energies

MDPI

Article

A Regression-Based Method for Monthly Electric Load Forecasting in South Korea

Geun-Cheol Lee

College of Business Administration, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea; gclee@konkuk.ac.kr

Abstract: In this study, we propose a regression-based method for forecasting monthly electricity consumption in South Korea. The regression model incorporates key external variables such as weather conditions, calendar data, and industrial activity to capture the major factors influencing electricity demand. These predictor variables were identified through comprehensive data analysis. Comparative experiments were conducted with various existing methods, including univariate time series models and machine learning techniques like Holt–Winters, LightGBM, and Long Short-Term Memory (LSTM). Additionally, ensemble methods combining two or more of these existing methods were tested. In the empirical analysis, the proposed model was used to forecast monthly electricity demand for a 24-month period (2022–2023), achieving a mean absolute percentage error (MAPE) of approximately 2%. The results demonstrated that the proposed method consistently outperforms all benchmarks tested in this study.

Keywords: mid-term load forecasting; regression; interaction effects; machine learning

Citation: Lee, G.-C. A Regression-Based Method for Monthly Electric Load Forecasting in South Korea. *Energies* **2024**, *17*, 5860. https://doi.org/10.3390/en17235860

Academic Editors: Riccardo Berta, Matteo Nardello and Luca Lazzaroni

Received: 25 October 2024
Revised: 18 November 2024
Accepted: 20 November 2024
Published: 22 November 2024

1. Introduction

In the past several decades, South Korea has seen remarkable economic growth. According to the recent World Bank report, one of the key drivers of this success has been Korea's strategic investments in infrastructure [1], particularly in the reliable supply of electricity, which has supported the competitiveness of its manufacturing sector. Korea ranks first among high-income OECD (Organisation for Economic Co-operation and Development) countries in the "Getting Electricity" category of the World Bank's Doing Business Index [2], emphasizing the importance of stable power infrastructure to overall national competitiveness. Effective and efficient management of the power system is essential for maintaining this competitive edge, with the first critical decision being accurate load forecasting. Accurate electricity demand forecasting offers multiple benefits, such as reducing investment costs, enabling more efficient scheduling of power plant development, and improving the planning of distribution and transmission grids [3]. In this context, this study proposes a method for load forecasting in South Korea. We focus on mid-term load forecasting, using a regression-based method that incorporates key external variables influencing electricity demand.

Load forecasting can be categorized into short-term, mid-term, and long-term, each serving distinct purposes in the management and planning of power systems. Short-term load forecasting (STLF), which typically spans hours to weeks, is essential for operations such as balancing electricity supply and demand, scheduling unit commitments, and managing the grid's stability [3,4]. Mid-term load forecasting (MTLF), which covers periods from several months to a year, is used for optimizing maintenance schedules, planning fuel supply, and preparing for seasonal peaks in demand [4]. Finally, long-term load forecasting (LTLF) is necessary for strategic decisions regarding investments in infrastructure, such as building new power plants and expanding transmission networks [4]. Given the various applications of these forecasting categories, this study focuses on MTLF in South Korea and proposes a method to forecast the monthly electricity consumption of the country.

The literature on load forecasting is extensive, with several survey papers having been published on the topic. Among the recent surveys, the works by Wang et al. [3] and Kuster et al. [4] are introduced. According to Kuster et al.'s 2017 survey, which provided a comprehensive review of electrical load forecasting models, research on MTLF was underrepresented compared to studies on STLF and LTLF [4]. However, by 2022, the increasing importance of MTLF had become evident, as seen in the survey by Wang et al., which focused specifically on MTLF [3]. This growing trend reflects the expanding recognition of the important role that MTLF plays in enhancing overall power system productivity.

Several recent studies have contributed to the advancement of MTLF, and most of them utilized sophisticated machine learning and hybrid models to improve accuracy. Rubasinghe et al. [5] employed a CNN (Convolutional Neural Network)-LSTM hybrid model to forecast monthly peak loads over three years using data from New South Wales, Australia. Jain and Gupta [6] compared various machine learning approaches, including LSTM, RNN (Recurrent Neural Network), SVM (Support Vector Machine), and deep learning models, and found that LSTM outperformed others in forecasting demand for a 12-month period in Chandigarh, India. Li et al. [7] introduced a mid-to-long-term forecasting model using an improved sparrow search algorithm (ISSA) combined with SVM, applying it to five years of monthly load data in China. Jung et al. [8] proposed a deep neural network (DNN) model with transfer learning for monthly load forecasting in Seoul, Korea. Liu et al. [9] developed a hybrid model combining ensemble empirical mode decomposition (EEMD) and Random Forest to forecast China's power consumption over six months, outperforming traditional methods. These studies share a common focus on leveraging nonlinear modeling techniques and hybrid approaches, demonstrating the efficacy of advanced machine learning models in the field of MTLF.

Despite the extensive research on load forecasting in South Korea, the majority of studies are published in Korean, and most of them focus on STLF. In this paper, we introduce several recent studies published in English, contributing valuable insights to enhance understanding of this field. Lee [10] applied regression analysis for daily load forecasting and provided a comprehensive review of various STLF methods in Korea. Lee and Cho [11] employed a hybrid SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors)-LSTM model for daily forecasts, demonstrating superior performance. Baek [12] extended the forecast horizon to mid-term (four weeks) using daily data, with the RANN (Robust Adaptive Neural Network) model achieved the best results. Ryu et al. [13] focused on day-ahead forecasts using 24 h load data, with DNN outperforming other models. Except for Lee's study, machine learning-based approaches consistently outperformed traditional methods. Notably, weather data, calendar data, and historical load data are common predictors across these studies. These factors will also be prioritized in this study on South Korea's load forecasting.

In summary, while many studies pointed out the importance of MTLF and its role, research on load forecasting in South Korea remains primarily focused on STLF. In this study, we focus on MTLF in South Korea, considering a forecasting horizon of two years. To accomplish this, we first identify key factors that influence electricity demand. Based on these findings, we propose a regression model that incorporates selected independent variables representing these influential factors. We will then demonstrate the superiority of the proposed model through comparative computational experiments with various existing methods. Consequently, this study proposes the most suitable method for MTLF in South Korea.

The remainder of this paper is organized as follows. The next section covers the data analysis, presenting a basic time series analysis of the electricity demand and introducing various external factors that could influence the demand. Section 3 presents the regression model constructed using the independent variables selected through the data analysis. Section 4 provides an empirical analysis, where monthly electricity demand for the years 2022 and 2023 is predicted to validate the performance of the proposed model through

comparative experiments. Finally, the last section concludes the paper by summarizing key findings, discussing implications, and suggesting directions for future research.

