IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2018, accepted January 5, 2019, date of publication January 11, 2019, date of current version February 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2892475

Hybrid Load Forecasting for Mixed-Use Complex Based on the Characteristic Load Decomposition by Pilot Signals

KANGGU PARK, SEUNGWOOK YOON, (Student Member, IEEE), AND EUISEOK HWANG, (Member, IEEE)

Department of Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Euiseok Hwang (euiseokh@gist.ac.kr)

This work was supported in part by the GIST Research Institute (GRI) Grant funded by GIST in 2018, and in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of South Korea under Grant 20171210200810.

ABSTRACT In this paper, a characteristic load decomposition (CLD)-based day-ahead load forecasting scheme is proposed for a mixed-use complex. The aggregated load of the complex is composed of the mixtures of different electricity usage patterns, and short-term load forecasting can be implemented by summing disaggregated sub-load predictions. However, tracing all usage patterns of sub-loads for prediction may be infeasible because of limited resources for measurement and analysis. To prevent this infeasibility, the proposed scheme focuses on effective decomposition using the sub-loads of typical characteristic load profiles and their representative pilot signals. Separate forecasts are obtained for the decomposed characteristic sub-loads using a hybrid scheme, which combines day-type conditioned linear prediction with long short-term memory regressions. Complex campus load data are considered to evaluate the proposed CLD-based hybrid forecasting. The evaluation results show that the proposed scheme outperforms conventional hybrid or similar-day-based forecasting approaches. Even when sub-load measurements are available only for a limited period, the CLD scheme can be applied for the extended training data through virtual disaggregations.

INDEX TERMS Day-ahead load forecasting, time series analysis, long short-term memory, hybrid forecasting model, characteristic load decomposition, hierarchical load forecasting.

I. INTRODUCTION

Over the past several decades, electricity has become an essential part of modern life, contributing to growth of the global economy and energy consumption levels. In recent years, there have been rapid increases in renewable energy generation and individual energy consumption levels [1], resulting in increased power grid uncertainty. Hence, technology for the smart and efficient management of grid uncertainty has attracted research interest and several researches on building energy management system (BEMS) [2] and home energy management system (HEMS) [3] have been actively carried out. Various demand-side management schemes and supply-side controls have been intensively studied to reduce this power grid uncertainty, e.g., flexible vehicle-to-grid (V2G) coordination schemes were studied for energy cost reduction [4] and an energy controller algorithm was proposed to reduce the peak load in residential distribution networks comprising electric vehicles, photovoltaic units and battery energy storage systems [5]. In particular, load forecasting is a key enabler of smart grid applications, such as demand response, as it can reliably predict the demand flexibility and potential problems in a grid [6]. Similarly, grid stability can be improved through the prediagnosis of critical problems such as blackouts, frequency variations and generator drops [7]. In addition, load forecasting can contribute to the efficient integration and wide allocation of distributed energy resources and their coordination to accommodate supply and demand [8]. However, several challenges arise when the reliable forecasting of modern complex electrical loads is attempted, and considerable research efforts have been expended on overcoming these challenges by employing big data [9].

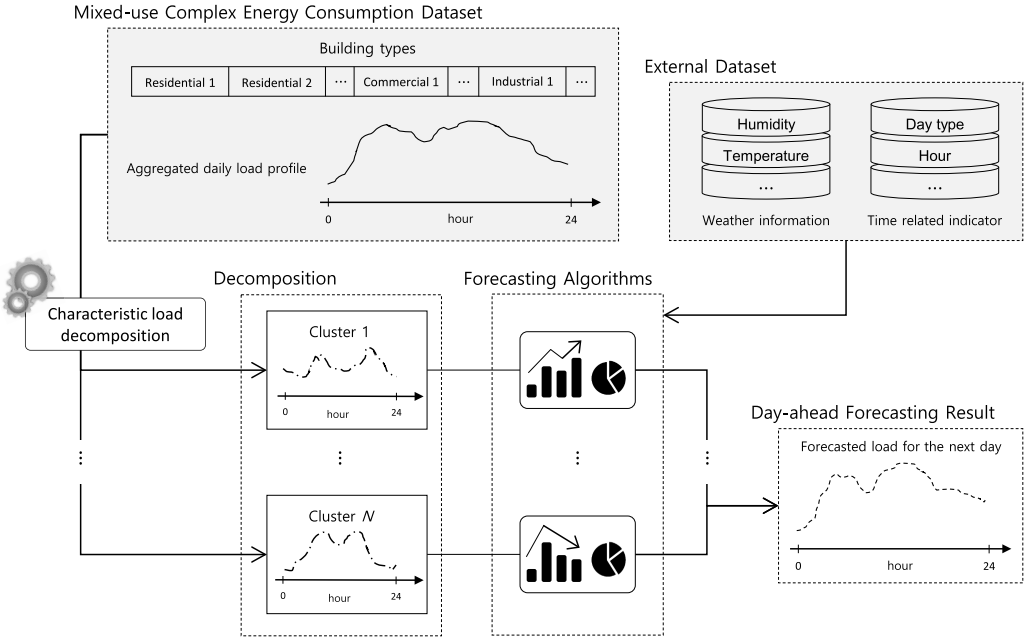


FIGURE 1. Overall flow of the proposed characteristic load decomposition (CLD)-based load forecasting scheme. Aggregated load can be decomposed to cluster N by the CLD method. Then, the decomposed loads are predicted by forecasting algorithms with external data to obtain the profile on the next day.

Among existing challenges, this study focuses on a novel forecasting scheme for an aggregated load with a hierarchical structure and mixed usage properties, such as that for a campus complex. As various power entities coexist, the aggregated load consists of the sum of various characteristic patterns, and prediction uncertainty is expected to increase with the variability of the patterns. To mitigate this uncertainty effect, sub-loads of similar behaviors can be clustered and predicted separately to incorporate their unique profiles. However, metering all sub-loads for cluster-based forecasting is infeasible because of the prohibitive overheads for installing, sensing and storing all measurements.

