energies

MDPI

Article

New Forecasting Metrics Evaluated in Prophet, Random Forest, and Long Short-Term Memory Models for Load Forecasting

Prajowal Manandhar [1], Hasan Rafiq [1], Edwin Rodriguez-Ubinas [1,*] and Themis Palpanas [2]

[1] DEWA R&D Centre, Dubai Electricity and Water Authority, Dubai P.O. Box 564, United Arab Emirates
[2] LIPADE, Université de Paris, 45 Rue Des Saints-Peres, 75006 Paris, France
* Correspondence: edwin.ubinas@dewa.gov.ae

Abstract: Data mining is vital for smart grids because it enhances overall grid efficiency, enabling the analysis of large volumes of data, the optimization of energy distribution, the identification of patterns, and demand forecasting. Several performance metrics, such as the MAPE and RMSE, have been created to assess these forecasts. This paper presents new performance metrics called Evaluation Metrics for Performance Quantification (EMPQ), designed to evaluate forecasting models in a more comprehensive and detailed manner. These metrics fill the gap left by established metrics by assessing the likelihood of over- and under-forecasting. The proposed metrics quantify forecast bias through maximum and minimum deviation percentages, assessing the proximity of predicted values to actual consumption and differentiating between over- and under-forecasts. The effectiveness of these metrics is demonstrated through a comparative analysis of short-term load forecasting for residential customers in Dubai. This study was based on high-resolution smart meter data, weather data, and voluntary survey data of household characteristics, which permitted the subdivision of the customers into several groups. The new metrics were demonstrated on the Prophet, Random Forest (RF), and Long Short-term Memory (LSTM) models. EMPQ help to determine that the LSTM model exhibited a superior performance with a maximum deviation of approximately 10% for day-ahead and 20% for week-ahead forecasts in the "AC-included" category, outperforming the Prophet model, which had deviation rates of approximately 44% and 42%, respectively. EMPQ also help to determine that the RF excelled over LSTM for the 'bedroom-number' subcategory. The findings highlight the value of the proposed metrics in assessing model performance across diverse subcategories. This study demonstrates the value of tailored forecasting models for accurate load prediction and underscores the importance of enhanced performance metrics in informing model selection and supporting energy management strategies.

Keywords: smart grid; load forecasting; machine learning; deep learning; time series; performance metrics

Citation: Manandhar, P.; Rafiq, H.; Rodriguez-Ubinas, E.; Palpanas, T. New Forecasting Metrics Evaluated in Prophet, Random Forest, and Long Short-Term Memory Models for Load Forecasting. *Energies* **2024**, *17*, 6131. https://doi.org/10.3390/en17236131

Academic Editor: Tek Tjing Lie

Received: 17 October 2024
Revised: 16 November 2024
Accepted: 28 November 2024
Published: 5 December 2024

1. Introduction

The evolution of the smart grid is pivotal in modernizing electrical power systems, enhancing efficiency, reliability, and sustainability. It enables real-time management of electricity flows and facilitates the incorporation of renewable energy sources while empowering consumers with better control over their energy usage [1]. Data mining is essential in this context, as it helps analyze the vast amount of data generated by smart meters and sensors to uncover patterns that improve demand forecasting, enhance grid reliability, and support proactive maintenance strategies, ultimately leading to more efficient energy management practices [2,3].

An important part of data mining for electricity utilities is the development and application of load forecasting models since they contribute to the optimal operation and planning of energy market systems, considering customers' behavior, hazardous environmental impacts, and the proportion of wasted energy [4]. Accurate predictions of electricity demand can lower the operating costs for energy systems by enabling more efficient electricity generation, transmission, and distribution. However, accurately predicting electricity demand can be

challenging due to variations in demand patterns [5], which are influenced by various factors, including weather, demographics, socioeconomic characteristics, technology, and tariffs [6,7]. Since ambient temperature is one of the most significant factors [8,9], the increase in global temperatures due to climate change is projected to increase the use of air conditioning systems, leading to increased energy demand in afternoons and evenings. [10]. Currently, heating, ventilation, and air conditioning (HVAC) systems account for more than half of the energy consumption of hot regions in the residential and commercial sectors [11]. In addition to the time of day and season, occupant behavior also plays a significant role in energy consumption, particularly in residential buildings [12]. Consequently, managing end-user demand is becoming increasingly critical for optimizing energy system operations [13], especially as more variable renewable energy sources, such as solar photovoltaics, are integrated into global energy systems [14]. There are several methods to analyze and model electricity demand, and that can be performed at multiple scales [15], including the following:

- Spatial resolution that considers geographical area or region.
- Temporal resolution corresponding to a data resolution, such as hourly, daily, or weekly.
- Temporal horizon that defines the scope of forecasting for short-term, medium-term, and long-term periods.

Various electric load forecasting methods have been proposed in the literature [16,17]. In relation to the temporal horizon, short-term load forecasting (STLF) refers to the use of load and other data, such as weather, economic, event, and time-related information, from prior periods to forecast electricity demand over a short period [15]. STLF permits electric utilities and grid operators to anticipate the amount of electricity needed to meet consumer demand, allowing them to efficiently plan and manage electricity generation, transmission, and distribution. Thus, STLF methods have recently attracted much attention in electric grids and markets because of their relatively high accuracy and reliability [16,18]. These STLF methods are conventional, machine learning-based artificial intelligence (AI), and hybrid methods [17,19]. The conventional methods address mainly time series [18,20] and regression analysis approaches [19–22]. On the other hand, regression methods seek to find the relationships between consumption target loads and input features by introducing statistical components. Regression methods have become popular among conventional methods because of their simplicity, good extrapolation performance, rapid forecasting speed, and relatively simple structure. However, simple linear regression methods have lower accuracy because of the non-stationarity and non-linearity of the loads, which limits their applicability in real-time.

Machine learning-based AI methods have been used in recent years to address the non-linear nature of electric loads. Behm et al. [21,23] proposed an integrated artificial neural network (ANN) approach for load forecasting that considers the effects of weather, calendar, and demographic information on the electric load. Shi et al. [22,24] proposed a deep recurrent neural network (Deep-RNN) by considering pooling to address the overfitting issue by increasing the data volume and diversity. Lazzari et al. [23,25] proposed Gaussian mixture clustering and the eXtreme gradient boosting classifier (XGBoost 1.5.1) to predict day-ahead behavior patterns. Then, they used this information to perform day-ahead forecasting of residential households via an ANN. He et al. [24,26] used parametric copula models within a deep belief network (DBN) to predict short-term loads. Similarly, Ouyang et al. [25,27] used a DBN architecture similar to that of copula models to perform forecasting. Deng et al. [26,28] used the Bagging-XGBoost method to perform STLF for a distribution transformer under extreme weather conditions. Jurado et al. [27,29] proposed an STLF method that uses total PV generation and electricity consumption. Furthermore, the authors extended their approach by proposing Monte Carlo dropout and kernel density estimation (KDE) to obtain probabilistic density forecasts. Deng et al. [28,30] proposed a multilevel convolutional neural network to improve the accuracy of multi-step forecasting with the help of the time-cognition factor. Langevin et al. [29,31] proposed a two-stage approach for the STLF of residential households. In the first stage, the authors estimated the past and future consumption of household appliances through a non-intrusive load monitoring (NILM) approach based on variational encoders (VE) and a deep generative model. In

the second stage, aggregated and disaggregated appliance consumption data were used to train the temporal convolutional neural network model (T-CNN) to perform short-term load forecasts for individual households. These AI-based methods have powerful feature learning capabilities that enable them to improve forecasting results, for which hyperparameters must be appropriately tuned. Weak hyperparameter tuning of deep neural networks can slow down the learning process because of the convergence of local optimal values. The pooling layer in CNN leads to a significant loss of valuable information. Furthermore, when the network layer becomes too deep, the RNN's ability to process long sequences is hampered due to Gradient vanishing, resulting in poor performance.