2. Data Analysis

In this section, we perform a comprehensive analysis of monthly electricity demand in South Korea. First, we identify the key characteristics of the electricity demand time series and then proceed to analyze various external factors that may influence this demand. For this analysis, we use the monthly electricity consumption data for South Korea, which can be obtained from the IEA (International Energy Agency) website (https://www.iea.org/data-and-statistics/data-tools/monthly-electricity-statistics, accessed on 29 September 2024).

2.1. Electricity Demand

Figure 1 shows the time series of monthly electricity demand from 2012 to 2021, providing a visual representation of the overall trend and seasonal variations in South Korea's electricity consumption.

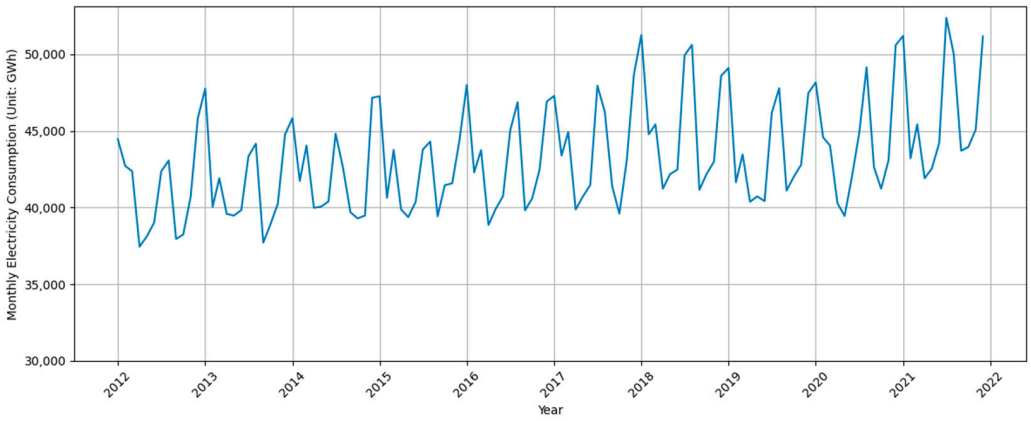

Figure 1. Monthly electricity consumption series of South Korea from 2012 to 2021.

As can be seen from the figure, the series exhibits both a clear upward trend and seasonal fluctuations. The demand typically peaks during the summer months, and a secondary peak is visible during the winter months. Over the observed period, the overall demand for electricity has steadily increased, consistent with South Korea's economic growth and industrial development. Notable drops in consumption in 2020 could correspond to the COVID-19 pandemic. From these observations, we recognize the importance of considering both trend and seasonal components in forecasting models, which will be further explored in the following sections. Among the several characteristics of the series, the next figure is prepared to highlight the seasonality of electricity demand.

Figure 2 illustrates the average monthly electricity consumption in South Korea, showing the distinct seasonality in demand. Notably, electricity consumption peaks during the summer months of July and August, as well as the winter months of December and January, reflecting increased usage for cooling and heating, respectively. The month of February shows lower electricity consumption compared to the surrounding months, which can be attributed to the fact that February has fewer days than other months, resulting in reduced overall consumption. The following graph further confirms the presence of seasonality.

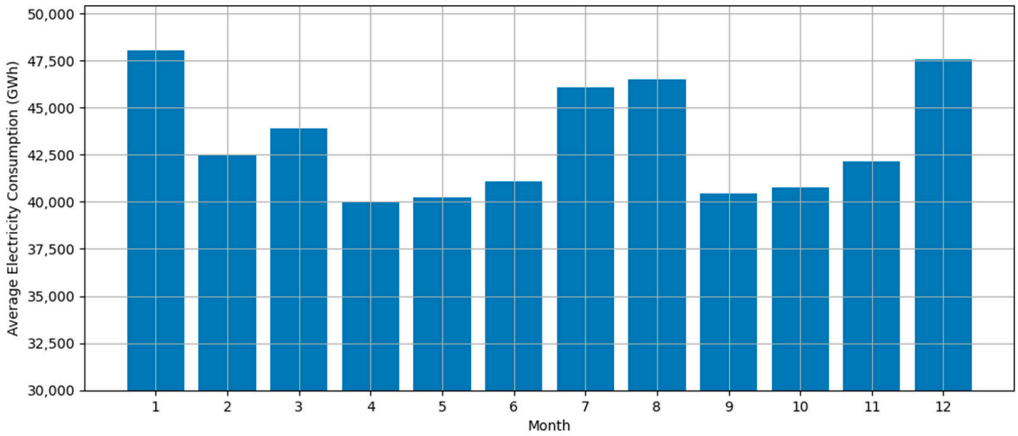

Figure 2. Average monthly electricity consumption in South Korea from 2012 to 2022.

In Figure 3, the ACF plot clearly confirms the presence of seasonality with a 12-month frequency corresponding to a yearly cycle. In the PACF plot, a prominent spike is observed at lag 12, which is larger than the spikes at any other lag. These aspects indicate that the electricity demand from the same month in the previous year is highly correlated with the current month's demand. It is particularly notable that the correlation with electricity demand from one year ago is much stronger than that with the demand from the immediate previous month. This observation suggests that incorporating the electricity consumption from the same month in the previous year as an explanatory variable in the forecasting model would be beneficial.

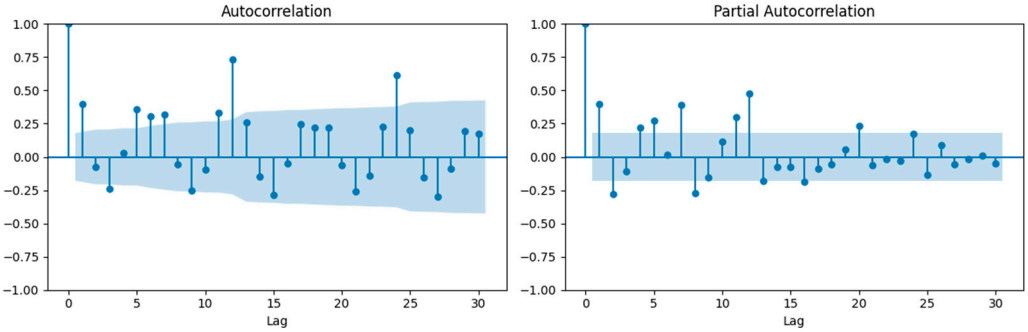

Figure 3. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of monthly electricity consumption.