In this paper, as an alternative approach for sub-load prediction, characteristic load decomposition (CLD) is proposed to estimate sub-cluster loads from an upper-level aggregated load for day-ahead load forecasting. The overall architecture of the proposed scheme is illustrated in Fig. 1. First, the aggregated load is decomposed into sub-cluster loads using only the representative individual load profile of each characteristic sub-cluster, referred to as the ''pilot load or pilot signal,'' and the decomposition weight vector based on the monthly overall relative power consumption. Then, the day-ahead forecasting scheme is separately applied to the decomposed clusters with side information such as weather data and time related indicators. Finally, the forecasting results of the clusters are combined for the aggregated load forecast. Note that the individual load data for decomposition may be restricted temporally by limited sensing and storage resources. To overcome this problem, the proposed CLD can be extended with a virtual disaggregation scheme based on a similar weather day approach. To evaluate the performance of the proposed scheme, day-ahead load forecasting is conducted using the aggregated load of a college campus complex in conjunction with the sub-loads of several pilot buildings in the complex. As electricity consumption is affected by various factors that have nonlinear impacts, such as external environment variables and historical load, a hybrid forecasting model is implemented in this study by combining customized linear and nonlinear prediction models. The aggregated load can be divided into two clusters through the analysis of the load characteristics of campus complex data, and CLD can be applied based on the loads of the representative buildings in each cluster. Simulation results show that the proposed CLD-based load forecasting provides improved prediction accuracy compared to conventional schemes such as similar-day-based load forecasting and hybrid approaches without decomposition. Furthermore, the proposed scheme outperforms the state-of-art method related to the decomposition of electricity feeders in terms of decomposition accuracy and forecasting result [10].

The remainder of this paper is organized as follows: Section II briefly introduces the related works and background information on load forecasting models. Section III presents the proposed CLD scheme for a mixed-use complex and its virtual extension scheme for training data. Section IV evaluates the proposed scheme using a campus complex data, and Section V concludes this paper.

II. RELATED WORKS AND BACKGROUND

The proposed CLD scheme is designed for short-term load forecasting (STLF) and is applied to allow for separate hybrid forecasting for the linear and nonlinear contributions to an

aggregated load. Conventional STLF and hybrid techniques are briefly introduced in this section.

A. SHORT-TERM LOAD FORECASTING (STLF)

Load forecasting methods are categorized based on the range of the forecasting time horizon [11]. Among these methods, STLF is crucial for the short-term scheduling of power supply and demand, which is required for various grid services. For instance, numerous independent system operators implement day-ahead demand response market [6] and generator planning to achieve power system stability [12]; these techniques can be applied based on STLF. However, forecasting errors lead to increased operating costs and power-grid instability. Thus, a large number of studies have been conducted on effective STLF schemes.

In particular, STLF for complex loads with hierarchical layers has been actively studied to overcome the challenges in handling the diverse characteristics of various sub-loads. A multiregion forecasting system has been introduced to find the optimal partitioning of a region according to weather and load conditions. Instead of aggregated load forecasting, predictions can be made separately for partitioned regions, yielding improved forecasting accuracy [13]. Similarly, the forecasting of an aggregated load with the weighted sum of sister loads has been introduced, where a sister load is modeled in each geographical zone [14]. In addition, an algorithm has been suggested for combining the synthesizing information from different layers obtained by a smart meter, which is different to from conventional bottom-up or top-down approaches [15]. Decomposition based on a load condition constitutes a method of predicting the day-ahead base, intermediate and peak loads from each suitable forecasting model. Power data are clustered using the k-means method to determine the daily base, intermediate and peak loads. Then, a neural network is utilized as a load forecasting model [16]. Recently, a categorical load decomposition approach was proposed, where quadratic programming (QP) was employed to separate categorical profiles from various mixtures of customer load profiles [10], and Gaussian mixture model and hierarchical clustering were applied to identify the categorical building load with two-step clustering [17]. Alternatively, a *Physarum*-based hybrid optimization algorithm was suggested, which provides adaptable solutions for the load-shedding problem in a microgrid system [18]. In addition, a hybrid algorithm that combines k-means-based similar day selection, empirical mode decomposition and long short-term memory (LSTM) was presented for STLF [19].

As electricity usage is influenced by various factors, it has high variability and nonlinearity. Thus, the selection of suitable variables is necessary for reliable load forecasting. The permutation importance scheme based on random forests evaluates prediction accuracy before and after permuting the variables averaged over all trees [20]. Furthermore, a conditional permutation importance scheme was suggested because a permutation importance scheme can be affected by a violation of either the independence of a permuted variable and an observed variable or the independence of a permuted variable and other unpermuted variables [21]. The conditional permutation scheme investigates permutation importance through categorization within groups. In addition, other related studies have been conducted to facilitate the selection of significant input variables, e.g., 7 importance methods for assessing the relative importance of independent variables in a multilayer perceptron neural network were compared [22] and an investigation of the selection performance of three input selection techniques was presented to select the appropriate inputs for the time series forecasting models [23].

Among the numerous studies on STLF, similar-day-based load forecasting schemes have been studied over several decades as they are simple and intuitive. In this study, two similar-day-based load forecasting schemes are utilized for comparison with the proposed scheme. In [24], a weighted nearest neighbor (WNN) method is introduced to day-ahead load forecasting. The WNN method identifies the k nearest neighbors (kNNs) of day d, where k is a number to be determined and the nearest neighbors in this context are selected according to Euclidean distance. The other model is a meteorological-forecast-based weighted nearest similar day method (MFWNS) [25]. This model is based on the similar day approach, and the nearest similar day set is determined by comparing the Euclidean distance between the meteorological forecast data of one day in the training set and the day to be predicted.

B. HYBRID STLF APPROACHES

One of the most common linear forecasting models is time series analysis, which is a prediction method for finding the causality between independent variables and observed values using previous information. In this study, an algorithm that makes predictions using different filter coefficients for each prediction time is applied in a manner similar to [26]. However, linear prediction models cannot capture the nonlinearity of electricity usage and other factors [27]. Therefore, several nonlinear models such as artificial neural networks [28], Gaussian process regression [29], recurrent neural networks (RNNs) [30] and LSTM [31] have been studied over the past decade to accommodate the nonlinearity of data in a hybrid forecasting model. In a hybrid forecasting structure, linear and nonlinear forecasting models are combined as follows:

$$\mathbf{y}_d = \mathbf{y}_d^{(L)} + \mathbf{y}_d^{(R)} \tag{1}$$

where $\mathbf{y}_d^{(L)}$ and $\mathbf{y}_d^{(R)}$ denote the linear predicted load and its residue load, respectively. The hybrid forecasting process consists of two steps. First, the linear part of data is estimated from a linear combination, and linear coefficient filters can be generated based on the linear minimum mean square error. In addition, as there is a relationship between human activity and electricity usage, it is important to know whether a day is a workday or holiday in load forecasting. To consider this property, the linear prediction model is classified according to day type, $h_d \in \{W, H\}$, denoting a workday (W) or

