ABSTRACT The accurate prediction approach of urban buildings’ electricity consumption is an important foundation for smart urban energy management. It provides a decision basis for reasonable electricity deployments upon different scenarios. Usually, a single model cannot solve linear and nonlinear problems that may occur in electricity consumption prediction effectively and may produce predictions with unsatisfactory accuracy and stability. Moreover, some prediction models are also poorly interpretable and generalized, which makes them difficult to be applied in practice. To overcome these problems, this paper proposes an ensemble prediction model called gravity gated recurrent unit electricity consumption model which integrates the gated recurrent unit model and the proposed logarithmic electricity consumption gravity model. The weights are derived from average mutual information and weighted entropy. We use two years (17,520 hours) electricity consumption of a five-star hotel building in Shanghai, China, as the study case to illustrate our approach, and apply nine common prediction models as the benchmarks to conduct the computational experiments and comparisons. Furthermore, we also employ the electricity consumption data of another type of building (office building) to evaluate the generalization capability of the proposed ensemble model. Our approach outperforms all benchmarks in terms of accuracy, stability, and generalization.

INDEX TERMS Data analysis, prediction algorithms, electricity consumption, ensemble model, information theory.

I. INTRODUCTION

Building energy consumption accounts for 30-45% of global energy consumption and buildings’ electricity consumption is a major part of building energy consumption [1]. According to the latest statistics from the World Bank, the global urbanization population has increased by about 3.9% and the per capita electricity consumption has increased by 1100 kWh (kilowatt hour) in the last decades [2]. The electricity consumption of urban buildings has risen sharply worldwide. Precisely forecasting hourly electricity consumption of urban buildings plays an important role in optimizing the usage of energy in urban buildings and realizing energy-saving operations [3], [4]. It can help electricity supply departments improve their deployment strategies and avoid electricity shortages at peak times [5]. This is necessary to formulate a reasonable urban energy production plan and reduce carbon emissions [6].

There exist some research results of prediction models for building electricity consumption. These studies mainly applied traditional and machine learning models of electricity consumption prediction [7]. The traditional methods include statistical models and time-series models such as widely applied autoregressive integrated moving average model (ARIMA) and regression-based approaches. The above-mentioned traditional models are able to achieve satisfactory results when solving linear problems. Lü et al. [8] proposed a prediction method of building energy demand by introducing random parameters into physical statistical models. A statistical time series model was established to
reflect uncertainty and individual heterogeneity in the underlying buildings, thereby improving the accuracy of the prediction. Shao et al. [9] presented a multivariate statistical and similarity measure based semiparametric modeling of the probability distribution. It identified effectively the pivotal aspects of electricity consumption fluctuation and anticipated the future trends. Kaur et al. collected the time series data from a health care institution Apollo Hospital, Ludhiana for the time period between April 2005 and February 2016. The analyses of time series data and prediction of electricity consumption performed using ARIMA model. The most suitable model for three kinds of time series (monthly, bimonthly, and quarterly) was selected to predict the electricity consumption in a health care institution [10]. Fumo et al. conducted the research on the influence of meteorological factors on residential buildings’ energy consumption and used a quadratic regression analysis approach to predict the demand of buildings’ energy [11].

Recently, some scholars have applied various machine learning techniques to predict electricity consumption. In contrast to traditional methods, they are suitable for non-linear cases. Machine learning models include neural network models, kernel function models such as SVR, deep learning models such as RNN, LSTM and so on. Some machine learning based approaches for predicting electricity consumption of buildings have been developed. Platon et al. compared the accuracy of neural network and Case-Based Reasoning for predicting the electricity consumption of buildings. The results had shown that the prediction errors of both approaches were within ASHRAE limits which stipulated that the error of the electricity consumption prediction was less than 30% but neural networks performed better [12]. Shen et al. applied a variable selection method to choose 18 key features from 48 features, and achieved the predictions based on SVR model [13]. Rahman et al. developed and improved the novel deep recurrent neural network (RNN) models aiming at medium- to long-term electric load prediction with one-hour resolution. And the corresponding performance of the model for different types of electricity consumption patterns were analyzed in [14]. Marino et al. used Long-Short-Term-Memory (LSTM) network to forecast a month-ahead electricity consumption [15].

However, the above-mentioned research faces some challenges. Some machine learning models such as neural networks are suitable for solving nonlinear problems and have high prediction accuracy, but are poorly interpretable and some of them have many parameters to tune. While certain traditional models are highly interpretable and suitable for predictions with linear property [16], [17], they are not suitable for solving nonlinear problems. Due to the complex nature in energy usage, the electricity consumption may vary linearly or nonlinearly with the influencing factors, hence a single method or model may not make predictions effectively [18]. In this case, an ensemble model making use of the advantages of different models is a promising approach to predict building electricity consumption. This paper proposes a novel ensemble-based approach for predicting electricity consumption of urban buildings. This model attempts to integrate machine learning and statistical models to improve the predicting accuracy of building’s electricity consumptions. And from the aspect of the information theory, the construction process of the ensemble model is explained, which makes the ensemble model more interpretable. Furthermore, in order to make the proposed model suitable for real applications, the proposed model pays particular attention to the stability of the predictions and the generalization capability of the model while pursuing the prediction accuracy. The main contributions of the paper are highlighted as follows:

- We propose a logarithmic electricity consumption gravity model (LEGRA) upon gravity model.
- We develop an ensemble prediction model for building’s electricity consumptions called gravity gated recurrent unit electricity consumption model (GRA_GRU) which integrates LERA and gated recurrent unit (GRU) model upon weighted entropy and average mutual information.
- We employ the real data of a hotel building in Shanghai, China to illustrate the forecasting process and validate the model. Furthermore, the real data of another type of building (office building) in Hangzhou, China is employed to evaluate model’s generalization ability. The commonly applied 9 prediction models are employed as the benchmarks to evaluate the proposed model.

This paper is organized as follows. Section 2 reviews the related work. Section 3 describes the process of data preparation. Section 4 presents our prediction approach, where the electricity consumption gravitation model and an ensemble prediction model are proposed. The experiment results and discussions are discussed in Section 5. The paper concludes with some remarks.

II. RELATED WORKS

There are two types of multi-model integration approaches for electricity consumption prediction. One is the method that integrates different models. The other incorporates multiple similar models.

The first type of multi-model integration approach avoids the shortcomings of a single model by exploring the advantages of different models, thereby improving the accuracy of prediction. For example, Dong et al. [19] proposed an integrated data-driven model and physics-based model to forecast the residential hourly electricity consumption. The data-driven model was used to predict the air conditioning electricity consumption, and the non-air-conditioning electricity consumption was predicted by the physical model. The final results were combined from both models, they are able to achieve the air conditioning load and non-air conditioning energy predictions simultaneously. Garshasbi et al. [20] presented a method combining Genetic Algorithm (GA) and Monte Carlo (MC) simulation approach. GA is used to predict the building electricity consumption affected by the common influence factors, and MC is used to predict the random fluctuation electricity consumption caused.