2.2. Weather Data

Weather information, including temperature, has been an essential component in all the studies forecasting electricity demand in South Korea [10–13], consistently being incorporated into prediction models. Among various weather variables, the following figure presents the relationship between temperature and electricity demand. In this study, all the weather data are collected from the Korean Meteorological Administration Weather Data Service website (https://data.kma.go.kr/, accessed on 5 October 2024).

In Figure 4, the blue circles represent the scatter between monthly electricity consumption and average temperature of the corresponding month, while the red stars show the relationship between monthly electricity consumption and the average highest temperature of the corresponding month. From visual information, it is clear that both temperature variables exhibit a non-linear relationship with electricity demand. As temperatures rise during the summer months, there is an increase in electricity consumption due to cooling needs. Conversely, lower temperatures in the winter correspond to increased electricity demand for heating. These patterns confirm the significant influence of temperature on electricity consumption in South Korea. Another weather variable recognized in this study is the

monthly total solar radiation, and the relationship between this variable and electricity consumption is presented in the following figure.

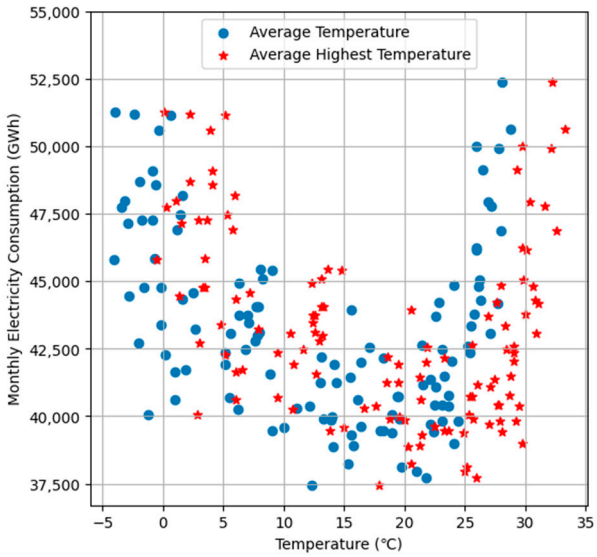

Figure 4. Relationship between monthly electricity consumption and two temperature data, i.e., average temperature and average highest temperature.

Figure 5 shows the scatter plot illustrating the relationship between monthly electricity consumption and total solar radiation of the corresponding month. Unlike temperature, solar radiation does not show a straightforward linear relationship with electricity consumption, making it harder to identify a direct correlation. In the plot, the data for July and August are displayed with red squares, showing a generally positive correlation between solar radiation and electricity consumption. In contrast, the data for the other months, represented by blue circles, reveal a roughly non-linear inverse relationship between the two variables. This suggests that in the forecasting model, total solar radiation should be considered with adjustments according to the month, as its impact on electricity demand varies between peak summer and other seasons.

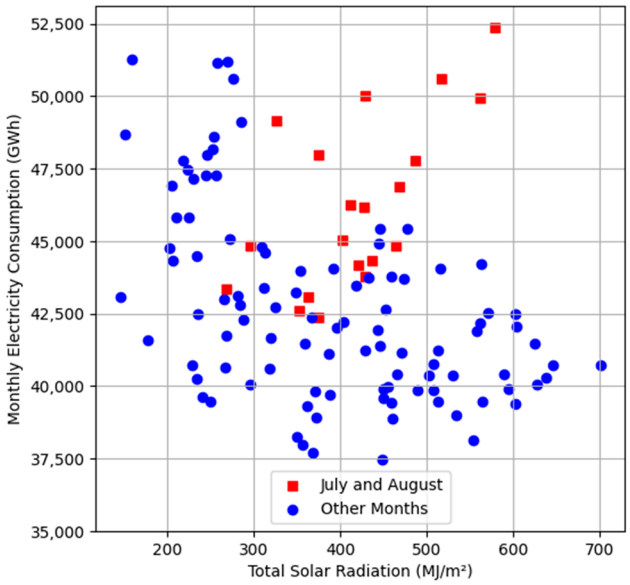

Figure 5. Relationship between monthly electricity consumption and total solar radiation.

2.3. Calendar Data

Among the introduced studies on monthly electricity demand forecasting earlier, it is noteworthy that Jung et al. [8] is the only study to utilize calendar data. In their study, variables such as the number of days in the month, the number of weekdays, the number of weekends, and the number of holidays on each weekday were incorporated. Considering the significant difference in electricity consumption between working and non-working days in South Korea, incorporating such information into forecasting models is reasonable. The next figure is a chart that compares monthly electricity consumption between months with relatively many non-working days and those with not many non-working days.

The calendar factor that influences electricity consumption is the number of non-working days, including holidays and weekends, in a given month. Figure 6 compares the average monthly electricity consumption based on the number of non-working days. The left bar represents months with less than 10 non-working days, while the right bar corresponds to months with at least 10 non-working days. The chart shows that electricity consumption tends to be slightly higher in months with fewer non-working days. This observation suggests that non-working days, when industrial and commercial activities are reduced, lead to a decrease in overall electricity consumption compared to months with more working days. Therefore, it is necessary to include calendar information related to non-working days as a predictor in the forecasting model.

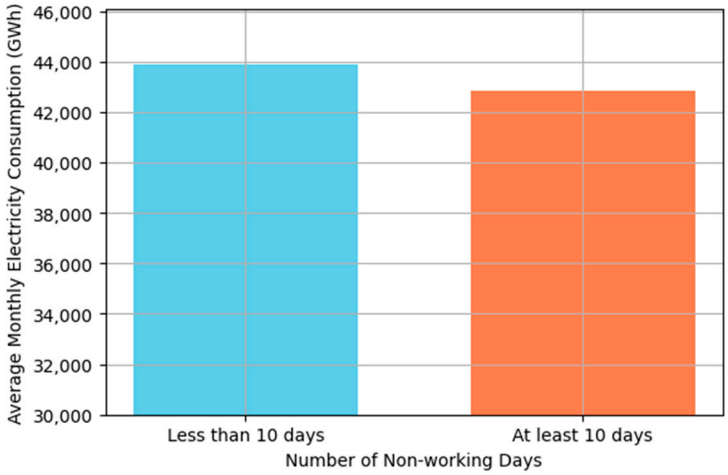

Figure 6. Comparison of average monthly electricity consumption between months with less than 10 non-working days and months with 10 or more non-working days.