a holiday (H). h_d represents the day type for the day of the week d, and special holidays such as Thanksgiving and the university anniversary day were considered along with Saturdays and Sundays. Thus, linear day-ahead load forecasting can be implemented as

$$\hat{\mathbf{y}}_d^{(L)} = \hat{\mathbf{A}}_{h_d|h_{d-1}}^T \cdot \mathbf{y}_{d-1} \tag{2}$$

$$\hat{\mathbf{a}}_{t,h_d|h_{d-1}} = (\mathbf{R}_{\mathbf{y}_{d-1},\mathbf{y}_{d-1}|h_{d-1},h_d})^{-1} \cdot \mathbf{R}_{\mathbf{y}_{d-1},y_{t,d}|h_{d-1},h_d} \tag{3}$$

where d is a day index and all hourly column vectors are essentially in 24 dimensions; $\hat{\mathbf{a}}_{t,h_d|h_{d-1}}$, which is the t-th column of $\hat{\mathbf{A}}_{h_d|h_{d-1}}$, represents the linear prediction coefficient filters for a particular time, t, of $\mathbf{y}_d$ from $\mathbf{y}_{d-1}$ representing the previous-day load profile. These linear prediction coefficients can be derived from the cross-correlation vector between $\mathbf{y}_{d-1}$ and $y_{t,d}$, denoted as $\mathbf{R}_{\mathbf{y}_{d-1},y_{t,d}|h_{d-1},h_d}$, divided by the auto-correlation matrix of $\mathbf{y}_{d-1}$, $\mathbf{R}_{\mathbf{y}_{d-1},\mathbf{y}_{d-1}|h_{d-1},h_d}$. In the second step, a nonlinear forecasting model is developed to model the residue of the linear prediction. In general, electricity usage is highly correlated with historical observation and strongly affected by different variables. Thus, LSTM, which adds an input gate, output gate and forget gate to the RNN model and has a memory cell that acts as an accumulator of state information, is suitable for a nonlinear forecasting model in a hybrid structure. By extending the conventional hybrid model [28], estimated linear forecast values are considered as input variables for the nonlinear forecasting model, as suggested in [32]. Multilayer perception is used to model the nonlinear and probable linear relationships that exist in the residue of the linear modeling and original data. The nonlinear data component can be expressed as $\mathbf{y}_d^{(R)} = \mathbf{y}_d - \mathbf{y}_d^{(L)}$, and it can be demonstrated by meaningful variables and linear forecasting values, which are regarded as input variables.

$$\hat{\mathbf{y}}_d^{(R)} = f\left(\mathbf{X}_d, \hat{\mathbf{y}}_d^{(L)}\right) \tag{4}$$

where $f(\cdot)$ represents the nonlinear functions determined by LSTM and $\mathbf{X}_d$ is a matrix with the t-th column of $x_{t,d}$ denoting the selected input variables that strongly affect the residue of the linear prediction.

III. CHARACTERISTIC LOAD DECOMPOSITION (CLD)-BASED HYBRID LOAD FORECASTING

A mixed-use complex with a hierarchical structure contains distinguishable loads. In a complex, sub-loads can be classified according to the characteristics of each load. The concept behind the decomposition-based load forecasting scheme is to decompose an aggregated load and separately predict decomposed loads. Then, all predicted results are summed to predict the aggregated load, as illustrated in Fig. 2.

An aggregated load can be divided into several clusters for practical implementation, as shown in Fig. 3. For instance, the loads of a mixed-use complex can be classified as commercial, residential and industrial load types. Therefore, when a typical load profile is predicted for each cluster, the performance is expected to be superior to that of aggregated load prediction alone as the complexity of the decomposed load pattern can be reduced. In addition, the probing of all nodes is required to accurately decompose the aggregated load, where additional monitoring, storage and communication devices are required for intrusive monitoring. However, the CLD scheme proposed in this study can decompose a load using only partial information; e.g., the representative individual load profile of each characteristic sub-cluster. The effectiveness of the CLD scheme was verified by comparing the results of CLD-based forecasting and all-node-based forecasting; this is detailed in section IV. In addition, the training data was extended based on the similar day approach by constructing the partially captured or missed training data to improve the forecasting performance in the hierarchical structure.

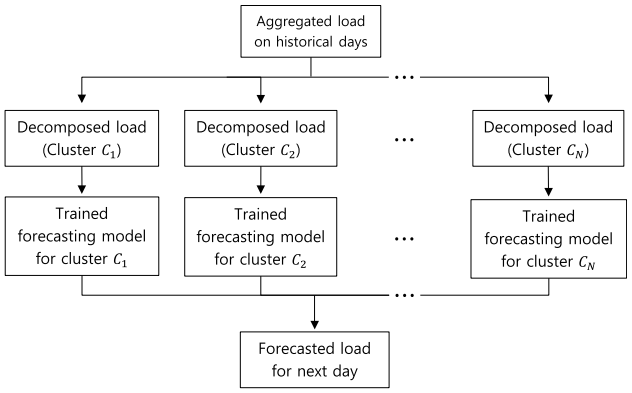


FIGURE 2. Decomposition-based forecasting scheme.

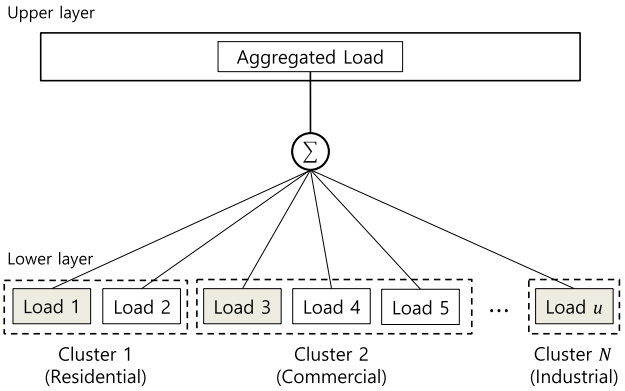


FIGURE 3. Example of hierarchical structure of mixed-use complex. (The colored boxes indicate the representative loads of each cluster).