88094 VOLUME 7, 2019
by the building’s own electricity generation and consumption. The method accurately monitors and reports the cumulative energy consumption and production of the cluster of Net Zero Energy Buildings and each individual building within the study area. Barak and Sadegh [21] developed a model called ARIMA-ANFIS to predict Iran’s annual energy consumption. ARIMA is used for linear predictions, while ANFIS is employed for nonlinear predictions in electricity consumption, and finally the results predicted by merging these two models. The method can effectively solve the energy consumption prediction even if the dataset is insufficient. The experimental results disclose that the prediction results of the model are better than the ones obtained by ARIMA or ANFIS model alone. Chen et al. [22] used the ridge regression model to integrate extreme gradient boosting forest and feedforward deep networks to predict household electricity usage. Experiments show that the method reduces the prediction error by 30% comparing to the classical regression model.

The second type of multi-model integration approach is the ensemble method which usually select two or more models for specific cases, and the integrations usually lead to achieve more satisfactory prediction accuracy. For example, Choi and Lee [23] studied an ensemble model upon dynamic adjustment of the weights on multiple LSTMs for forecasting electricity consumption. This ensemble model captures the nonlinear statistical nature of the underlying time series and improves the prediction accuracy. Galicia et al. [24] proposed an ensemble model integrating decision trees, gradient boosted trees, and random forests, whose performance was validated via a series of experiments. Wang et al. [25] developed an ensemble bagging tree model (EBT) to predict institutional building electricity demand. It is proven to be effective for short-term building energy prediction. Divina et al. [26] applied the stacking ensemble learning scheme to generate regression trees to forecast short-term electricity consumption. The accuracy of the ensemble approach is better than the one of the individual base algorithms. Alobaidi et al. [27] employed the multiple multi-layer perceptron (MLP) and the feed-forward artificial neural networks (ANNs) for predicting daily household’s energy consumption. Song et al. [28] developed a multi-resolution selective ensemble extreme learning machine (MRSE-ELM) model for time-series prediction. The proposed approach selects the better trained extreme learning machines (ELMs) from a set of ELMs with different numbers of hidden neurons for integrating. The overall performances of the proposed ensemble models are better than the ones of the corresponding base models. Ensemble models indeed prevail comparing to those underlying base models in terms of achieving accurate and stable forecasting results.

Both types of multi-model integration approaches have their own characteristics. The former is not limited by specific datasets and possesses good generalization capability. The latter can achieve quite accurate results albeit the generalization ability is not so satisfied. Inspired by these observations we choose both approaches to construct an ensemble model with the expectation of achieving solid generalization. Specifically, LE_GRA (statistical model) and GRU models (machine learning model) are assembled to solve the problems with linear and nonlinear properties encountered in the electricity consumption prediction. This model will be discussed in details after the data preparation is explained.

III. DATA PREPARATION

We collect the hourly electricity consumption data of a five-star hotel building in Shanghai, China from September 1st 2013 to August 31st 2015. For a single building, weather is a major factor impacting building electricity consumption because it affects the electricity usage in the building [29], [30]. Therefore, we obtain the weather data for every hour during the same time period including temperature, humidity, and wind speed. The weather datasets are published by the government’s meteorological department, and they are complete and clean. The electricity consumption data of the building is recorded by the sensor. Since the sensor is susceptible to interference, there may be some missing values and outliers in the actual electricity consumption data. Furthermore, the different units used for measuring temperature, humidity, and wind speed could result in large numerical variances, which in turn affects the accuracy of electricity consumption prediction. In order to resolve the above issues and improve the accuracy of electricity consumption prediction, we need to preprocess the data. The procedure of preprocessing is depicted in figure 1. The preprocessing of electricity consumption data includes completion of missing values, outlier detection and replacement. To address the different measuring units in weather datasets, the normalization is employed. The preprocessed datasets are then divided into training, validation, and testing sets for the future use.

A. FILLING MISSING VALUES

In the two-year electricity consumption data, there are 4 days of data with missing values as shown in figure 2(a). Since the electricity usage may be different on working days and
holidays, we select the electricity consumption at similar points in time of the previous week as the substitutes. For example, the electricity consumption of 2 AM on Tuesday, April 29, 2014 is missing. The electricity consumption of 2 AM on Tuesday, April 22, 2014 will be employed to fill the missing value. By applying this method, the missing values in figure 2(a) are filled as depicted in figure 2(b).

B. DETECTING AND REPLACING OUTLIERS

Among the collected electricity consumption data, some of them may be abnormal due to the interference of a sensor. We need to detect outliers and replace them with proper values.

1) DETECTING ABNORMALITY

The box plot in figure 3 displays the maximum, minimum, median, and quartile of the dataset. The upper quartile is called Q3 and the lower quartile is Q1. The interquartile range is called IQR, which is the difference between the two quartiles. Typically, the box plot defines an outlier when a value is less than Q1-1.5IQR or greater than Q3+1.5IQR [31]. Nevertheless, this box plot does not accurately filter out the outliers according to the actual data distribution. Other methodologies besides this box plot to discover the real outliers such as some square points in the plot are necessary. Due to the seasonal impacts, the difference in electricity consumptions is large. We construct a box chart of electricity consumption each month instead of whole time period (figure 3), where 24 box plots for 24-month of electricity consumption over that time period are shown. The scatter points represent possible outliers and some of them possess abnormally large values. The following procedure describes how to detect these outliers.

2) IDENTIFYING AND REPLACING OUTLIERS

Obtaining data distribution may help us to find out the outliers accurately. Figure 4 displays the distribution of the collected electricity consumptions, which is a positive skewed distribution. There are some huge values on the right side of the mode such as 131,831,1975. According to the experience, the electricity consumption per hour of a hotel building will never reach that much and an outlier appear on the right side of the mode at least statistically, an outlier is an observation point that is distant from most observations and there are such points on the right side of the mode (see figure 4).

According to the 3δ principle of normal distribution, 99.7% of the data is located in the interval \((μ−3σ, μ+3σ)\), \(μ\) is the mean (the mean, median and mode are the same in the case of normal distribution), and \(σ\) is the standard deviation. That is, the outliers may account for up to 0.3% of all data and the right side of the mode has been determined to have an outlier at least (i.e. 131, 831, 1975). In order to avoid a data point being detected as an outlier mistakenly, the proportion of the outlier is set to 0.15% of the total data points and an outlier may appear on the right side of the mode in our study.

EllipticEnvelope model is an abnormality detection model based on Gaussian distribution [32]. It assumes that a dataset follows Gaussian distribution and tries to define the “ellipse shape” of the data. The outliers can be then defined as
FIGURE 4. The distribution of electricity consumptions and it is the positive skewed distribution.

FIGURE 5. 23 outliers detected using the EllipticEnvelope model. It can be seen that the values of all detected abnormal electricity consumptions are large, and they deviate from the normal data a lot.

the observations standing far enough from the fitted elliptic shape. In other words, it fits an ellipse to the central data points while ignoring those outside of the central portion. According to our above discussions, we set the EllipticEnvelope model outlier ratio to 0.15% for identifying outliers. The scatter plot of the hourly electricity consumptions for each month over 2 years is presented in figure 5, and the square dots on the left are the detected outliers. We detect in total 23 outliers by applying the EllipticEnvelope model. We find that the detected outliers are far from most data points and this is consistent with our consensus. We remove all 23 outliers and repair the dataset by filling the missing (removed) values as described in part A of section 3.

