher load during weekdays and residential halls have higher load during weekends. *Even though the load pattern varies a lot, it is observed that M10$_B$ presents superior performance in every month, every day of the week, and at every hour of the day than the best models in other groups.*

5. Conclusion

This paper developed and compared short-term load forecasting (STLF) with different aggregation strategies, including information aggregation (IA), model aggregation (MA), and hierarchy aggregation (HA). The three aggregation strategies integrated distinct objectives at different stages in the forecasting process. STLF-IA aggregates more informative and better-organized data. STLF-MA aggregates forecasts of different ML models and takes advantage of their learning abilities. STLF-HA aggregates lower-level forecasts into higher level forecasts in the hierarchical structure. Case studies based on 2-year of hierarchical smart meter data showed that:

(i) STLF with the three aggregation strategies improved forecasting accuracy, compared with benchmarks without aggregation.
(ii) STLF-I presented superior performance than STLF with weather data and STLF with individual load data consistently.
(iii) MA improved the STLF robustness by reducing the risk of unsatisfactory single-algorithm STLF models.
(iv) HA produced the most accurate forecasts while keeping hierarchical aggregate consistency in distinctive load pattern scenarios caused by calendar effects.



(a) Forecasting errors by month

(b) Forecasting errors by day of the week

(c) Forecasting errors by hour of the day

Fig. 10. Calendar effects on forecasting errors.

CRediT authorship contribution statement

Cong Feng: Data curation, Formal analysis, Investigation, Methodology, Validation, Writing - original draft. **Jie Zhang:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing - review & editing.

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.

Acknowledgments

This work is supported by the International Institute of Forecasters-SAS (IIF-SAS) Research Award.

References

[1] T. Hong, S. Fan, Probabilistic electric load forecasting: a tutorial review, Int. J. Forecast. 32 (3) (2016) 914–938.
[2] K.G. Boroojeni, M.H. Amini, S. Bahrami, S. Iyengar, A.I. Sarwat, O. Karabasoglu, A novel multi-time-scale modeling for electric power demand forecasting: from short-term to medium-term horizon, Electr. Power Syst. Res. 142 (2017) 58–73.
[3] J. Xie, T. Hong, Temperature scenario generation for probabilistic load forecasting, IEEE Trans. Smart Grid 9 (3) (2016) 1680–1687.
[4] J. Xie, Y. Chen, T. Hong, T.D. Laing, Relative humidity for load forecasting models, IEEE Trans. Smart Grid 9 (1) (2016) 191–198.
[5] W. Kong, Z.Y. Dong, D.J. Hill, F. Luo, Y. Xu, Short-term residential load forecasting based on resident behaviour learning, IEEE Trans. Power Syst. 33 (1) (2018) 1087–1088.
[6] P. Jiang, F. Liu, Y. Song, A hybrid forecasting model based on date-framework strategy and improved feature selection technology for short-term load forecasting, Energy 119 (2017) 694–709.
[7] M. Saviozzi, S. Massucco, F. Silvestro, Implementation of advanced functionalities for distribution management systems: load forecasting and modeling through artificial neural networks ensembles, Electr. Power Syst. Res. 167 (2019) 230–239.
[8] R. Zhang, Z.Y. Dong, Y. Xu, K. Meng, K.P. Wong, Short-term load forecasting of Australian national electricity market by an ensemble model of extreme learning machine, IET Gener. Transm. Distrib. 7 (4) (2013) 391–397.
[9] L. Xiao, W. Shao, T. Liang, C. Wang, A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting, Appl. Energy 167 (2016) 135–153.
[10] S. Hassan, A. Khosravi, J. Jaafar, Examining performance of aggregation algorithms for neural network-based electricity demand forecasting, Int. J. Electr. Power Energy Systems 64 (2015) 1098–1105.
[11] M.H. Alobaidi, F. Chebana, M.A. Meguid, Robust ensemble learning framework for day-ahead forecasting of household based energy consumption, Appl. Energy 212 (2018) 997–1012.
[12] A. Khwaja, A. Anpalagan, M. Naeem, B. Venkatesh, Joint bagged-boosted artificial neural networks: using ensemble machine learning to improve short-term electricity load forecasting, Electr. Power Syst. Res. 179 (2020) 106080.
[13] C. Feng, J. Zhang, Reinforcement learning based dynamic model selection for short-term load forecasting, 2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), IEEE, 2019, pp. 1–5.
[14] C. Feng, M. Sun, J. Zhang, Reinforced deterministic and probabilistic load forecasting via q-learning dynamic model selection, IEEE Trans. Smart Grid 11 (2) (2019) 1377–1386.
[15] R. Sevlian, R. Rajagopal, A scaling law for short term load forecasting on varying levels of aggregation, Int. J. Electr. Power Energy Systems 98 (2018) 350–361.
[16] C.E. Borges, Y.K. Penya, I. Fernandez, Evaluating combined load forecasting in large power systems and smart grids, IEEE Trans. Ind. Inf. 9 (3) (2013) 1570–1577.
[17] C. Feng, J. Zhang, Short-term load forecasting with different aggregration strategies, ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, 2018.
[18] Y. Wang, Q. Chen, M. Sun, C. Kang, Q. Xia, An ensemble forecasting method for the aggregated load with sub profiles, IEEE Trans. Smart Grid 9 (4) (2018) 3906–3908.
[19] R.J. Hyndman, R.A. Ahmed, G. Athanasopoulos, H.L. Shang, Optimal combination forecasts for hierarchical time series, Comput. Stat. Data Anal. 55 (9) (2011) 2579–2589.
[20] S.L. Wickramasuriya, G. Athanasopoulos, R. Hyndman, Forecasting hierarchical and grouped time series through trace minimization, Technical Report, Monash University, Department of Econometrics and Business Statistics, 2015.
[21] D. Yang, H. Quan, V.R. Disfani, L. Liu, Reconciling solar forecasts: geographical hierarchy, Sol. Energy 146 (2017) 276–286.
[22] M.T. Hagan, S.M. Behr, The time series approach to short term load forecasting, IEEE Trans. Power Syst. 2 (3) (1987) 785–791.
[23] S. Fan, L. Chen, W.-J. Lee, Short-term load forecasting using comprehensive combination based on multimeteorological information, IEEE Trans. Ind. Appl. 45 (4) (2009) 1460–1466.
[24] C. Zhang, Y. Li, Z. Yu, F. Tian, Feature selection of power system transient stability assessment based on random forest and recursive feature elimination, Power and Energy Engineering Conference (APPEEC), 2016 IEEE PES Asia-Pacific, IEEE, 2016, pp. 1264–1268.
[25] O. Kramer, N.A. Treiber, M. Sonnenschein, Wind power ramp event prediction with support vector machines, International Conference on Hybrid Artificial Intelligence Systems, Springer, 2014, pp. 37–48.
[26] C. Feng, M. Cui, B.-M. Hodge, J. Zhang, A data-driven multi-model methodology with deep feature selection for short-term wind forecasting, Appl. Energy 190 (2017) 1245–1257.
[27] C. Feng, M. Cui, B.-M. Hodge, S. Lu, H. Hamann, J. Zhang, Unsupervised clustering-based short-term solar forecasting, IEEE Trans. Sustain. Energy (2018) 2174–2185.
[28] C. Feng, J. Zhang, Hourly-similarity based solar forecasting using multi-model machine learning blending, IEEE PES General Meeting 2018, IEEE PES, 2018.
[29] Generalized Least Squares, ([Online]. Available at: https://en.wikipedia.org/wiki/