A. LOAD DECOMPOSITION FOR MIXED-USE COMPLEX

The proposed load decomposition technique is based on an aggregated load, load types, the sub-loads of each cluster and monthly power consumption. Essentially, an aggregated load with a certain complexity can be divided into several clusters according to load characteristics. Thus, the aggregated load, $\mathbf{y}_d$, can be expressed as

$$\mathbf{y}_d = \mathbf{y}_d^{C_1} + \mathbf{y}_d^{C_2} + \cdots + \mathbf{y}_d^{C_N} \tag{5}$$

where $\mathbf{y}_d^{C_n}$ is the sub-load of cluster n on day d. In this study, it is assumed that the total load of clusters can be expressed as the matrix multiplication of the aggregated load and weight factors. Therefore, the proposed CLD can be derived as follows:

$$\mathbf{y}_d = \mathbf{Y}_d \cdot (\mathbf{w}_d^{C_1} + \mathbf{w}_d^{C_2} + \cdots + \mathbf{w}_d^{C_N}) \tag{6}$$

$$\mathbf{w}_d^{C_n} = r^{C_n}(\mathbf{Y}_d)^{-1} \cdot \mathbf{p}_d^{C_n} \tag{7}$$

$$r^{C_n} = \frac{\sum_{i \in C_n} m_i^{C_n}}{m^{C_n}} \tag{8}$$

Here, $\mathbf{Y}_d$ denotes the diagonal matrix of the hourly aggregated load, diag($y_{t,d}$); $y_{t,d}$ is the load at a particular time, t, on day d; $\mathbf{w}_d^{C_n}$ is the hourly weight column vector in cluster n, where $\mathbf{w}_d^{C_1} + \mathbf{w}_d^{C_2} + \cdots + \mathbf{w}_d^{C_N} = \mathbf{1}_{24}$ and $\mathbf{1}_j$ denotes a length j column vector of one; $m_i^{C_n}$ and m^{C_n} are the monthly power consumptions of the i-th sub-load and the representative load in cluster n, respectively; $\mathbf{p}_d^{C_n}$ is the column vector of the representative hourly load in cluster n on day d; N is the number of clusters.

The monthly power consumption and pilot load of each cluster are utilized to obtain the weight column vector; this is the key enabler of CLD. The monthly power consumption can be easily obtained without additional facilities because the monthly power consumption of each load is measured to determine monthly power cost. This is also proportional to the load scale. Therefore, the monthly power consumption is suitable for deriving the relation factors among the sub-loads in a cluster. r^{C_n} in (8) indicates how the total load of cluster n can be expressed by the pilot load of the cluster. As representative pilot load has a typical pattern in cluster n, the load that is expected to consume the most power can be deemed the pilot load. In this study, it is generally assumed that a larger building size corresponds to higher typical power consumption, and the load with the widest gross area in the mixed-use complex is considered as the representative pilot load. In (7), the hourly weight factors are calculated as the ratio of the aggregated load to the pilot load of cluster n. The decomposition result is derived by substituting (7) into (6). The basis of these formulas is the extent to which the load of each cluster occupies the weight of the aggregated load. Then,

$$\hat{\mathbf{y}}_d^{C_n} = \hat{\mathbf{y}}_d^{(L),C_n} + \hat{\mathbf{y}}_d^{(R),C_n} \tag{9}$$

where $n \in \{1, 2, \cdots, N\}$; the forecasted values of the linear and residue components of linear prediction in cluster n are denoted by $\mathbf{y}_d^{(L),C_n}$ and $\mathbf{y}_d^{(R),C_n}$, respectively. As shown in Fig. 2, the linear and residue components of the linear part for each cluster are estimated after decomposing the aggregated load based on CLD. Finally, the forecasted values of the aggregated load are derived by summing both parts.

B. TRAINING DATA EXTENSION FOR CLD

Data errors or shortage can be caused by a number of reasons such as communication errors, hardware defects and absence of equipment. Essentially, predictive technologies are based on historical data and forecasting performance can be inferior if historical data are not sufficient or normal. Therefore, historical data reliability is one of the important factors that influence predictive technology performance. In a hierarchical structure, missing data can be reconstructed using normal data at other levels. In this study, the data missing caused by the absence of sub-load measurement equipment are restored based on a similar day approach. To complete the missing data, the extended weight factors of a sub-load cluster are calculated based on the similar day approach and an extended training data is derived from the aggregated load and extended weight factors. The extended training data is derived from the following equations:

$$\mathbf{w}_{d_e}^{C_n} = \frac{1}{g} \sum_{d_i \in S_{d_e}} \mathbf{w}_{d_i}^{C_n} \tag{10}$$

$$\mathbf{y}_{d_e}^{C_n} = \mathbf{y}_{d_e} \cdot \mathbf{w}_{d_e}^{C_n} \tag{11}$$

where d_e is one day to be extended; S_{d_e} = {set of g days, $d_1, d_2, \cdots, d_g$, closest to d_e} such that D_o is the set of observed days, $d_i \in D_o$, and $h_{d_i} = h_{d_e}$. The similar day set is arranged in accordance with the small order of error between the forecast temperature profiles on d_e and d_i with the same day type. In (10), the extended weight factors on day d_e in cluster n are derived from the weight factors of the g most similar day candidates. Finally, the extended loads in cluster n are derived from the aggregated load and extended weight factors, as in (11).

IV. NUMERICAL EVALUATIONS

For the numerical evaluations conducted in this study, aggregated load data were acquired from the Korea Electric Power Corporation (KEPCO) i-smart system of the Gwangju Institute of Science and Technology (GIST) in South Korea. The loads of 10 buildings at GIST were accumulated from the monitoring systems in each building [33]. The Korea Meteorological Administration (KMA) website [34] provides meteorological forecast data for every city in South Korea. As the published data are organized in different time units, the data are decimated or linearly interpolated on an hourly basis.

A. CHARACTERISTIC LOAD ANALYSIS AND DECOMPOSITION

The load patterns of the GIST complex were analyzed. The profiles of four buildings for one week are plotted in Fig. 4. CA and CB correspond to lecture-hall buildings, while FA and DA are residential buildings. The figure shows that the lecture buildings have peak loads in the daytime, whereas the residences have peak loads in the evening; these profiles correspond to human activity in general. Hence, it is apparent that similar types of buildings have similar load patterns.

Furthermore, based on the cross-correlation analysis between the individual building profiles in the GIST complex, the complex can be categorized into two clusters i.e., residential buildings (FA, DA and DB) and non-residential buildings

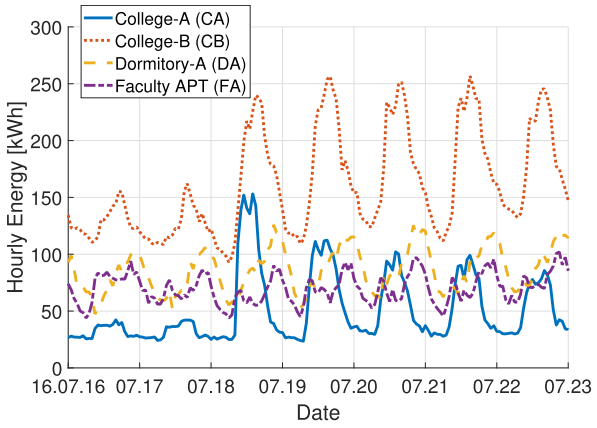


FIGURE 4. Load profiles of four individual buildings for one week in July, 2017.