C. NORMALIZING DATA

Since weather data includes temperature, wind speed and humidity, their measuring units are different and their values are quite different. This will cause the influence factors to be significantly different in predictions. That is, the influence factors with smaller values may play insignificant role in the prediction while the influence factors with larger values impact the prediction at the greater degree. To overcome this obstacle, we normalize the underlying dataset using equation (1) so that the resultant values will range between 0 and 1. Normalization also contributes to the prediction accuracy of the model [33]. In Equation (1), $X$ is the data value that needs to be normalized, $V_{\text{min}}$ is the minimum value in the dataset, $V_{\text{max}}$ is the maximum value of the dataset, and $X_{\text{norm}}$ is the normalized value of $X$.

$$X_{\text{norm}} = \frac{X - V_{\text{min}}}{V_{\text{max}} - V_{\text{min}}} \quad (1)$$

D. PARTITIONING DATASETS

We divide the 24-month of data into training, validation, and testing datasets. The short-term prediction period of building electricity consumption is from 1 to 6 months [34]. In order to fully evaluate the predictive capability of a model, we set various test datasets to fit the short-term prediction. As shown in table 1, data for 1-, 2-, 3-, 4-, 5- and 6-month are used as the testing sets to evaluate the short-term prediction. We set the size of the validation dataset to be the same as the testing one. In table 1, the proportion refers to the ratio of the size of a training/validating/testing dataset to one of the whole data. For example, the proportion of a 2-month dataset is 2/24.

IV. PREDICTION APPROACH

In order to improve the accuracy of predicting electricity consumptions and to address the linear and nonlinear natures occurring in the prediction, we propose an ensemble model integrating statistical model and machine learning model via the information entropy-based weighting method. As shown in figure 6, the logarithmic electricity consumption gravity model named LE_GRA is developed based on the gravity model. LE_GRA is utilized as the statistical base model in the ensemble. The GRU model is applied as the machine learning base model in the ensemble. After LE_GRA and GRU models are trained independently, the average mutual information and weighted entropy are employed to determine the weights of LE_GRA and GRU models upon the validation set. Finally, an ensemble model called GRA_GRU that produces the final prediction results by weighing prediction results yielded by LE_GRA and GRU.

A. LE_GRA MODEL

Since the gravitational model is suitable for analyzing the flow changes between two different locations and has good generalization, it is widely used in various areas such as immigration, transportation, tourism, etc., even though its parameters and variables need to be appropriately tuned [35]–[37]. We can revise a gravity model such as a trade gravity model for the purpose of predicting electricity consumptions.

$$y_{ij} = a_0 x_i^a x_j^b Z_{ij}^c \quad (2)$$
TABLE 1. The dataset partitioning of short-term prediction.

| Experiment ID | Training data (months) | Proportion of training data | Validating data (months) | Proportion of validating data | Testing data (months) | Proportion of testing data |
|---------------|------------------------|------------------------------|--------------------------|------------------------------|-----------------------|---------------------------|
| A             | 22                     | 92%                          | 1                        | 4%                           | 1                     | 4%                        |
| B             | 20                     | 84%                          | 2                        | 8%                           | 2                     | 8%                        |
| C             | 18                     | 75%                          | 3                        | 12.5%                        | 3                     | 12.5%                     |
| D             | 16                     | 66%                          | 4                        | 17%                          | 4                     | 17%                       |
| E             | 14                     | 58%                          | 5                        | 21%                          | 5                     | 21%                       |
| F             | 12                     | 50%                          | 6                        | 25%                          | 6                     | 25%                       |

FIGURE 6. The process of electricity consumption prediction using our approach. The GRA, GRU model produces the final predictions by weighing the results obtained by LE_GRA and GRU models.

Equation (2) is a trade gravity model used to predict the amount of imports between two regions [38]. The parameters in equation (2) are:

- $y$: the import quantity,
- $x$: influence factors that could impact trade such as GDP,
- $Z$: a special event factor such as the number of trading days in two regions or the number of days of tariff increase,
- $d$: the distance between the two regions,
- $a$ and $b$: coefficients.

The electricity consumption of a building is affected by many factors, and the electricity consumption over a period of time can be derived from the difference in meter readings at two points in time. If different points in time are treated as different “locations” in time, and the increase or decrease of electricity consumption is viewed as a type of flow change, then above mentioned difference (in meter readings) reflects the flow change between different locations. Therefore, electricity consumption can be determined by the gravity model (equation (2)) with proper modifications. The electricity consumption gravitational model is presented in equation (3) and the associated parameters are:

- $E$: the electricity consumption,
- $x$: influence factors affecting the electricity consumption such as temperature,
- $Z$: a special event factor such as the number of days for the building renovation and/or major activities in the building etc.,
- $n$: the number of factors
- $m$: the number of special events,
- $a$, $b$, $c$: coefficients
- $d$: the duration between two points in time (hours).

$$E_{ij} = a_0 + x_1^{d_{ij}^{e_1}}x_2^{d_{ij}^{e_2}}...x_n^{d_{ij}^{e_n}}Z_1^{b_1}Z_2^{b_2}...Z_m^{b_m}$$

(3)

The electricity consumption represented by equation (3) will vary with the influencing factors, and the proportion of these changes may be different. If the dispersion of $E$ gradually increases (decrease) with the increase (decrease) of $x$, then the collected data is heteroscedastic [39]. In order to detect the presence of heteroscedasticity in the collected dataset, we extract a subset including 744 hours of temperature and electricity consumption data for 31 days in August 2015. To facilitate the observation, we sort the 744-hour of temperatures ascendingly and the electricity consumption data is reordered accordingly. A scatter plot can be generated as shown in figure 7 after the sorting procedure.