Generalized_least_squares). [Accessed 18 May 2018].
[30] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[31] J.Z.C. Feng, Short-term load forecasting data with hierarchical advanced metering infrastructure and weather features, 2019. 10.21227/jdw5-z996.
[32] S.B. Taieb, J. Yu, M.N. Barreto, R. Rajagopal, Regularization in hierarchical time series forecasting with application to electricity smart meter data, AAAI, (2017), pp. 4474–4480.
[33] T. Hong, P. Pinson, S. Fan, Global energy forecasting competition 2012, Int. J. Forecast. 30 (2) (2014) 357–363.
[34] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R.J. Hyndman, Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond, Int. J. Forecast. 32 (3) (2016) 896–913.
[35] C. Feng, E.K. Chartan, B.-M. Hodge, J. Zhang, Characterizing time series data diversity for wind forecasting, Big Data Computing Applications and Technologies (BDCAT), 2017 IEEE/ACM 4th International Conference on, IEEE, 2017.
[36] C.N. Bergmeir, J.M. Benítez Sánchez, Neural networks in R using the Stuttgart neural network simulator: RSNNS, J. Stat. Softw. 46 (7) (2012) 1–26.
[37] C. Feng, M. Cui, M. Lee, J. Zhang, B.-M. Hodge, S. Lu, H.F. Hamann, Short-term global horizontal irradiance forecasting based on sky imaging and pattern recognition, IEEE PES General Meeting, IEEE, 2017.
[38] L. Xiao, W. Shao, M. Yu, J. Ma, C. Jin, Research and application of a hybrid wavelet neural network model with the improved cuckoo search algorithm for electrical power system forecasting, Appl. Energy 198 (2017) 203–222.