Hybrid methods have been proposed to solve problems that combine two or more AI methods to perform STLF. Wan et al. [30,32] used a hybrid combination of CNN, LSTM network, and attention mechanisms to resolve the issue of information loss because of excessively long input time series data. Wei et al. [31,33] used a detrending singular spectrum fluctuation analysis approach with the LSTM model to forecast long-range correlation components. They then combined the trend, periodic, and long-range correlation forecasted components to obtain the final forecasting results. Li et al. [32,34] proposed a hybrid STLF model based on multiple seasonal patterns and a modified firefly algorithm. Similarly, Kim et al. [33,35] and Sekhar et al. [34,36] proposed a hybrid CNN-LSTM model to perform STLF for residential loads. In contrast, Sadaei et al. [35,37] proposed an integrated method that combines fuzzy time series and CNNs. More recently, Ran et al. [36,38] proposed a hybrid STLF approach that combines complete ensemble empirical model decomposition with adaptive noise (CEEMDAN), sample entropy (SE), and transformer (TR) models. The study performed STLF at 4 h, 8 h, 12 h, and 24 h horizons for New York City. The results of the hybrid CEEMDAN-SE-TR approach outperformed those of other machine learning methods. By working on the same objective, Tong et al. [37,39] proposed an attention-based temporal–spatial convolutional network (ACN) for feature learning. They combined it with a multi-head attention mechanism method to develop an ultra-short-term forecasting model. Yang et al. [38,40] demonstrated the superiority of their proposed dynamic decomposition–reconstruction method with an ensemble technique from safe operation and rational dispatching. A summary of the key literature is shown in Table 1.

The performance evaluation of forecasting algorithms is essential in developing accurate forecast models. Domain-specific metrics provide insights that help users better understand the performance of these models. The most commonly used evaluation metrics for forecasting are the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). Even though these metrics provide information about the algorithm's performance in forecasting electricity consumption compared with actual consumption, these metrics do not indicate how much the forecasted consumption matches the actual consumption. Many forecasting methods may be biased and prone to under-forecast (when the forecasted consumption is less than the actual consumption at a given forecasting horizon) or over-forecast (forecasted consumption is greater than the actual consumption). To calculate this bias, the forecasting bias [39,41] or mean bias error (MBE) [40,42] is used. Forecast bias calculates the average bias of the model, and it can be either positive or negative. A positive forecast bias indicates over-forecasting of consumption. Similarly, a negative forecast bias indicates under-forecasting. However, reporting only positive or negative bias is insufficient, as further quantification of forecast bias is required to make meaningful changes to the model. To the best of the authors' knowledge, no study in the literature has proposed quantifying forecasting bias to evaluate short-term load forecast methods. Therefore, the main contribution of this work is to propose new detailed performance metrics (i.e., EMPQ) for the evaluation of forecasting methods that enable quantifying forecasting bias by calculating the maximum and minimum differences between forecasted and actual consumption and then reporting the differences in how much and in which direction (i.e., higher or lower) the forecasted consumption deviates from the actual consumption.

Table 1. Summary of reviewed literature regarding proposed algorithms, key findings, and evaluation metrics used.