2.4. Other External Factors

In mid-term load forecasting, aside from weather and calendar information, economic variables are often used as predictors [5,7,9]. While macroeconomic variables are essential for long-term electricity demand forecasting, they are considered somewhat indirect predictors for mid-term load forecasting. Therefore, this study aims to identify external factors that have a more direct influence on mid-term electricity demand. Looking at where electricity is consumed in South Korea, approximately 50% of the country's electricity usage has consistently come from the manufacturing sector for several decades. In exploring specific statistical indicators that could measure the scale of the manufacturing sector, which accounts for half of the nation's electricity consumption, we identified the number of registered factories, which was collected from the Korean Statistical Information Service site (https://kosis.kr). The following figure presents a chart showing the relationship between annual electricity consumption and the number of registered factories for the corresponding year.

Figure 7 illustrates the relationship between annual electricity consumption and the number of registered factories in South Korea from 2012 to 2021. Although the relationship is not perfectly proportional, there is a noticeable positive correlation between the two

variables. Because the number of registered factories is available on a semi-annual basis, the data can be converted into monthly data using interpolation methods when applied to a forecasting model.

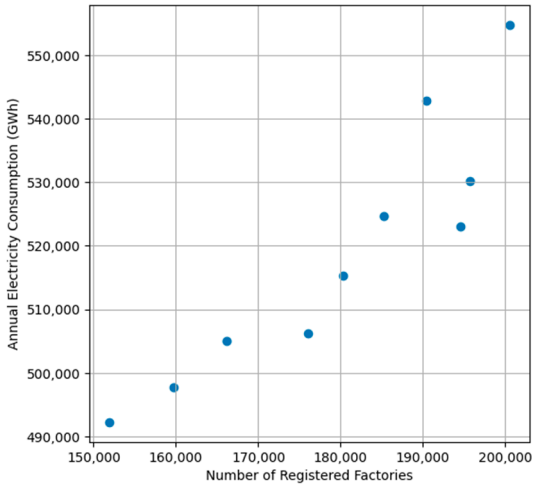

Figure 7. Annual electricity consumption versus the number of registered factories in South Korea (2012–2021).

To examine the impact of another external factor, particularly the COVID-19 pandemic, we compared monthly electricity consumption with the number of confirmed COVID-19 cases in South Korea. Figure 8 presents a time series chart of monthly electricity consumption (blue line) and confirmed COVID-19 cases (red line) from January 2020 to December 2021. Due to the rapid increase in confirmed cases, the COVID-19 data have been log-transformed and are displayed as log (monthly confirmed cases + 1). Surprisingly, the trends in confirmed COVID-19 cases and electricity consumption exhibit a certain degree of similarity. Contrary to the initial expectation that an increase in confirmed cases would lead to a decline in electricity demand, the data show that while electricity consumption decreased in the early months of 2020, it gradually recovered as the country adapted to the pandemic. This suggests that, over time, electricity consumption rebounded despite the continued rise in confirmed cases, reflecting a shift in consumption patterns as businesses and individuals adjusted to the prolonged impacts of COVID-19. Statistics of the confirmed cases are available at the World Health Organization (WHO) COVID-19 Dashboard (https://data.who.int, accessed on 5 October 2024).

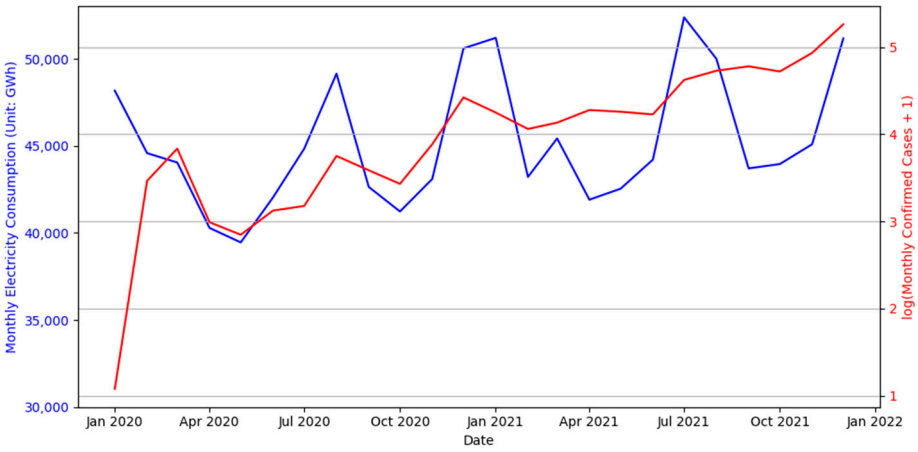

Figure 8. Electricity consumption versus log-transformed COVID-19 confirmed cases in South Korea from January 2020 to December 2021.

2.5. Correlation Analysis

Based on the above data analysis, we identified explanatory variables that can measure each influencing factor. These variables include temperature, total solar radiation, calendar-related factors, and economic indicators such as the number of registered factories. To quantify the relationship between each derived variable and electricity demand, we calculated the Pearson correlation coefficients between the monthly electricity consumption and each variable at the corresponding month. The statistical significance of each correlation coefficient was also assessed to ensure the reliability of the results. The following table summarizes the Pearson correlation coefficients and their *p*-values for all tentative explanatory variables.

As shown in Table 1, the variables "Month (numerical)", "Number of Saturdays", "Number of Sundays", and "Average cloud cover" were not found to be significant, so they are excluded from further analysis. Additionally, we examine the correlations among the explanatory variables that passed the significance test. If severe multicollinearity is suspected among variables, it may be necessary to remove some of them. The results are visualized and summarized in the figure below.

Table 1. Results of correlation analysis of factors influencing electricity consumption.

Influential Factors	Tentative Explanatory Variables	Correlation Coefficients	*p*-Value
Trend	Time index	0.4166	<0.0001
Seasonality	Month (numerical)	−0.0153	0.8686
Autocorrelation	Load of the same month in the previous year	0.8713	<0.0001
Calendar Data	Number of days	0.3513	0.0001
	Number of Saturdays	0.0372	0.6865
	Number of Sundays	0.0279	0.7622
	Number of holidays on weekdays	−0.2318	0.0108
Weather Data	Average temperature (°C)	−0.2432	0.0074
	Average highest temperature (°C)	−0.2801	0.0019
	Average cloud cover (1/10)	0.0483	0.6000
	Total solar radiation (MJ/m^2)	−0.335	0.0002
Other Factors	Number of registered factories	0.4111	<0.0001
	COVID-19 confirmed cases (log-transformed)	0.3093	0.0006

As shown in Figure 9, some pairs of explanatory variables exhibit very high correlations. Specifically, "Time index" and "Number of registered factories", as well as "Average temperature" and "Average highest temperature", have correlation coefficients that are nearly equal to 1. Consequently, we have decided to exclude "Time index" and "Average highest temperature" from further analysis to avoid multicollinearity issues. Additionally, "Total solar radiation" shows a high correlation with temperature variables, indicating potential redundancy in information. Therefore, we will also exclude "Total solar radiation". In summary, the proposed model in the following chapter will utilize only the following variables: load of the same month in the previous year; number of days; number of holidays on weekdays; average temperature (°C); number of registered factories; COVID-19 confirmed cases (log-transformed).