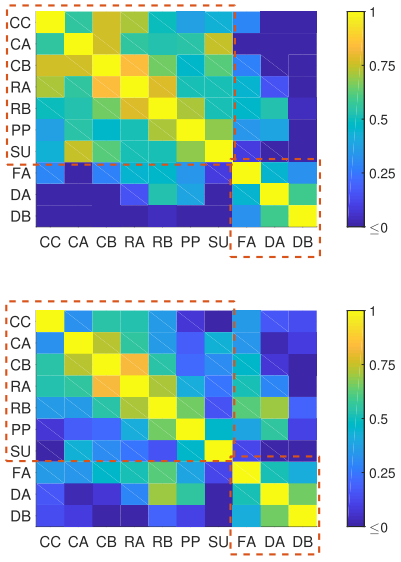


FIGURE 5. Cross-correlation coefficient-based classification on workdays (Top) and holidays (Bottom).

(CC, CA, CB, RA, RB, PP and SU), as illustrated in Fig. 5. Hence, it is determined that the load pattern of each building is similar to the load patterns of the buildings of the same type and that the buildings can be categorized according to their profile patterns.

As the representative load should have a typical pattern in cluster n, the building with the widest gross area at GIST was designated for the pilot load in each cluster. Thus, DA and RA were identified as having the pilot loads in the residential and non-residential clusters, respectively. As explained in previous section, the proposed CLD is derived from the aggregated load, pilot loads and monthly total power consumption. In Fig. 6, the solid lines denote the actual data measured from the 10 individual buildings summed for the two categories and the dotted lines represent the CLD results obtained by the measured data for the aggregated load and two pilot loads only. Even though CLD was applied based

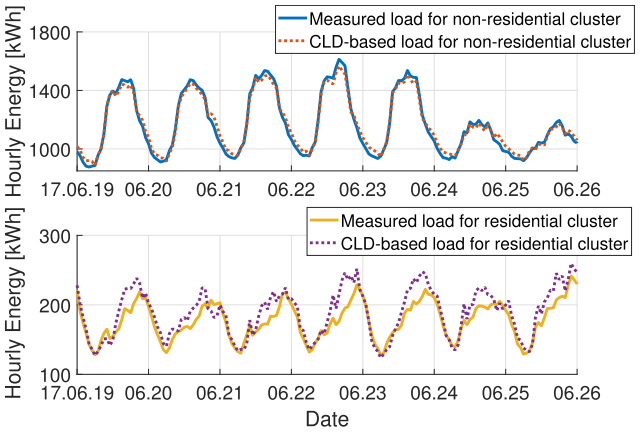


FIGURE 6. Reliability comparison of measured data with pilot-based decomposition results in case of non-residential (Top) and residential load (Bottom) for one week in June, 2017.

on partial information, the results in Fig. 6 indicate that the proposed decomposition scheme is quite reliable.

In the dataset, the aggregated load spans three years from Jan. 2015 to Dec. 2017. However, the load data of 10 buildings for only 1 year and 7.5 months are incorporated due to the absence of measurement equipment. Thus, training and validation data are insufficient for the individual buildings in cast that the test period is set to one year from Jan. 1st to Dec. 31st, 2017. However, the data can be extended based on the similar day approach, as given in (10) and (11). As the test set should not be known beforehand, the extended data were derived from known data for the period from May 15th to Dec. 31st, 2016. The similar day set was arranged based on the temperature profile, and the weight factors of each cluster were derived from the similar day set. The extended decomposition loads were derived from the aggregated load and extended weight factors. Hence, the dependent variables for electricity usage such as temperature, hour indicators, day type and seasonality could be considered. Fig. 7 shows the represented extended data and CLD-based decomposed load for two clusters, i.e., non-residential and residential loads.

B. SELECTION OF MEANINGFUL VARIABLES

Several different environment variables and previous observed data affect electricity usage [35]. Thus, selecting meaningful variables is necessary for precise prediction. In this study, the conditional permutation importance score is applied to assess the importance of a variable in a complicated estimation problem [21]. Table 1 presents the input variable candidates that may affect electricity usage. There are environmental variables, time indicators and the results of linear prediction. In addition, the conditional permutation importance scores of all candidates are presented in Table 1. Note that the importance scores in terms of the linear prediction residue were evaluated as the nonlinear forecasting target of the hybrid model. Among the input candidates, the local temperature forecast (LTF), local humidity

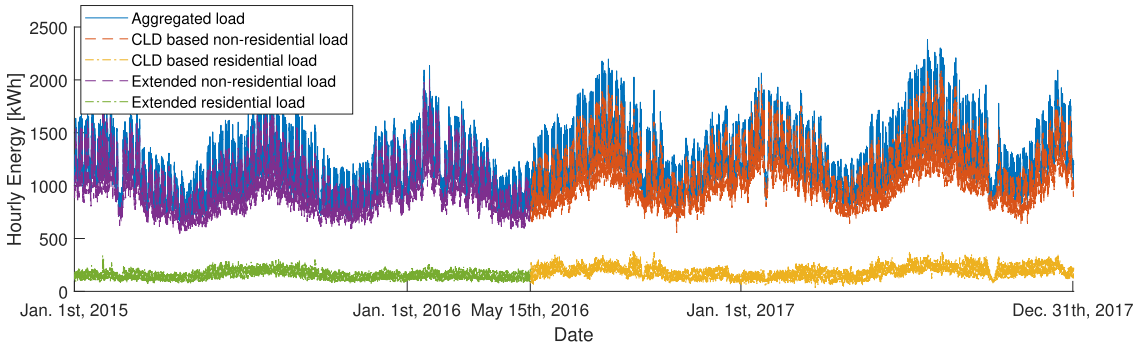


FIGURE 7. Aggregated loads for entire days from Jan. 1st, 2015 to Dec 31st, 2017. CLD-based loads from May 15th, 2016 to Dec. 31st, 2017 and extended load data from Jan. 1st, 2015 to May 15th, 2016.

TABLE 1. Input variable candidates and their importance scores.