Upon figure 7 it is observed that as the temperature rises from 22 to $30\, ^\circ C$, the dispersion of electricity consumption gradually increases. When the temperature continues rising (from $33\, ^\circ C$ and beyond), the dispersion of electricity...
consumption decreases rapidly. This fully demonstrates the collected data is heteroscedastic. The heteroscedasticity of data leads to a deterioration in prediction accuracy. To address this issue, the logarithmic transformation is one of the effective means to reduce the heteroscedasticity [40]. Therefore, in order to mitigate the heteroscedasticity effect in the dataset, we take the logarithm on both sides of equation (3) and obtain the logarithmic electricity consumption gravity model called LE_GRA, as shown in equation (4). The meanings of the parameters in equation (4) are the same as those in equation (3). As described above, the influencing factors in our study are temperature, humidity, and wind speed. There is no influencing factor for any special event. Substituting these relevant influencing factors into equation (4) we will be obtaining equation (5). In equation (5), \( x_t \), \( x_h \) and \( x_s \) are temperature, humidity, and the wind speed respectively; \( \gamma \) is a constant.

\[
\begin{align*}
\ln E_{ij} &= \ln a_0 + \sum_{j=1}^{n} a_k \ln x_k + \sum_{j=1}^{m} b_k \ln z_k + c \ln \frac{1}{d_{ij}} \quad (4) \\
\ln E_{ij} &= \alpha_t \ln x_t + a_h \ln x_h + a_s \ln x_s + \gamma, \\
(\gamma &= \ln a_0 + c \ln \frac{1}{d_{ij}}) \quad (5)
\end{align*}
\]

B. GRU MODEL

The GRU is a type of recurrent neural networks (RNN) with a complex gate structure shown in figure 8 [41]. Each neuron in the current hidden layer needs to process two inputs including one of the current neural unit \( x_t \) and the previous state information \( h_{t-1} \), \( x_t \) as well as \( h_{t-1} \) constitute the candidate output state information \( \tilde{h}_t \) of the current neural unit via \( \text{Tanh} \). The reset rate \( r_t \) (between 0 and 1) is applied to control the amount of \( h_{t-1} \) fed to \( \tilde{h}_t \) and \( \tilde{h}_t \) compose the current state information \( h_t \) by updating the gate \( z_t, \tilde{z}_t \) is used to control the amount of \( h_{t-1} \) brought into \( h_t \). The product of \( h_t \) and the weight is fed to the sigmoid function \( \sigma \) to obtain the output \( y_t \) of the neuron in the current hidden layer.

The loss function presented in (6) is the mean squared logarithmic error (MSLE), where \( p_t \) is the predicted value, \( a_i \) is a real value, and \( n \) is the number of samples in the whole dataset.

\[
\text{MSLE} = \frac{1}{n} \sum_{i=1}^{n} (\log (p_i + 1) - \log(a_i+1))^2 \quad (6)
\]

C. ENSEMBLE MODEL AND WEIGHTS

In this section, we are going to propose a novel ensemble prediction model called GRA_GRU that integrates LE_GRA and GRU models. Equation (7) is the mathematical expression of GRA_GRU model, \( V \) is the predicted electricity consumption, and \( w \) is the weight.

\[
V_{\text{GRA-GRU}} = w_1 V_{\text{LE-GRA}} + w_2 V_{\text{GRU}} \quad (7)
\]

In the information theory, entropy quantifies the amount of uncertainty involved an information source [42]. In equation (7), LE_GRA and GRU models are two sources of information, and their information is actually the prediction accuracy (e.g. 1%, 2%, \ldots, 100%). Based on the entropy definition, in our study the higher the accuracy of an information source, the smaller entropy. However, entropy is symmetric, that is, the poor accuracy of the information source also leads to small entropy. To address this issue, the weighted entropy is applied to deduct the importance of the information source with lower accuracy. In addition, the average mutual information measures how much knowing one of these information sources reduces uncertainty about the other. Specifically, there are two information sources in our study, and thus their average mutual information is the same. It is conceivable that if the ratio of the average mutual information of one information source to its weighted entropy is greater, then the corresponding source has more accurate information. In this case, the associated weight of this source in the ensemble model should be greater considering the ensemble model needs to integrate more accurate information from different models.

Based on the above discussion, the weights are determined as follows. We first use the training dataset to complete the training for both LE_GRA and GRU models, then apply the trained LE_GRA and GRU models to conduct forecast using the validation dataset separately. The accuracies of the outcomes obtained by these two models are calculated based on equation (8). In equation (8), \( a_{ij} \) is the accuracy of predicting the \( j^{th} \) validation data using the \( i^{th} \) model (i.e., the \( j^{th} \) information source). \( R_t \) represents the real electricity consumption of the \( j^{th} \) validation data. \( F_{ij} \) indicates the electricity consumption predicted by the \( i^{th} \) model for the \( j^{th} \) validation data.

\[
a_{ij} = 100\% (1 - \left| \frac{R_j - F_{ij}}{R_t} \right|) \quad (8)
\]

For \( m \) items of validation data, the \( j^{th} \) model will produce \( m \) corresponding accuracy values. If the column vector \( A_j^T = (a_{j1}, a_{j2}, \ldots, a_{jm})^T \) is used to represent the accuracy of the \( j^{th} \) model, then the accuracy values of all models
can be represented by matrix $A_{mn}$ as shown in equation (9).

$$A_{mn} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad (i \leq m, j \leq n) \quad (9)$$

We count the number of occurrences of each accuracy value in matrix $A_{mn}$ (note: an accuracy value is a real number, therefore we only consider its integer part such as 87% instead of 87.15% while counting the number of occurrence of an accuracy value) and obtain $R_{mn}$ illustrated in equation (10), where $r_{ij}$ represents the number of occurrences of $a_{ij}$ (its integer part) in the $j^{th}$ column.

$$R_{mn} = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix}, \quad (i \leq m, j \leq n) \quad (10)$$

Each element in matrix $R_{mn}$ will be fed to equation (11) to obtain the weighted information entropy of LE_GRA and GRU models respectively. As discussed above, the weighted entropy is applied to make the information sources with higher-accuracy more important in the ensemble model. And the ASHRAE standard stipulates that the error of the electricity consumption prediction was less than 30%. Therefore, when a source has more than 70% of accuracy, its weighted entropy should be smaller. We can calculate the weighted entropy upon equation (11). In equation (11), $E_j$ is the weighted information entropy of the $j^{th}$ model (i.e., the $j^{th}$ information source), $w'_{ij} \log p_{ij}$ is the weight corresponding to $p_{ij} \log p_{ij}$, $N_j$ represents the number of $a_{ij}$ greater than 70% in the $j^{th}$ column of matrix $A_{mn}$, $p_{ij}$ is the probability of occurrence of $r_{ij}$ where $M$ indicates the sum of $r_{ij}$ on the $j^{th}$ column.

$$E_j = -\sum_{i=1}^{m} w'_{ij} \log p_{ij}, \quad (11)$$

where,

$$w'_{ij} = \begin{cases} 1, & a_{ij} < 70\% \\ 1 - \frac{N_j}{\sum_{i=1}^{m} N_j}, & a_{ij} \geq 70\%, \end{cases}$$

$$p_{ij} = \frac{r_{ij}}{M}, \quad \text{where} \quad M = \sum_{i=1}^{m} r_{ij}$$

The process of computing the average mutual information used for determining the weights in the ensemble model is follows: define a unit step function $UN(a_k, a_l) = 1$ if $a_k = a_l$, and 0 otherwise, where $a_k$ and $a_l$ ($k \neq l$) are two rows in matrix $A_{mn}$. We define $c_i = 1 + \sum_{j=1}^{m} UN(a_k, a_l) \quad (1 \leq i \leq m)$. The vector $c_i^T = (c_1, c_2, \ldots, c_m)^T$ can be then obtained. Equation (12) is applied to obtain the average mutual information of two information sources, in which $J$ and $J'$ represent two information sources respectively in our case.

$$I(J; J') = \sum_{y \in J'} \sum_{x \in J} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}, \quad (12)$$

where,

$$P(x, y) = \frac{c_i}{\sum_{i=1}^{m} c_i}, \quad P(x) = \frac{r_{ij}^{(J)}}{\sum_{i=1}^{m} r_{ij}^{(J)}}, \quad P(y) = \frac{r_{ij}^{(J')}}{\sum_{i=1}^{m} r_{ij}^{(J')}}$$

According to our previous discussion, the weight of each information source is the ratio of the average mutual information to its entropy, which is named as RAMIE weights determination method. The weight of the $j^{th}$ information source will be calculated by equation (13). In equation (13), $I$ is the average mutual information while $E_j$ represents the weighted entropy of the $j^{th}$ source, $Z$ is the parameter used to normalize the weights of all base algorithms to ensure the sum of all weights to be 1.

$$w_j = \frac{I}{ZE_j} \quad (13)$$

The pseudocode in figure 9 describes the prediction process of GRA_GRU model.