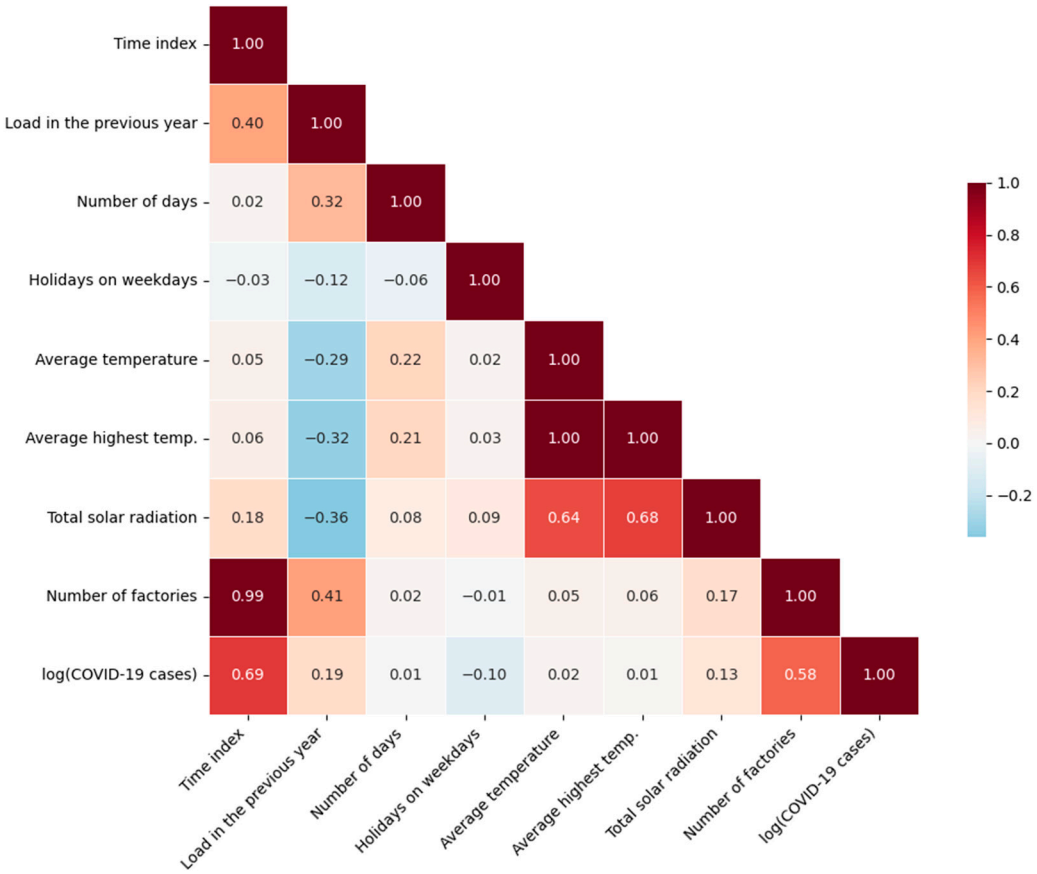

Figure 9. Heatmap of the correlation matrix for explanatory variables.

3. Methodology

In this section, we propose a regression-based model for forecasting monthly electricity demand in South Korea, using the findings from the above data analysis. The model incorporates the key characteristics of the electricity demand time series, such as trend and seasonality, along with external factors that are expected to influence electricity consumption. The mathematical formulation of the regression model proposed in this study is as follows:

$$Y_t = \beta_0 + \beta_1 Y_{t-12} + \beta_2 M_t + \beta_3 T_t + \beta_4 D_t + \beta_5 O_t + \beta_6 F_t + \beta_7 F_t \cdot M_t + \beta_8 log(C_t + 1) + \epsilon_t \quad (1)$$

where Y_t is the monthly electricity consumption at time t;

Y_{t-12} represents the electricity consumption for the same month from the previous year of time t;

M_t is a categorical variable representing the month of time t;

T_t represents the average temperature at time t;

D_t is the number of days in time t;

O_t is the number of holidays that falls on weekdays during time t;

F_t is the number of registered factories at time t;

C_t is the number of COVID-19 confirmed cases at time t;

β_0 is the intercept;

β_i are the coefficients for the independent variables, $i = 1, 2, \ldots, 8$, and

ϵ_t is the error term at time t.

Here is a detailed explanation of each term in the regression model. Y_t is the dependent variable, representing the monthly electricity consumption at time t, which we aim to forecast based on the following independent variables. Y_{t-12} is the lagged term that accounts for the strong seasonality with frequency of 12 month. M_t is the categorical

variable representing month of time t. As can be seen from Figure 2, different months exhibit distinct consumption patterns due to varying seasonal conditions. T_t representing the average temperature during month t, captures a non-linear relationship between temperature and electricity demand shown in Figure 4. The next two variables, i.e., D_t and O_t, are related to calendar data. As confirmed in Figure 6, electricity consumption tends to be relatively low on non-working days. In this study, the number of days (D_t) and number of holidays on weekdays (O_t) are used to represent the calendar feature of each month.

The external factors included as independent variables in the model are the number of registered factories (F_t) and the number of COVID-19 confirmed cases (C_t). The number of registered factories is provided on a semi-annual basis, with data collected only in June and December; thus, interpolation methods were employed to estimate the data for the remaining months. Since the number of registered factories can have different impacts on electricity consumption across months, an interaction term between the number of registered factories (F_t) and the month (M_t) was also included in the model. The monthly number of COVID-19 confirmed cases was log-transformed before being incorporated into the model to ensure proper scaling. This transformation allows for better handling of the sharp increase in confirmed cases and ensures that the variable is appropriately scaled for the regression model.