Variables	Unit/Index	Importance score for residue part
LTF	°C	38.42
LHF	%	18.31
WS	m/s	5.83
RF	mm	0.33
HOD	Hour	32.50
WHOD	W-W-W: 1, H-W-W: 2, ⋯, H-H-H: 8	14.23
DSMI	Jan. 2015: 1, Feb. 2015: 2, ⋯, Jan. 2016: 13 ⋯,	16.30
LCL	kWh	26.37

forecast (LHF), hour of day (HOD), day type, i.e., workday or holiday (WHOD), distinguishing sequential month index (DSMI) and linear component of load (LCL) were the most informative variables according to the conditional permutation importance score. Correlations exist between the weather variables and load because of the seasonal effect. In addition, the HOD and WHOD are significant variables as electricity usage is strongly dependent on human activity. Unlike the linear prediction model, WHOD is set to $h_{d-1} - h_d - h_{d+1}$, which denotes the day type of yesterday, today and tomorrow, to enable the consideration of more cases in the nonlinear prediction model. The DSMI is an important input variable because the number of facilities and electricity usage generally increase over time. The LCL can be directly related to the residue of linear prediction, as suggested in [32]. These variables were applied as the input of the nonlinear forecasting model.

C. DAY-AHEAD LOAD FORECASTING RESULTS

This subsection reports a comparison of the predictive capabilities of the proposed scheme with other forecasting models in terms of the similar day approach, different amounts of node information and the extended training set using

TABLE 2. Linear forecasting results with different amounts of node information.

	AGG w/ 1 node, 1 cluster	IND w/ 10 nodes, 2 clusters	CLD w/ 3 nodes, 2 clusters
MAPE [%]	4.07	3.89	3.84
(STD) [%]	(1.95)	(1.87)	(1.92)
MAE [kWh]	54.15	51.38	51.36
(STD) [kWh]	(27.38)	(25.95)	(26.63)

GIST load data. To determine the parameters of each forecasting model, it is important to investigate the model performance for varying parameters, which should be found for a known dataset. For evaluation, two forecasting error criteria, i.e., the mean absolute percentage error (MAPE) and mean absolute error (MAE), were used to assess the load forecasting reliability of the forecasting methods. Standard deviation and a mean metric were compared because lower mean and variance of error ensure a more reliable forecasting model. For all models, the method of updating the training set was applied to the accumulated update and the test period was set to 1 year from Jan. 1st to Dec. 31st, 2017.

Table 2 lists the results of the day-type conditioned linear prediction (DTLP) model, which was trained from the middle of May, 2016. Here, AGG, IND and CLD denote the forecasting results of the aggregated load with the measurement of one node, all measured building loads for two clusters, and the characteristic load decomposition with two clusters, respectively. Because the determination of the degree of recent data is an important part of the linear prediction performance, the time lags of DTLP were set as parameters with the minimum error in the validation set. The values ranged from 7 to 12 for each DTLP model. Overall, the forecasting result obtained through decomposition exhibits superior performance, as detailed in Table 2. In addition, even though the prediction obtained with CLD is based on partial sensing information, the performance is similar to that of a technique that utilizes all information. Therefore, it is verified

TABLE 3. Linear and hybrid forecasting results obtained with extended training set.

	WNN	MFWNS	AGG, DTLP	CLD, DTLP	AGG, Hybrid	CLD, Hybrid
MAPE [%]	5.28	5.15	3.84	3.71	3.61	3.38
(STD) [%]	(2.87)	(3.32)	(1.93)	(1.87)	(1.70)	(1.57)
MAE [kWh]	70.26	66.28	52.15	49.65	47.72	44.79
(STD) [kWh]	(37.80)	(38.41)	(26.73)	(26.99)	(22.99)	(21.42)

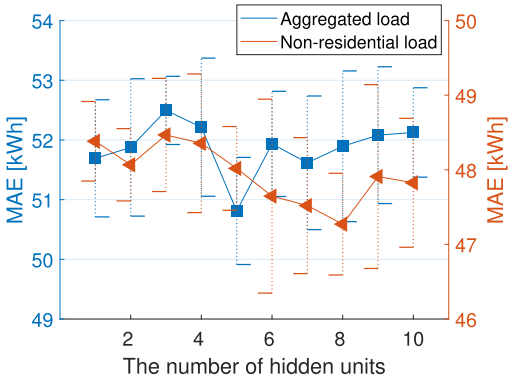


FIGURE 8. Average MAE and 1-sigma in the training set with different numbers of hidden units in case of aggregated and non-residential loads.

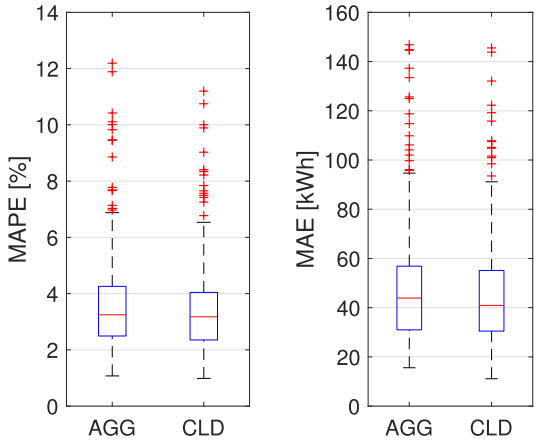


FIGURE 9. Box plots of daily MAPE (Left) and MAE (Right) for aggregated (AGG) and CLD-based (CLD) load forecasting with the test set.

that the proposed CLD is a reliable and efficient method for improving load forecasting performance.

Generally, the load time series exhibits an upward trend phenomenon when power equipment is added or new buildings are constructed. Therefore, in electricity usage data, the trend, which is one of the time series characteristics, could be considered together. However, this upward trend cannot be captured by a similar day scheme because it is derived from the weighted sum of previous similar day loads only. Thus, the bias terms from historical data were manually applied to alleviate the upward trend effect in the similar day approach. The numbers of similar days in WNN and MFWNS were 7 and 6, respectively. The results of two similar-day-based forecasting models are presented in Table 3.