V. RESULTS AND DISCUSSION

In order to evaluate the qualities of the prediction results, we utilize MAPE (Mean Absolute Percentage Error) [43] illustrated by equation (14), where $A_t$ is the actual value and $F_t$ is the forecast value.

$$M = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t} \quad (14)$$

In the computational experiments, we employ 9 commonly applied forecasting models (shown in table 2) as the benchmarks. Some applications of these models can be found in [44], [45].

They represent 7 different types of models including generalized linear regression, support vector machine, nearest neighbor, gaussian process, decision tree, ensemble method, neural network, time series models and Long Short-Term Memory. In this case the proposed GRA_GRU model can be evaluated comprehensively. The datasets presented in table 1 will be used to conduct the experiments.

The benchmarks listed in table 2 are trained by employing the training dataset and the testing datasets are applied to obtain the prediction results. The proposed GRA_GRU model needs to use the training datasets in table 1 to train the embedded LE_GRA and GRU models. The validation dataset is applied to determine the weights of LE_GRA and GRU models.

The testing dataset is employed to run GRA_GRU model to perform the predictions. Furthermore, we use some practical experience and grid search method [46], [47] to tune all parameters of the benchmarks and GRA_GRU so that their predictions are best used for reasonable comparisons.

In order to comprehensively evaluate the predicting performance of the proposed model GRA_GRU, we compare the prediction results from five different aspects:

- comparing GRA_GRU to 9 benchmarks’ prediction results,
Algorithm 1. The prediction process using GRA_GRU model

Input:
\[ D_o \] — training dataset
\[ D_v = (E_v, T_v, H_v, S_v) \] — validation dataset, \( m \) items of data in \( D_v \)
\[ D_t = (E_t, T_t, H_t, S_t) \] — testing dataset, \( n \) items of data in \( D_t \)
\[ E_v, T_v, H_v, S_v \] — Electricity consumption, temperature, humidity, wind speed in validation dataset
\[ E_t, T_t, H_t, S_t \] — Electricity consumption, temperature, humidity, wind speed in testing dataset

GRA_GRU — GRA_GRU model

Output:
\[ E_v \] — The \( k \)-th prediction results using GRA_GRU model on testing dataset

1. LE_GRA' — train LE_GRA upon \( D_o \)
2. GRU' — train GRU upon \( D_v \)
3. For \( i = 0 \ldots n \)
4. \[ E_{LE, GRA} \] — input \( T_v, H_v, S_v \) to LE_GRA'
5. \[ E_{GRA, GRU} \] — input \( E_{LE, GRA} \) and \( E_v \) to Accuracy function
6. \[ A_{GRA} \] — return
7. \[ E_{GRU} \] — input \( T_v, H_v, S_v \) to GRU'
8. \[ E_{GRU} \] — input \( E_v \) and \( E_v \) to Accuracy function
9. \[ A_{GRU} \] — return
10. For \( i = 0 \ldots n \)
11. \[ p_{LE, GRA} = \frac{N(A_{LE, GRA})}{\sum_{i=1}^{N} (A_{i, GRA})} \]
12. \[ p_{GRU} = \frac{N(A_{GRU})}{\sum_{i=1}^{N} (A_{i, GRA})} \]
13. If \( A_{LE, GRA} < 70\% \) or \( A_{GRU} < 70\% \)
14. \[ W_{GRU} \] — input
15. else
16. \[ W_{GRU} = \frac{1 - N(A_{GRU})}{N(A_{LE, GRA}) + N(A_{GRU})} \]
17. \[ W_{LE, GRA} = \frac{1 - N(A_{LE, GRA})}{N(A_{GRU}) + N(A_{LE, GRA})} \]
18. \[ Entropy = W_{LE, GRA} \times \log p_{LE, GRA} \]
19. \[ Entropy = W_{GRU} \times \log p_{GRU} \]
20. Return Entropy_LE_GRA = sum(Entropy)
21. Return Entropy_GRU = sum(Entropy)
22. For \( i = 0 \ldots n \)
23. \[ p_{LE, GRA} = \frac{N(A_{LE, GRA})}{\sum_{i=1}^{N} (A_{i, GRA})} \]
24. \[ Log = \log \frac{p_{LE, GRA}}{p_{GRU}} \]
25. \[ I = \sum \log \frac{p_{LE, GRA}}{p_{GRU}} \]
26. \[ W_{LE, GRA} = \frac{1}{W_{GRU}} \]
27. \[ W_{GRU} = \frac{1}{W_{LE, GRA}} \]
28. For \( k = 0 \ldots n \)
29. \[ Input T_v, H_v, S_v \] to LE_GRA'
30. \[ Return E^{LE, GRA} \]
31. \[ Input T_v, H_v, S_v \] to GRU'
32. \[ Return E^{GRU} \]
33. \[ Output: E_v = \sum (W_{LE, GRA} \times E^{LE, GRA}, W_{GRU} \times E^{GRU}) \]

The pseudocode of the prediction process using GRA_GRU model.

- comparing GRA_GRU to its base models’ prediction results,
- comparing optimization methods such as SGD to RAMIE method (our method) from the aspect of determining weights,
- motivation of building GRA_GRU model
- examining the generalization capability of GRA_GRU model using the dataset collected from another building.

A. COMPARING GRA_GRU TO 9 BENCHMARKS

1) OBSERVATIONS OF THE PREDICTION RESULTS

Figure 10 shows the MAPE for each model’s prediction results in various experiments (dataset divisions of various experiments are shown in table 1). This helps us to observe the generalization capabilities of benchmarks and the proposed model. The MAPEs of the predicted results of all models (including GRA_GRU) are oscillating in all experiments (A to F) (i.e. the time periods of testing data range are from 1- to 6-month). In the experiment B (i.e. the time period of testing data is 2-month), the prediction results of these models are the worst comparing to other experiments except ARIMA. However, the predicted results of GRA_GRU appear to be less fluctuant than the predictions of 9 benchmarks in various experiments.