This model aims to capture the key factors influencing electricity consumption, incorporating both temporal dependencies and external variables such as weather, calendar data, and industrial activity. To identify the explanatory power of each variable, an analysis of variance (ANOVA) was performed after fitting the regression model to the training data, i.e., monthly electricity consumption data from 2012 to 2021. The results of the ANOVA are summarized in Table 2.

Table 2. ANOVA table obtained from fitting the training data with the proposed regression model.

Sources	Sum of Squares	*d.f.*	*F*-Value	*p*-Value
M_t	122,677,400	11	8.28	<0.000
Y_{t-12}	4,928,347	1	3.66	0.059
D_t	85,445	1	0.06	0.802
O_t	2,519,024	1	1.87	0.175
T_t	772,887	1	0.57	0.451
F_t	45,698,480	1	33.95	<0.000
$M_t \cdot F_t$	36,876,620	11	2.49	0.009
$log(C_t + 1)$	4,060,276	1	3.02	0.086
Residual	122,499,000	91		

As you can see from the table, the ANOVA results indicate that the month variable and the number of registered factories are the most significant predictors of electricity consumption, with extremely low p-values. These findings confirm the importance of including seasonal and industrial activity variables in the model to accurately capture variations in electricity demand.

4. Empirical Analysis

In this section, we evaluate the performance of the proposed regression model by comparing it with various existing forecasting methods. The analysis uses monthly electricity consumption data from 2012 to 2021 as the training set, while predictions are made for the years 2022 and 2023. In the following table, a subset of the training data is presented. Table 3 displays a subset of the training data, including entries from the beginning and the end of the training period. The notations in the header are consistent with those used in the explanation of the regression model in Section 3.

Table 3. A subset of the training data for monthly load forecasting in South Korea (2012–2021).

t	Year	Y_t	Y_{t-12}	M_t	T_t	D_t	O_t	F_t	C_t
1	2012	44,466.3	47,315.6	January	−2.8	31	2	147,600	0
2	2012	42,717.7	38,033.8	February	−2	29	0	148,071	0
3	2012	42,368.9	42,013.7	March	5.1	31	1	148,512	0
4	2012	37,448.9	37,712.0	April	12.3	30	0	148,984	0
...	...	...	...	...	...	...	...	...	...
117	2021	43,704.7	42,637.9	September	22.6	30	3	199,460	60,332
118	2021	43,952.7	41,230.4	October	15.6	31	2	199,815	52,613
119	2021	45,088.3	43,095.5	November	8.2	30	0	200,181	85,961
120	2021	51,165.7	50,595.7	December	0.6	31	0	200,535	182,904

The forecasting performance of each method is assessed using three widely accepted evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and mean absolute percentage error (MAPE). Here are the equations for the three performance metrics:

$$\text{MAE} = \sum_{i=1}^{n} \frac{|A_i - F_i|}{n} \tag{2}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(A_i - F_i)^2}{n}} \tag{3}$$

$$\text{MAPE} = \left\{ \sum_{i=1}^{n} \left(\frac{|A_i - F_i|}{A_i} \right) / n \right\} \times 100\% \tag{4}$$

where A_i represents the actual value and F_i denotes the forecast at i-th time point in the validation period, respectively. n is the number of months in the validation period. In this test, n = 24.

For the performance comparison, we used three univariate forecasting methods and five machine learning approaches. The univariate methods include Holt–Winters, SARIMA, and Prophet, while the machine learning models consist of XGBoost, Random Forest, LightGBM, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). Each method is briefly described below:

- Holt–Winters [14]: A triple exponential smoothing method that accounts for level, trend, and seasonality in time series data.
- SARIMA (Seasonal Autoregressive Integrated Moving Average) [15]: An extension of the ARIMA model that includes seasonal components. In this test, the SARIMA model is configured through a model identification process, where the parameters $(p,d,q)(P,D,Q)$ are set to (1,1,1)(1,1,1).
- Prophet [16]: A forecasting tool developed by Meta (formerly Facebook) that is known for effectively handling time series data with daily, weekly, and yearly seasonality.
- Random Forest [17]: A widely used ensemble method constructing multiple decision trees reduces overfitting and improves generalization through bagging and random feature selection.
- XGBoost [18]: Another ensemble learning method based on decision trees. It uses boosting to improve prediction accuracy by combining weak learners into a strong model.
- LightGBM [19]: A highly efficient gradient boosting framework that is optimized for speed and performance.
- RNN (Recurrent Neural Network) [20]: A type of neural network designed for sequential data. RNNs can capture temporal dependencies by using loops within the network structure, making them suitable for time series forecasting.
- LSTM (Long Short-Term Memory) [21]: A special kind of RNN that is capable of learning long-term dependencies in time series data.

Before presenting the results of the computational experiments, various details about the experiments conducted in this study are summarized in Table 4.

Table 4. Experimental configuration for electric load forecasting in this study.

Category	Description
Forecast Target	Monthly electric load of South Korea
Forecast Horizon	24 months of years 2023 and 2024
Training Period	120 months of years from 2012 to 2022
Data Sources	- Monthly electricity consumption data (2012–2023) from IEA - Weather data from Korean Meteorological Administration - Number of registered factories from the Korean Statistical Information Service - COVID-19 data from WHO Dashboard
Resources	- Software: Python 3.10 - Hardware: Intel Core i7 processor with 16GB RAM.
Python Libraries	The proposed method: LinearRegression; Holt–Winters: ExponentialSmoothing; SARIMA: SARIMAX; Prophet: Prophet; Random Forest: RandomForestRegressor; XGBoost: XGBRegressor; LightGBM: LGBMRegressor; RNN: Sequential; LSTM: LSTM
Hyperparameters	- XGBoost: n_estimators = 100, learning_rate = 0.1, max_depth = 3, random_state = 42. - Random Forest: n_estimators = 100, max_depth = 5, random_state = 42. - LightGBM: n_estimators = 100, learning_rate = 0.1, max_depth = 3, random_state = 42. - RNN: 1 hidden layer with 50 neurons, adam optimizer, 50 epochs, batch_size = 32. - LSTM: 1 LSTM layer with 50 units, 1 Dense layer, adam optimizer, 50 epochs, batch_size = 32.

In this study, a total of eight existing forecasting methods were used as benchmarks, as mentioned above. Additionally, we used ensemble methods, which combine two or more of these individual forecasting techniques, also as benchmarks. The following table summarizes the performances of the forecasting results for monthly electricity demand in South Korea for the years 2022 and 2023, using the proposed regression model and the eight individual benchmarks.