			Metrics Used				Ref.
Year	**Algorithm Used**	**Key Findings**	**MAE, MAPE, RMSE, and Its Variants ***	**R^2**	**Mean Square Error (MSE)**	**Hit Rate (HR)**	
2020	Integrated ANN	Improved forecasts with weather and demographic info	•	•			[23]
2017	Deep-RNN	Addressed overfitting with pooling to diversify data	•				[24]
2022	Gaussian Mixture Clustering + XGBoost + ANN	Behavior patterns aid day-ahead forecasting	•				[25]
2017	Parametric Copula + DBN	Used copula models for accurate load prediction	•	•			[26]
2019	DBN	Similar DBN models achieved accurate forecasting	•			•	[27]
2022	Bagging-XGBoost	Bagging-XGBoost excelled under extreme weather	•				[28]
2023	Monte Carlo Dropout + KDE	Probabilistic density forecasts using KDE	•				[29]
2021	Multi-level CNN	Time-cognition factor improved multi-step forecasting	•				[30]
2023	Variational Encoders + TCN	NILM with VE improved household forecasting	•				[31]
2023	CNN + LSTM + Attention	Attention improved long time series data handling	•				[32]
2022	LSTM + Detrend Singular Spectrum Analysis	Combining trend and long-range correlation improved the accuracy	•	•			[33]
2020	Multiple Seasonal Patterns + Modified Firefly Algorithm	Modified firefly algorithm handled seasonal patterns	•				[34]
2019	Hybrid CNN-LSTM	Hybrid CNN-LSTM enhanced load forecasting	•		•		[35]
2023	Hybrid BiLSTM-CNN	Proposed hybrid method outperforms standard standalone LSTM, CNN models	•		•		[36]
2019	Fuzzy Time Series + CNN	CNN and fuzzy time series improved load forecasting	•				[37]
2023	CEEMDAN + SE + Transformer	CEEMDAN-SE-TR outperformed other methods	•	•			[38]
2023	Attention-based Temporal–Spatial CNN + Multi-head Attention	Attention mechanism enhanced ultra-short-term forecasting	•	•			[39]
2023	Dynamic Decomposition-Reconstruction with Ensemble	Dynamic decomposition improved safe operation	•				[40]

* Variants were Normalized RMSE, Normalized MAPE, Absolute Percentage Error (APE), and Mean Arctangent APE.

The proposed metrics provide both holistic insights and more comprehensive details about meeting the actual target and provide a better understanding of the tendencies of the developed models. They indicate how much the forecasts displace the actual value. This information helps address power distribution/generation cases when less demand or surplus is required on occasions, such as special days (e.g., public holidays) than on regular days. Thus, an illustration of the proposed metrics is provided with a comparative analysis of the STLF results from advanced approaches comprising traditional time series techniques, machine learning, and deep learning-based forecasting techniques. Hence, evaluation metrics that show under-forecasting and over-forecasting in time series load forecasting provide a comprehensive understanding of forecast performance, which aids in making informed decisions, optimizes resource allocation, and helps mitigate economic impacts. It also supports the continuous improvement of forecasting models, leading to more reliable and efficient operations.

The next section of this paper describes the materials and methods used in this study; Section 3 discusses the results; and Section 4 summarizes the overall work presented.

2. Materials and Methods

Figure 1 shows an overview of the methods followed for the data used and their processing performed in this study. First, electricity consumption data, occupant profile data, and weather data were collected, followed by preprocessing procedures. The feature selection, machine learning models, and cross-validation schemes were then defined in the following subsections.