We use a heat map to depict the MAPE of each model in different experiments as shown in Figure 11. In Figure 11, we use warm colors to identify three best predictions for each experiment and the warmer the color, the better the accuracy (the smaller the MAPE). Conversely, we used the cool color to identify three worst predictions in each experiment and the colder the color, the worse the accuracy (the larger the MAPE). White represents neither the best nor the worst. We can conclude that the greater color difference of the heat map of a model, the worse the stability of the model, which will make it difficult if not impossible to apply in practice. For example, the predictions of DTR and proposed GRA_GRU are top three in 4 out of the 6 experiments. However, the predictions of DTR in 2 out of 6 experiments were among the worst three. Therefore, although DTR model

| Table 2. 9 benchmarks’ names and types. |
|----------------------------------------|
| Type of model | Full name of model | Abbreviation of model |
|---------------|-------------------|----------------------|
| Generalized Linear Regression | Ordinary Least Squares Regression | OLSR |
| | Bayesian Ridge Regression | BRB |
| | Stochastic Gradient Descent Regression | SGDR |
| Support Vector Machine | Support Vector Regression | SVR |
| Nearest Neighbor | Nearest Neighbors Regression | NNR |
| Decision Tree | Decision Tree Regression | DTR |
| Neural network model | Multi-layer Perceptron Regression | MLP |
| Time series model | Autoregressive Integrated Moving Average model | ARIMA |
| Recurrent Neural Network | Long Short-Term Memory | LSTM |
performs well in certain experiments, it is not stable and the quality of its prediction results varies dramatically. Similarly, the prediction quality of LSTM model fluctuates and some of its predictions are poor. In contrast, the proposed GRA_GRU model is capable of achieving the predictions with more reasonable accuracy and stability. To further illustrate the advantages of GRA_GRU model, we are going to apply the metrics discussed below to evaluate GRA_GRU model.

2) METRICS OF PREDICTION RESULTS
We use four metrics including maximum MAPE, minimum MAPE, average MAPE and average variance to evaluate the predictions of each model in different experiments. Figure 12 presents the metrics of prediction results obtained by 9 benchmarks and GRA_GRU. In terms of the minimum MAPE, GRA_GRU model is ranked fourth among all models. However, the proposed GRA_GRU model performs the best among all models considering the maximum MAPE and average MAPE. The maximum MAPE of GRA_GRU is 3.76% lower than the second one and the average MAPE of GRA_GRU is 1.8% lower than the second one. This demonstrates that the proposed GRA_GRU model has overall better accuracy.

We use average variance to evaluate the stability of the prediction results for all models. Table 3 shows the average variances of the prediction results of 9 benchmarks and GRA_GRU model across 6 experiments. The GRA_GRU performs the best in terms of average variance, which indicates that GRA_GRU is able to produce reasonable results relatively consistently.

Taking the maximum MAPE, average MAPE, and average variance into account, the proposed GRA_GRU model...
outperforms 9 benchmarks. In general, the proposed model has better accuracy and stability.

**B. COMPARING GRA_GRU TO ITS BASE MODELS**

GRA_GRU integrates LE_GRA and GRU models. Similar to the experiments in part A of section 5, we conduct a series of computational experiments to observe their performances in various predictions.

As shown in Figure 13, in terms of MAPE, the proposed GRA_GRU model performs better than LE_GRA model in experiments A, C, E and F. It performs better than GRU in experiments B, D and F. The proposed approach dominates both LE_GRA and GRU model in experiment F. In other words, GRA_GRU model outperforms LE_GRA model in 4 out of 6 experiments and GRU model in 3 out of 6 experiments. Regarding the average MAPE, GRA_GRU outperforms both its base models.

Table 4 presents the average variances of the prediction results of GRA_GRU and its base models. With regard to the average variance across 6 experiments, GRA_GRU dominates. The computational experiments reveal that the proposed model outperforms its base models concerning the capability of achieving reasonable results with consistency in general.

In summary, the above computational experiments demonstrate that the proposed model performs the best with regard to the average MAPE and variance of the predicted results comparing to its base models and 9 benchmarks. The accuracy and stability of the proposed model for the short-term prediction of building electricity consumption are validated. GRA_GRU model is expected to be applicable to various short-term electricity consumption predictions.

**C. COMPARING DIFFERENT APPROACHES FOR DETERMINING WEIGHTS**

Certain optimization algorithms such as SGD (Stochastic gradient descent) etc. are often used to determine weights in machine learning models. However, in addition to being black box approaches, they are easily trapped at local optimum [48]. It is difficult if not impossible to obtain the optimal weights, which leads to a compromise in the prediction accuracy of the underlying ensemble model. We use SGD algorithm to determine the weights in equation (7) where two loss functions including MSLE and MSE (mean-square error) are employed. The results obtained by this method will be compared to the ones yielded by GRA_GRU model whose weights are determined by RAMIE (our approach).

In figure 14, GRA_GRU indicates that the ensemble model uses RAMIE to determine the weights whilst SGD_mse represents that SGD with MSE loss function is applied to determine the weights for the ensemble model. SGD_msle indicates that SGD with MSLE loss function is employed to determine the weights. The MAPE of the predicted results of GRA_GRU in 4 out of 6 experiments are better than SGD_mse and SGD_msle. The average MAPE of GRA_GRU also is the lowest. As shown in table 5,
TABLE 5. The comparison of predictions’ variance upon different weight determination methods.

|               | GRA_GRU  | SGD_mse  | SGD_mse  |
|---------------|----------|----------|----------|
| Experiment A  | 0.0086   | 0.0119   | 0.0111   |
| Experiment B  | 0.0118   | 0.0155   | 0.0155   |
| Experiment C  | 0.0184   | 0.0166   | 0.0171   |
| Experiment D  | 0.0171   | 0.0224   | 0.0224   |
| Experiment E  | 0.0166   | 0.0170   | 0.0170   |
| Experiment F  | 0.0159   | 0.0160   | 0.0160   |
| average       | 0.0147   | 0.0166   | 0.0165   |

GRA_GRU performs the best in terms of prediction’s variance. The outcome demonstrates that the information theory method (our method) indeed helps in obtaining the ideal weights.

D. ENSEMBLING LE_GRA AND GRU

An ensemble model improves the predictive performance of a single model by training multiple models and fusing their predictions. Due to its superiority in the generalization capability and prediction capability, we are motivated to combine LE_GRA and GRU to build an ensemble model that follows two key principles [49] including diversity and predictive performance.

1) DIVERSITY

Diversity means that the base models of combination are different. Therefore, we conduct statistical tests of the hypothesis that the electricity consumption’s prediction results of LE_GRA and GRU are statistically significant. Since the distribution of the prediction results obtained by the algorithms is unknown, a non-parametrical test should be used. Similar to Ablanedo-Rosas and Rego [50], we use the Wilcoxon rank-sum statistic test to determine whether two selected samples have the same distribution. Particularly, the null hypothesis assumes that the populations of prediction results obtained by both LE_GRA and GRU are identical.

The open source scipy software is used to conduct the test by calling stats.ranksums, which is provided with the results from all 6 experimental results. The pvalue of this test is 2.505e-306 (< 0.01). Therefore, we reject the null hypothesis and accept the alternative that LE_GRA and GRU are different.