Table 5 presents the comparison of the forecasting performance between the proposed regression model and various existing forecasting methods based on three performance metrics. Overall, the proposed regression model outperforms all other methods across the board, achieving the lowest MAE, RMSE, and MAPE. In terms of MAE, the proposed regression model is the only method that achieves an average error below 1000 GWh. Furthermore, in terms of MAPE, the proposed model stands out as the only method to achieve an error rate in the 2% range. None of the benchmark methods achieve a 2% MAPE, which shows the superior predictive accuracy of the proposed model. Among the benchmark methods, RNN achieved the best overall performance, outperforming other machine learning models, while the model still did not surpass the accuracy of the proposed regression model. When using the machine learning methods for forecasting, the same independent variables from the proposed regression model were utilized as predictors. These variables include Y_{t-12}, M_t, T_t, D_t, O_t, F_t, and C_t. For technical purposes, M_t was converted into a numerical variable, and C_t was log-transformed appropriately before being used in the methods. To further examine the detailed forecast results, a time series chart has been plotted comparing the forecasts from the proposed method and the several best-performing benchmark methods against the actual values during the validation period.

Table 5. Results of the comparison test between the proposed model and the eight existing methods.

Forecasting Methods		MAE	RMSE	MAPE
Proposed Regression Model		921.4	1245.7	2.01%
Existing Methods	Holt–Winters	1588.2	1865.6	3.54%
	SARIMA	2095.6	2451.6	4.64%
	Prophet	1714.3	2107.0	3.81%
	Random Forest	1713.01	2342.91	3.85%
	XGBoost	1606.16	2113.06	3.62%
	LightGBM	1579.28	1963.11	3.50%
	RNN	1456.13	1863.89	3.28%
	LSTM	1871.12	2335.57	4.20%

Figure 10 illustrates the comparison between actual monthly electricity consumption (black solid line) and the forecasts from the proposed model (red dashed line) and three selected benchmark methods: Holt–Winters (green solid line), LightGBM (purple dashed line), and RNN (blue dotted line) for the validation period. As you can see from the table, it is clear that the proposed regression model tracks the actual values more closely than the benchmark methods. The proposed model shows consistent alignment with the actual data across most of the months, demonstrating its robustness for mid-term load forecasting.

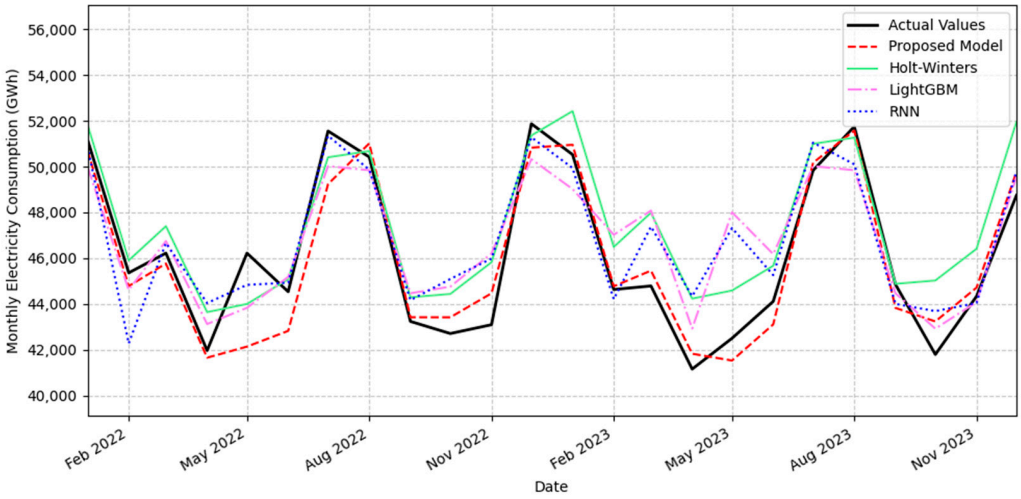

Figure 10. Actual values vs. forecasts (2022–2023) for the proposed model and selected benchmark methods.

Another reason for the lower performance of the machine learning methods lies in the characteristics of the monthly load forecasting problem, where only a limited number of data points are available for training. In cases with such limitations, complex machine learning models are often disadvantaged due to their higher data requirements. This limitation demonstrates the effectiveness of our proposed regression model, which is carefully designed with well-selected variables and structured to perform effectively even with a small data sample.

To enhance the predictive performance of the benchmark models, ensemble methods were applied by averaging forecasts from pairs and groups of three forecasting methods. A total of 28 combinations ($_8C_2$) and 56 combinations ($_8C_3$) were tested. While most combinations did not surpass the best-performing individual models—Holt–Winters, LightGBM, and RNN—a few combinations, specifically Holt–Winters + LightGBM, Holt–Winters + RNN, and Holt–Winters + LightGBM + RNN, showed improved performance. The results of these ensemble tests are summarized in the table below.

As seen in Table 6, although the ensemble methods improved the performance of the individual existing methods, the proposed regression model continues to demonstrate

significantly superior predictive accuracy. The selected ensemble combinations share a few key characteristics: they combine univariate methods with machine learning approaches, and the best-performing individual methods—Holt–Winters, LightGBM, and RNN—are consistently part of the most effective combinations. This suggests that blending the strengths of traditional time series forecasting methods with the adaptability of machine learning models can result in enhanced predictive performance, but it is still not sufficient to outperform the proposed regression model.

Table 6. Results of the comparison test between proposed model and the selected ensemble methods.

Forecasting Methods		MAE	RMSE	MAPE
Proposed Regression Model		921.4	1245.7	2.01%
Selected Ensemble Methods	Holt–Winters + LightGBM	1448.6	1767.4	3.24%
	Holt–Winters + RNN	1420.7	1716.0	3.19%
	Holt–Winters + LightGBM + RNN	1420.4	1733.8	3.19%

Several factors contribute to the superiority of the proposed regression model. First, independent variables were selected based on a thorough data analysis, identifying factors that have a direct influence on electricity consumption. Notably, variables reflecting South Korea's characteristics, particularly the significant electricity consumption by the manufacturing sector and the variable capturing the impact of COVID-19, played a significant role in improving model accuracy. Another factor is the treatment of the month variable as a categorical variable, which seems to have contributed to more precise predictions by capturing seasonal patterns more effectively. Moreover, considering interaction terms between the month variable and other factors, such as the number of registered factories, allowed the model to account for two-dimensional relationships between these variables and electricity consumption.