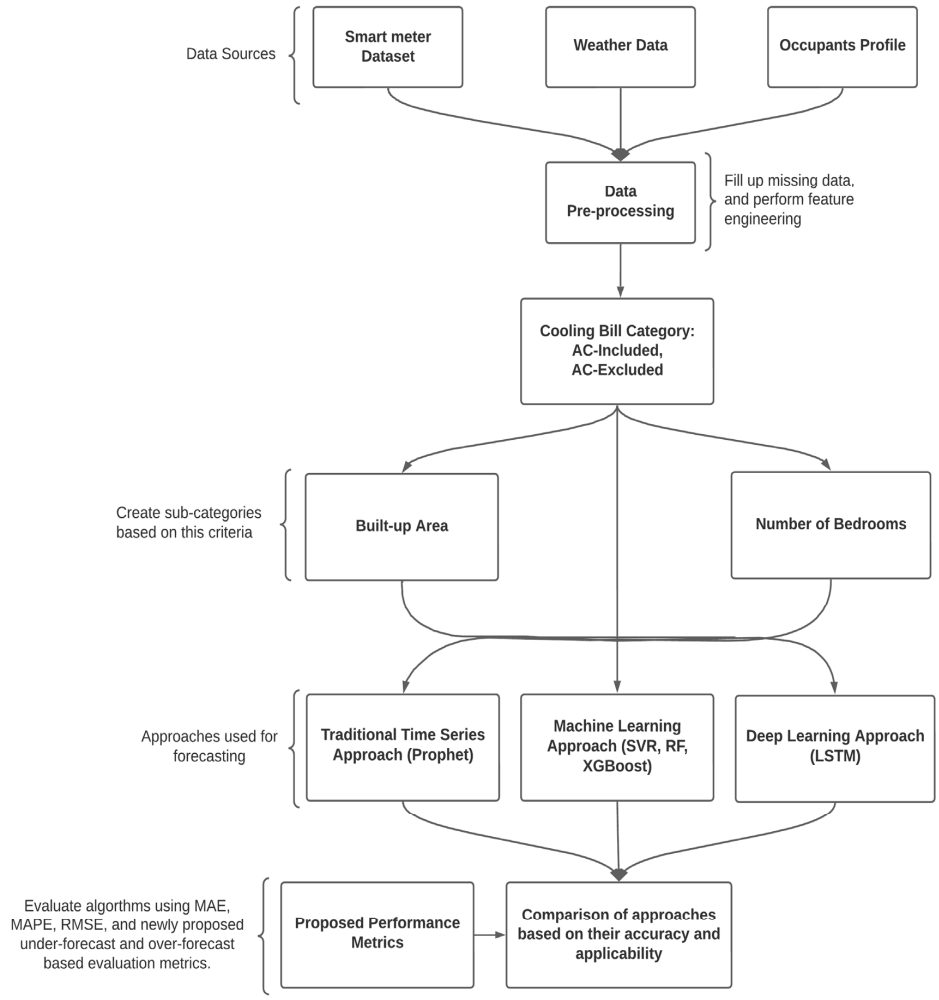

Figure 1. Flowchart of the method utilized to process the data used in this study.

2.1. Study Area

Dubai is the second-largest Emirate and economic capital (latitude: 25.25 N, longitude: 55.33 E) of the United Arab Emirates (UAE). It is part of the northern desert belt subregion and has an arid climate [43]. Dubai is home to approximately 3.41 million people, of whom 88.5% are expatriates and 11.5% are Emiratis residing across its 24 regions. Dubai has highly variable electricity consumption due to its scorching weather in summer and diverse demographics. Dubai's residential sector accounts for 30.4% of the Emirate's total electricity consumption, with air conditioning being the primary contributor to its energy demand. On the other hand, as per the Dubai Supreme Council of Energy, the market penetration of the district cooling sector in 2021 reached 25.6% and is expected to increase to 40% by 2030. Therefore, it was necessary to distinguish between customers whose utility smart meters record their cooling electricity consumption (those who used their own air conditioning) and those whose cooling energy consumption is not reflected in the utility smart meter readings. This latter group comprises apartments that receive the cooling services from the building's general air conditioning system (paid as part of the rent) and dwellings connected to a district cooling network.

2.2. Data Sources

This study used data from three sources: residential smart meters, weather data, and a customer survey. The collected data were used to develop short-term load forecasting models using advanced time series, deep learning, and other machine learning techniques. The following subsections provide the details of the smart meter dataset, weather data, and survey records.

2.2.1. Electricity Smart Meter Dataset

The analysis in this study was based on anonymized electrical smart meter data with a 15-min resolution from 8000 residential customers of the Dubai Electricity and Water Authority (DEWA). This dataset includes four years of data, from January 2018 to December 2021.

2.2.2. Weather Data

Dubai has an arid desert climate with two main seasons, extremely hot summers and warm winters, which are separated by two short transitional periods. The two transitional periods (April–May and October–November) are characterized by high variability and rapid weather changes [42,44].