2) PREDICTIVE PERFORMANCE

As described in section 2, LE_GRA model is applied to solve linear problems. As shown in figure 10, and comparing to other models presented in table 2 including OLSR, BRR and SGDR that are able to solve linear problems relatively well, LE_GRA performs best in 6 experiments. GRU model is employed to address nonlinear issues. Based on the benchmarks, GRU over-performs the neural network models including MLP and LSTM in 6 experiments. The results inspire us to ensemble the best performing models capable of handling different types of problems. With regard to the average accuracy and variance across 6 experiments the ensemble model GRA_GRU indeed performs the best comparing to other 9 benchmarks and its base models.

E. GENERALIZATION CAPABILITY OF GRA_GRU

One of the key metrics for evaluating the building electricity consumption prediction models is the generalization capability. It is not enough that a model works perfectly in predicting the electricity consumption prediction for just one type of buildings. Ideally it should be applicable to the electricity consumption predictions of different types of buildings. Therefore, in order to evaluate the generalization capability of the proposed GRA_GRU model, we use another type of building in another city to conduct the experiments. The dataset contains one-year electricity consumptions of an office building in Hangzhou, China. These data elements were recorded manually and accurately without any outliers. We conduct 3 experiments using 1-, 2-, 3-months of consumption data as testing data.

As shown in Figure 15 and table 6, with regard to the average accuracy and variance GRA_GRU model performs the best comparing to other 9 benchmarks and its base models. It confirms further the outstanding performance of the proposed GRA_GRU model.

VI. CONCLUSION

This paper presents a study attempting to predict building electricity consumptions effectively. The linear and nonlinear
issues need to be addressed in the prediction of electricity consumption, so a single model may not deliver satisfactory predictions. To solve this problem, we adapt two outstanding base models including LE_GRA and GRU to solve linear and nonlinear problems, and ensemble them to form GRA GRU. To avoid the local optima in setting the weights for the ensemble model, we use average mutual information and weighted entropy to determine the weights of LE_GRA and GRU in the ensemble model, which not only addresses the problem of local optimal solution but also has better interpretability and makes the model with higher-accuracy more important.

The experimental results reveal some interesting facts. Due to the complexity of electricity usage behavior in buildings, none of the prediction models included in the study could dominate in all experiments. Even some state-of-art models such as LSTM did not perform overwhelmingly in most experiments. Therefore, proposing a model that performs better than most models in general might be an effective way to predict building electricity consumption in practice. In addition, it is not sufficient to judge the ability of a model merely based upon the prediction accuracy. It would be more reasonable to evaluate a model based on overall accuracy and variance of all prediction outcomes. It is crucial for a model to possess good generalization capability so that it can be applied to different types of building electricity consumption predictions in real applications. GRA GRU model proposed in this study is a new prediction model for building electricity consumption. In spite of its relative simplicity, it is more interpretable. Furthermore, the comprehensive computational experiments demonstrate that GRA GRU model delivers the prediction results with better average accuracy and variance for different types of buildings in different cities comparing well-known 9 benchmarks and its base models. This indicates that GRA GRU model possesses impressive prediction and generalization capabilities. In summary, GRA GRU model is more suitable for practical applications, especially for short-term electricity consumption predictions.

In the future study we are planning to collect more datasets to evaluate the capability of the proposed model for long-term predictions and electricity consumptions of other buildings than hotel ones. Finer granularity data such as electricity consumptions every 10 minutes could be interesting to investigate, because it may help us further improving the prediction accuracy. More potential influence factors other than three factors employed in the paper will be studied with the aim of enhancing the model and improving the prediction accuracy.

REFERENCES

[1] D. A. Asimakopoulos, M. Santamouris, I. Farrou, M. Laskari, M. Salari, G. Zanis, G. Giannakidis, K. Tiganis, J. Kapsomenakis, C. Douvis, S. C. Zeeceos, T. Antonakaki, and G. Giannakopoulos, “Modelling the energy demand projection of the building sector in Greece in the 21st century,” Energy Buildings, vol. 49, pp. 488–498, Jun. 2012.

[2] Indicators. (2018). The World Bank. [Online]. Available: http://data.worldbank.org/indicator

[3] Z. Shao, F. Chao, S. L. Yang, and K. L. Zhou, “A review of the decomposition methodology for extracting and identifying the fluctuation characteristics in electricity demand forecasting,” Renew. Sustain. Energy Rev., vol. 75, pp. 123–136, Aug. 2017.

[4] J. K. Gruber, M. Prodanovic, and R. Alonso, “Estimation and analysis of building energy demand and supply costs,” Energy Procedia, vol. 83, pp. 216–225, Dec. 2015.

[5] J. A. Dirks, W. J. Gorissen, J. H. Hathaway, D. C. Skorski, M. J. Scott, T. C. Pulsipher, M. Huang, Y. Liu, and J. S. Rice, “Impacts of climate change on energy consumption and peak demand in buildings: A detailed regional approach,” Energy, vol. 79, pp. 20–32, Jun. 2015.

[6] R. Pino-Mejias, A. Perez-Fargallo, C. Rubio-Bellido, and J. A. Pulido-Arcas, “Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions,” Energy, vol. 118, pp. 24–36, Jan. 2017.

[7] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gomez-Romero, and M. J. Martin-Bautista, “Data science for building energy management: review,” Renew. Sustain. Energy Rev., vol. 70, pp. 598–609, Apr. 2017.

[8] X. Lu, T. Lu, C. J. Kibert, and M. Viljanen, “Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach,” Appl. Energy, vol. 144, pp. 261–275, Apr. 2015.

[9] Z. Shao, F. Gao, Q. Zhang, and S.-L. Yang, “Multivariate statistical and similarity measure based semiparametric modeling of the probability distribution: A novel approach to the case study of mid-long term electricity consumption forecasting in China,” Appl. Energy, vol. 156, pp. 502–518, Oct. 2015.

[10] H. Kaur and S. Ahuja, “Time series analysis and prediction of electricity consumption of health care institution using ARIMA model,” in Proc. 6th Int. Conf. Soft, Singapore, Apr. 2017, pp. 347–358.

[11] N. Fumo and M. A. R. Biswas, “Regression analysis for prediction of residential energy consumption,” Renew. Sustain. Energy Rev., vol. 47, pp. 332–343, Jul. 2015.

[12] R. Platon, V. R. Dehkordi, and J. Martel, “Hourly prediction of a building’s electricity consumption using case-based reasoning, artificial neural networks and principal component analysis,” Energy Buildings, vol. 92, pp. 10–18, Apr. 2015.

[13] M. Shen, H. Sun, and Y. Lu, “Household electricity consumption prediction under multiple behavioural intervention strategies using support vector regression,” Energy Procedia, vol. 142, pp. 2734–2739, Dec. 2017.

[14] A. Rahman, V. Srikumar, and A. D. Smith, “Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks,” Appl. Energy, vol. 212, pp. 372–385, Feb. 2018.