5. Conclusions

In this study, we proposed a regression-based forecasting method for mid-term electricity load forecasting in South Korea. The proposed model integrates both the temporal characteristics of electricity consumption, such as seasonal patterns and trends, as well as external variables, including weather, calendar data, and industrial activity. Through comparative experiments, we demonstrated that the proposed model outperforms various existing forecasting methods, including Holt–Winters, SARIMA, Prophet, and several machine learning-based approaches such as XGBoost, Random Forest, LightGBM, RNN, and LSTM, as well as ensemble methods of pairs or triplets of the existing methods. In particular, the proposed model achieved the lowest MAE, RMSE, and MAPE, with MAPE close to 2%, a result that none of the benchmark combinations were able to achieve.

One of the primary contributions of this study is the careful selection and inclusion of external variables that have a direct impact on electricity consumption, such as the number of registered factories and the impact of COVID-19. These factors, along with the categorical treatment of the month variable and the consideration of interaction effects, enhanced the model's ability to capture the complex relationships between electricity consumption and its influencing factors. This demonstrates the importance of considering both external and temporal factors in load forecasting models, particularly in countries like South Korea, where industrial electricity consumption is a major component of overall demand. It is worth noting, however, that in the future, the relationship between factory count and electricity consumption may weaken as self-reliant factories utilizing solar and other renewable energy sources become more common. Therefore, caution should be exercised when using this variable for long-term forecasts, as the increasing prevalence of self-reliant energy practices could diminish the predictive power of factory count in determining electricity demand.

This study also showed that regression models remain highly effective for load forecasting. While regression models are often considered classical approaches, the results confirm that they remain robust, flexible, and effective forecasting tools when suitable independent variables are introduced. Another key advantage is that regression models offer interpretability, a feature that distinguishes them from artificial neural network-based methods. For simpler time series predictions, such as monthly demand forecasting considered in this study, regression models may actually be more advantageous due to their interpretability.

Despite the promising results, there are several avenues for future research. First, further exploration into the use of additional external variables, such as economic indicators, could provide further improvements in forecasting accuracy. Additionally, hybrid models that combine the strengths of both statistical and machine learning approaches could be developed to enhance model robustness and flexibility. Moreover, studies on proper handling methods for predictor variables that require forecasting themselves, such as weather data, are also needed. Finally, the application of adaptive learning algorithms could further improve the model's performance in dynamic environments where electricity consumption patterns are subject to frequent changes.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study can be accessed from publicly available sources. The monthly electricity consumption data is available through the International Energy Agency (IEA) at https://www.iea.org/data-and-statistics/data-tools/monthly-electricity-statistics. Weather data were sourced from the Korean Meteorological Administration (KMA) and can be accessed via https://data.kma.go.kr/. Factory registration data were obtained from the Korean Statistical Information Service (KOSIS) at https://kosis.kr. COVID-19-confirmed case data used in this research are publicly available from the WHO COVID-19 Dashboard at https://data.who.int.

Acknowledgments: In this paper, AI-assisted tools were utilized during the preparation of the manuscript. Particularly OpenAI's ChatGPT 4o was used to assist with English writing.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. World Bank. *World Development Report 2024: The Middle-Income Trap*; World Bank: Washington, DC, USA, 2024. [CrossRef]
2. World Bank Group. Doing Business Archive. Available online: https://archive.doingbusiness.org/en/rankings?region=oecd-high-income (accessed on 6 October 2024).
3. Wang, H.; Alattas, K.A.; Mohammadzadeh, A.; Sabzalian, M.H.; Aly, A.A.; Mosavi, A. Comprehensive Review of Load Forecasting with Emphasis on Intelligent Computing Approaches. *Energy Rep.* **2022**, *8*, 13189–13198. [CrossRef]
4. Kuster, C.; Rezgui, Y.; Mourshed, M. Electrical Load Forecasting Models: A Critical Systematic Review. *Sustain. Cities Soc.* **2017**, *35*, 257–270. [CrossRef]
5. Rubasinghe, O.; Zhang, X.; Chau, T.K.; Chow, Y.H.; Fernando, T.; Iu, H.H.-C. A Novel Sequence to Sequence Data Modelling Based CNN-LSTM Algorithm for Three Years Ahead Monthly Peak Load Forecasting. *IEEE Trans. Power Syst.* **2024**, *39*, 1932–1947. [CrossRef]
6. Jain, A.; Gupta, S.C. Evaluation of Electrical Load Demand Forecasting Using Various Machine Learning Algorithms. *Front. Energy Res.* **2024**, *12*, 1408119. [CrossRef]
7. Li, J.; Lei, Y.; Yang, S. Mid-Long Term Load Forecasting Model Based on Support Vector Machine Optimized by Improved Sparrow Search Algorithm. *Energy Rep.* **2022**, *8*, 491–497. [CrossRef]
8. Jung, S.-M.; Park, S.; Jung, S.-W.; Hwang, E. Monthly Electric Load Forecasting Using Transfer Learning for Smart Cities. *Sustainability* **2020**, *12*, 6364. [CrossRef]
9. Liu, D.; Sun, K.; Huang, H.; Tang, P. Monthly Load Forecasting Based on Economic Data by Decomposition Integration Theory. *Sustainability* **2018**, *10*, 3282. [CrossRef]
10. Lee, G.-C. Regression-Based Methods for Daily Peak Load Forecasting in South Korea. *Sustainability* **2022**, *14*, 3984. [CrossRef]
11. Lee, J.; Cho, Y. National-Scale Electricity Peak Load Forecasting: Traditional, Machine Learning, or Hybrid Model? *Energy* **2022**, *239*, 122366. [CrossRef]
12. Baek, S.-M. Mid-Term Load Pattern Forecasting With Recurrent Artificial Neural Network. *IEEE Access* **2019**, *7*, 172830–172838. [CrossRef]
13. Ryu, S.; Noh, J.; Kim, H. Deep Neural Network Based Demand Side Short Term Load Forecasting. *Energies* **2017**, *10*, 3. [CrossRef]
14. Winters, P.R. Forecasting Sales by Exponentially Weighted Moving Averages. *Manag. Sci.* **1960**, *6*, 324–342. [CrossRef]

15. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis Forecasting and Control*, 4th ed.; John Wiley and Sons: Hoboken, NJ, USA, 2008.
16. Taylor, S.J.; Letham, B. *Forecasting at Scale*; e3190v2; PeerJ Inc.: San Francisco, CA, USA; London, UK, 2017. [CrossRef]
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
19. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
20. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]