Training data were extended from the synthesized weight factors and aggregated load to alleviate the lack of data, as suggested in previous section. Table 3 presents the forecasting results trained using the extended data. For linear prediction, the filter lengths of the aggregated, non-residential and residential load forecasting in DTLP were set as 12, 12 and 15, respectively. The structure of LSTM is comprised of an input layer, an LSTM layer and a fully connected layer. In addition, as LSTM is sensitive to the scale of data, input and target data were normalized based on the mean and standard deviation of the training set. Standard backpropagation could be applied to train the network using an adaptive moment estimation method known as the Adam optimizer [36]. The number of hidden units with the minimum error in the validation set was found by conducting iteration tests with varying number of hidden units, as illustrated in Fig. 8. The number of hidden units of the aggregated, non-residential and residential loads was set as 5, 8 and 8, respectively. Table 3 proves that the extended training data is reliable and effective, as the results obtained with the extended training data indicate better performance compared to the results obtained without data extension (Table 2). The performance of the proposed decomposition scheme is also improved for DTLP and hybrid models. In addition, the hybrid forecasting performance with the proposed CLD is improved by approximately 6.45% compared with the performance achieved using only the aggregated load (Table 3). The distribution of the iteration test should be checked to demonstrate the reliability of the results because LSTM model is based on the Adam optimizer. Fig. 9 shows the box plot for the forecasting results of the daily error with the test set. It also shows that the forecasting performance when CLD is employed improves as the variation and range of the outlier is reduced. Fig. 10 summarizes the daily forecasting results presented in Tables 2 and 3 in bar plots. It is also necessary to analyze the forecasting results of the peak hours because the reliability of load prediction during the peak hours is important with regard to power coordination. The peak hour is set from 9 am to 8 pm because our dataset consists of residential and non-residential buildings. The prediction errors are 4.17 % (57.94 kWh) and 3.86 % (54.38 kWh) for AGG and CLD respectively. Although the performance of CLD is better than that of AGG, the overall error is increased. It can be seen that the prediction error is large for daytime than for nighttime, as the electricity usage variation is comparatively smaller during nighttime.

D. DECOMPOSITION PERFORMANCE EVALUATION

To practically assess the performance of the proposed CLD, QP-based categorical decomposition (QPD) [10] was implemented and evaluated for comparison. The goal of QPD is to

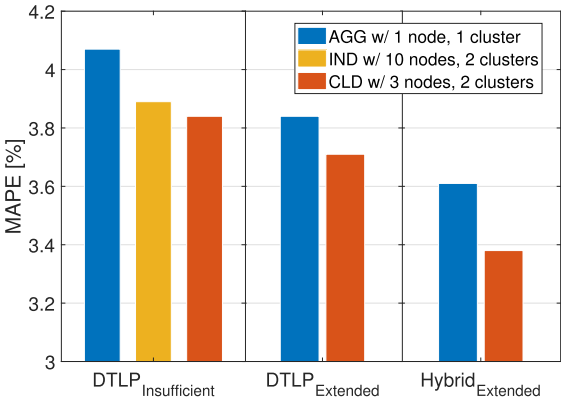


FIGURE 10. Summary of daily error in linear forecasting with relatively insufficient training data in terms of different amounts of node information (Left), linear and hybrid forecasting with extended training set (Middle, Right); AGG (The aggregated load with measurement of one node), IND (All measured building loads for two clusters), CLD (Characteristic load decomposition with two clusters).

TABLE 4. Load forecasting results based on QPD [10] and the proposed CLD.

	QPD, DTLP	CLD, DTLP	QPD, Hybrid	CLD, Hybrid
MAPE [%]	3.80	3.71	3.47	3.38
(STD) [%]	(1.94)	(1.87)	(1.60)	(1.57)
MAE [kWh]	50.43	49.65	45.97	44.79
(STD) [kWh]	(27.13)	(26.99)	(21.83)	(21.42)

find the elementary profiles that are common to each category using an optimization method with time-invariant weights, whereas the CLD scheme derives time-varying weights for each category based on the representative pilot signal through the proposed equation. For the evaluation, load decomposition and prediction results were implemented in equivalent evaluation environments. First, the decomposition performance was compared in terms of the root mean square error from the measured profiles. The QPD yields 55.52 kWh and 46.23 kWh for non-residential and residential loads respectively, whereas the proposed CLD yields 38.16 kWh and 28.02 kWh respectively for the non-residential and residential load pair. As QPD utilizes the proportions of elements and time-invariant weight, the dynamic load properties of each building may not be considered. On the contrary, as the proposed CLD is based on the representative load profiles and thus time-varying weight, the characteristic of the load can be captured according to the building types. Consequently, the linear and hybrid forecasting performance measures of the proposed CLD are better than those of QPD, as presented in Table 4. From the perspective of computational complexity, CLD is comparatively advantageous because QPD is based on quadratic optimization programming whereas CLD uses matrix equations for load decomposition.

V. CONCLUSION

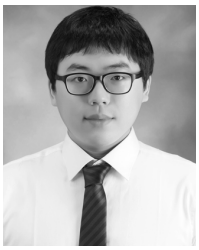
This paper proposes the CLD-based day-ahead forecasting of the aggregated load of a mixed-use complex such as a mixture of residential and commercial buildings. In this approach, complicated power usage patterns can be clustered according to their characteristic profiles and the aggregated load can be decomposed into sub-loads of the clusters. Then, hybrid forecasting models are applied to each sub-load with particular characteristics, and the aggregated load is predicted as the sum of the predicted cluster loads. In addition, in this study, the potential spatial and temporal limitations that may occur as a result of restricted sensing and storage resources in actual applications can be overcome by employing the proposed technique using representative pilot signals and similar-day-based extensions. The proposed schemes were evaluated for the GIST campus load data with a hierarchical structure. Hybrid forecasting that combined CLD-based linear prediction with LSTM exhibited superior performance over conventional approaches in terms of forecasting accuracy. To further assess the effectiveness of our proposed scheme, we compared it with another method that is applied to scenarios similar to that addressed in our study. Considering the expandability of the proposed scheme, it can be used for the short-term load forecasting of various aggregated quantities that involve complex properties. In other words, because CLD is a preprocessing step for predicting a mixed-use complex load, it is applicable to any dataset with a hierarchical structure suitable for partial power measurement.