[15] D. L. Marino, K. Amarasinghe, and M. Manic, “Building energy load forecasting using deep neural networks,” in Proc. 42nd Annu. Conf. Ind. Electron. Soc. (IECON), Florence, Italy, Oct. 2016, pp. 7046–7051.

[16] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” SIAM Rev., vol. 60, no. 2, pp. 223–311, 2018.

[17] M. A. M. Daut, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah, and F. Hussin, “Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review,” Renew. Sustain. Energy Rev., vol. 70, pp. 1108–1118, Apr. 2017.

[18] Z. Wang and R. S. Srinivasan, “A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models,” Renew. Sustain. Energy Rev., vol. 75, pp. 796–808, Aug. 2017.

[19] B. Dong, Z. Li, S. M. M. Rahman, and R. Vega, “A hybrid model approach for forecasting future residential electricity consumption,” Energy Buildings, vol. 117, pp. 341–351, Apr. 2016.

[20] S. Garshabi, I. Kurnitski, and Y. Mohammadi, “A hybrid Genetic Algorithm and Monte Carlo simulation approach to predict hourly energy consumption and generation by a cluster of net zero energy buildings,” Appl. energy, vol. 179, pp. 626–637, Oct. 2016.

[21] S. Barak and S. S. Sadegh, “Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm,” Int. J. Electr. Power Energy Syst., vol. 82, pp. 92–104, Nov. 2016.

[22] K. Chen, J. Jiang, F. Zheng, and K. Chen, “A novel data-driven approach for residential electricity consumption prediction based on ensemble learning,” Energy, vol. 150, pp. 49–60, May 2018.

[23] J. Y. Choi and B. Lee, “Combining LSTM network ensemble via adaptive weighted learning for improved time series forecasting,” Math. Problems Eng., vol. 2018, Aug. 2018, Art. no. 2470171.

[24] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, and F. Martinez-Álvarez, “Multi-step forecasting for big data time series based on ensemble learning,” Knowl-Based Syst., vol. 163, pp. 830–841, Jan. 2019.
[25] Z. Wang, Y. Wang, and R. S. Srinivasan, “A novel ensemble learning approach to support building energy use prediction,” Energy Buildings, vol. 159, pp. 109–122, Jan. 2018.

[26] F. Divina, A. Gilson, F. Gomez-Vela, M. G. Torres, and J. F. Torres, “Stacking ensemble learning for short-term electricity consumption forecasting,” Energies, vol. 11, no. 4, p. 949, Apr. 2018.

[27] M. H. Alobaidi, F. Chebana, and M. A. Meguid, “Robust ensemble learning framework for day-ahead forecasting of household based energy consumption,” Appl. Energy, vol. 212, pp. 997–1012, Feb. 2018.

[28] H. Song, A. K. Qin, and F. D. Salim, “Multi-resolution selective ensemble extreme learning machine for electricity consumption prediction,” in Proc. Int. Conf. Neural Inf. Process., Siem Reap, Cambodia, Oct. 2017, pp. 600–609.

[29] S. Jovanović, S. Savić, M. Bojić, Z. Djordjević, and D. Nikolić, “The impact of the mean daily air temperature change on electricity consumption,” Energy, vol. 88, pp. 604–609, Aug. 2015.

[30] A. Prabakar, L. Wu, L. Zwanepol, N. Van Velzen, and D. Djiairam, “Applying machine learning to study the relationship between electricity consumption and weather variables using open data,” in Proc. IEEE PES Innov. Smart Grid Technol. Conf. (ISGT), Sarajevo, Bosnia-Herzegovina, Oct. 2018, pp. 1–6.

[31] H. Hofmann, H. Wickham, and K. Kafadar, “Letter-value plots: Boxplots for large data,” J. Comput. Graph. Statist., vol. 26, no. 3, pp. 469–477, 2017.

[32] Practical Data Science, Apress, Berkeley, CA, USA, 2018, pp. 527–684.

[33] D. Pfeffer, Data Mining and Statistics for Decision Making. Hoboken, NJ, USA: Wiley, 2011.

[34] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” PloS ONE, vol. 13, no. 3, Mar. 2018, Art. no. e0194889.

[35] R. Ramos and J. Suriñach, “A gravity model of migration between the ENC and the EU,” Tijdschrift Voor Economische en Sociale Geografie, vol. 108, no. 1, pp. 21–35, Feb. 2017.

[36] Z. Wang, Y. Wang, and R. S. Srinivasan, “A novel ensemble learning framework for day-ahead forecasting of household based energy consumption,” Energy, vol. 212, pp. 997–1012, Feb. 2018.

[37] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1724–1734.

[38] R. M. Gray, Entropy and Information Theory, Springer Science & Business Media, 2011.

[39] R. L. Moy, L.-S. Chen, and L. J. Kao, “Multiple linear regression,” Study Guide for Statistics for Business and Financial Economics, Cham, Switzerland: Springer, 2015, pp. 223–240.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine learning in Python,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1724–1734.

[41] O. Sagi and L. Rokach, “Ensemble learning: A survey,” Wiley Interdiscip. Rev., Data Mining Knowl. Discovery, vol. 8, no. 4, Jul. 2018, Art. no. e1249.

[42] J. H. Ablanedo-Rosas and C. Rego, “Surrogate constraint normalization for the set covering problem,” Eur. J. Oper. Res., vol. 205, no. 3, pp. 540–551, Sep. 2010.

SHUBING SHAN received the bachelor’s degree in computer science from Northeastern University, Shenyang, China, in 2006, and the master’s degree from the School of Software Engineering, University of Science and Technology of China, Hefei, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Software Engineering, Tongji University, Shanghai, China.

From 2011 to 2014, he was an R&D Manager with the Shanghai Software Center, Shanghai, China. In 2017, he studied data mining with The Australian National University as a Visiting Doctoral Student. His research interests include data mining, machine learning, energy computing, and urban computing.

BUYANG CAO received the bachelor’s and master’s degrees from the University of Shanghai for Science and Technology, Shanghai, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of the Federal Armed Forces Hamburg, Hamburg, Germany, in 1994.

From 1994 to 2008, he was an R&D Manager with ESRI, Inc., Redlands, CA, USA. Since 2009, he has been a Professor with the School of Software Engineering, Tongji University, China, also with the College of Architecture and Urban Planning, Tongji University. His research interests include machine learning, data mining, urban computing, energy computing, GIS, and operations research. He received the Franz Edelman Finalist Award from INFORMS twice, in 1998 and 2002.

ZHIQIANG WU received the bachelor’s and master’s degrees from Tongji University, Shanghai, China, in 1982 and 1985, respectively, and the Ph.D. degree from Technische Universität Berlin, Germany, in 1994.

Since 1996, he has been a Professor with the College of Architecture and Urban Planning, Tongji University, China, where he has been the Vice President, since 2011. He is also an Academician of the Chinese Academy of Engineering, an Academician of the Royal Swedish Academy of Engineering Sciences (IVA), and an Academician of the German Academy of Science and Engineering (acatech). His research interests include city planning, smart urban planning, urban computing, and artificial intelligence.

* * *