REFERENCES

[1] *Annual Energy Outlook 2018*, U.S. Energy Inf. Admin., U.S. Dept. Energy, Washington, DC, USA, Feb. 2018. [Online]. Available: https://www.eia.gov/aeo/
[2] M. A. Hannan *et al.*, "A review of Internet of energy based building energy management systems: Issues and recommendations," *IEEE Access*, vol. 6, pp. 38997–39014, 2018.
[3] H. Shareef, M. S. Ahmed, A. Mohamed, and E. A. Hassan, "Review on home energy management system considering demand responses, smart technologies, and intelligent controllers," *IEEE Access*, vol. 6, pp. 24498–24509, 2018.
[4] S. Yoon, K. Park, and E. Hwang, "Connected electric vehicles for flexible vehicle-to-grid (V2G) services," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, 2017, pp. 411–413.
[5] K. Mahmud, M. J. Hossain, and G. E. Town, "Peak-load reduction by coordinated response of photovoltaics, battery storage, and electric vehicles," *IEEE Access*, vol. 6, pp. 29353–29365, 2018.
[6] B. Cornelusse, "How the European day-ahead electricity market works," Inst. Montefiore, Liège, Belgium, Tech. Rep. ELEC0018-1, 2014.
[7] S.-Y. Son, S.-H. Lee, K. Chung, and J. S. Lim, "Feature selection for daily peak load forecasting using a neuro-fuzzy system," *Multimedia Tools Appl.*, vol. 74, no. 7, pp. 2321–2336, 2014.
[8] A. Kaur, L. Nonnenmacher, and C. F. M. Coimbra, "Net load forecasting for high renewable energy penetration grids," *Energy*, vol. 114, pp. 1073–1084, Nov. 2016.
[9] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. zhang, "Energy big data: A survey," *IEEE Access*, vol. 4, pp. 3844–3861, 2016.
[10] A. Gerossier, T. Barbier, and R. Girard, "A novel method for decomposing electricity feeder load into elementary profiles from customer information," *Appl. Energy*, vol. 203, pp. 752–760, Oct. 2017.
[11] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.
[12] Y. Shu and Y. Tang, "Analysis and recommendations for the adaptability of China's power system security and stability relevant standards," *CSEE J. Power Energy Syst.*, vol. 3, no. 4, pp. 334–339, 2017.

[13] S. Fan, K. Methaprayoon, and W. J. Lee, "Multiregion load forecasting for system with large geographical area," *IEEE Trans. Ind. Appl.*, vol. 45, no. 4, pp. 1452–1459, Jul. 2009.
[14] J. Nowotarski, B. Liu, R. Weron, and T. Hong, "Improving short term load forecast accuracy via combining sister forecasts," *Energy*, vol. 98, pp. 40–49, Mar. 2016.
[15] S. B. Taieb, J. W. Taylor, and R. J. Hyndman, "Hierarchical probabilistic forecasting of electricity demand with smart meter data," pp. 1–30, 2017. [Online]. Available: https://robjhyndman.com/papers/HPFelectricity.pdf
[16] L. C. P. Velasco, N. R. Estoperez, R. J. R. Jayson, C. J. T. Sabijon, and V. C. Sayles, "Day-ahead base, intermediate, and peak load forecasting using K-means and artificial neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, pp. 62–67, 2017.
[17] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering," *Appl. Energy*, vol. 231, pp. 331–342, Dec. 2018.
[18] C. Gao, S. Chen, X. Li, J. Huang, and Z. Zhang, "A *physarum*-inspired optimization algorithm for load-shedding problem," *Appl. Soft Comput.*, vol. 61, pp. 239–255, Dec. 2017.
[19] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, pp. 1168–1187, 2017.
[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[21] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinform.*, vol. 9, no. 307, pp. 1–11, 2008.
[22] O. M. Ibrahim, "A comparison of methods for assessing the relative importance of input variables in artificial neural networks," *J. Appl. Sci. Res.*, vol. 9, no. 11, pp. 5692–5700, 2013.
[23] H. D. Tran, N. Muttil, and B. J. C. Perera, "Selection of significant input variables for time series forecasting," *Environ. Model. Softw.*, vol. 64, pp. 156–163, Feb. 2015.
[24] M. E. El-Hawary, *Advances in Electric Power and Energy Systems*. Hoboken, NJ, USA: Wiley, 2017.
[25] T. Hong, "Energy forecasting: Past, present, and future," *Foresight, Int. J. Appl. Forecasting*, no. 32, pp. 43–48, 2014. [Online]. Available: https://ideas.repec.org/a/for/ijafaa/y2014i32p43-48.html#author-abstract
[26] A. E. Clements, A. S. Hurn, and Z. Li, "Forecasting day-ahead electricity load using a multiple equation time series approach," *Eur. J. Oper. Res.*, vol. 251, no. 2, pp. 522–530, 2016.
[27] J. Zheng, C. Xu, Z. Zhang, and X. Li, "Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, 2017, pp. 1–6.
[28] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.
[29] J. Song and E. Hwang, "Hybrid day-ahead load forecasting with atypical residue based Gaussian process regression," in *Proc. 9th ACM Int. Conf. Future Energy Syst. (ACM e-Energy)*, 2018, pp. 631–634.
[30] S.-L. YU and Z. Li, "Stock price prediction based on ARIMA-RNN combined model," in *Proc. 4th Int. Conf. Social Sci. (ICSS)*, 2017, pp. 17–25.
[31] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models," *Expert Syst. Appl.*, vol. 103, pp. 25–37, Aug. 2018.
[32] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2664–2675, Mar. 2011.
[33] *GIST Building Power Consumption Status (Korean)*. Accessed: Jan. 21, 2019. [Online]. Available: http://www.gist.ac.kr/kr/html/sub06/0606.html
[34] *Korea Weather Information (Korean)*. Accessed: Jan. 21, 2019. [Online]. Available: https://data.kma.go.kr/cmmn/main.do
[35] Y. T. Chaea, R. Horeshb, Y. Hwangb, and Y. M. Lee, "Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings," *Energy Builidngs*, vol. 111, pp. 184–194, Jan. 2016.
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

KANGGU PARK received the B.S. degree from the School of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul, South Korea, in 2016, and the M.S. degree from the School of Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2018. His research interest includes information and signal processing, with an emphasis on predictive analysis and energy informatics.



SEUNGWOOK YOON received the B.S. degree from the Department of Electric Engineering, Kwangwoon University, Seoul, South Korea, in 2014. He is currently pursuing the integrated M.S. and Ph.D. degrees with the School of Mechatronics, Gwangju Institute of Science and Technology, Gwangju, South Korea. His research interests include energy informatics, vehicle grid integration, and data channel array signal processing.

EUISEOK HWANG received the B.S. and M.S. degrees from the School of Engineering, Seoul National University, Seoul, South Korea, in 1998 and 2000, respectively, and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2010 and 2011, respectively. He was with the Digital Media Research Center, Daewoo Electronics Co., Ltd., South Korea, from 2000 to 2006, and was with the Channel Architecture Group, LSI Corporation (now Broadcom), San Jose, CA, USA, from 2011 to 2014. Since 2015, he has been an Assistant Professor with the School of Mechatronics, Gwangju Institute of Science and Technology, South Korea. He holds 21 granted U.S. patents. He has over 70 journal and conference papers in information and signal processing issues. His research interests include signal disaggregation, equalization, and coding for data communication channels and emerging large-scale information processing applications, such as smart grids.