Owing to its proximity to the Gulf, Dubai's relative humidity is higher than that of other cities at the center of the Arabian Peninsula. Therefore, hot and humid air masses affect cities, especially in summer. The present study utilized weather data collected from the Dubai International Airport weather station during the same period as the smart meter data (2018–2021) [45]. The temperature and relative humidity were used to determine the intensity and variability correlation between weather, electricity consumption, and input features for STLF modeling.

2.2.3. Survey Data

The smart meters and weather data were complemented with the information collected from the anonymized household data survey conducted by DEWA's "My Sustainable Living Program". This program targeted residential customers to support them in living a sustainable lifestyle and improving their energy demand behavior [40]. The survey data were utilized to define the customers' profiles, considering the type of dwelling, total occupants, built-up area, cooling provider, and bedroom number.

2.3. Data Processing

2.3.1. Data Sampling and Preprocessing

Out of the initial 8000 customers, only 586 fully completed the survey. Of these, only 439 had smart meter data across the entire study period (four years). Thus, 439 households

were ultimately considered for further analysis. The 15-min-resolution smart meter data were resampled to one-hour intervals for this study. The consumption data were merged with the survey data to extract the occupant's profile for each consumption. Moreover, weather station temperature and relative humidity data were also incorporated into the resulting dataset. The utilized smart meter dataset includes approximately 35,064 observation records from every 439 consumers with continuous records over the study period.

2.3.2. Data Subsets

Segmenting the data into relevant subsets improves load forecasting accuracy and offers utilities more actionable insights. Therefore, the author divided the datasets into two main subsets: 'AC-included' (297 dwellings conditioned using their systems) and 'AC-excluded' (142 dwellings conditioned by a general system or a district cooling network). This initial division was based on differing energy demand patterns and weather dependencies. Dwellings with individual air conditioning systems have more variable energy demand and higher peaks. Their greater variability is evidenced throughout the hours of the day and the months of the year. Due to the particularities of these two groups, this segmentation is critical for generating accurate forecasting models and managing peak demand.

Socio-demographic factors also impact the households' energy consumption levels and patterns. Therefore, to capture some of these factors, each of the two main subsets was further subdivided based on built-up area and number of bedrooms, allowing for even more precise energy demand forecasting. The 'built-up area' subgroup was divided into six subsets (0–50, 51–100, 101–150, 151–200, 201–300, and over 300 square meters). On the other hand, the 'number of bedrooms' subgroup was divided into four subsets (zero-studio, one, two, and three).

2.3.3. Feature Selection

The feature selection process enables identifying the most relevant features from the available list. It facilitates isolating profitable features to ensure quality in the underlying information. In addition, feature selection helps reduce the dimensionality of the available data.

The ranker-based feature selection approach presents many advantages, including simplicity, efficiency, scalability, interpretability, and the ability to improve model performance by focusing on the most relevant features of the dependent target variable. It generates the rank of the features via a sorted list based on various scores, such as distance, correlation, information gain, and consistency measures [46]. Therefore, considering their benefits, the authors used this approach to evaluate and identify features based on their importance to the target variable [43,45].

2.3.4. Data Splitting

The dataset was trained and validated using samples representing the entire population to create a robust forecasting model. In this work, three initial years of data (2018, 2019, and 2020) were used to train the models using a cross-validation approach, and the evaluation metrics were calculated on the test set of the fourth year (2021), the data used for prediction.

2.4. Short-Term Forecasting Methods

Based on the literature review's findings, summarized in the introduction, the authors decided to validate the proposed new metrics by comparing the performance of three of the most commonly used models for short-term load forecasting: Prophet (advanced time series approach), Random Forest (machine learning), and LSTM (deep learning). They were utilized in two forecasting scenarios: one day and one week. However, all the metrics were reported for one whole year of test data.

The following sections provide more information about these models and insights into tuning hyperparameters to enhance the models' performance. Including the details of the hyperparameters' tuning enhances the models' transparency and supports their reproducibility and reliability.

2.4.1. Time Series Analysis Approach

Time series analysis involves time series data and trend analysis. Time series data generally follow periodic time intervals collected at particular time intervals. The applications of time series analysis include understanding the underlying data patterns and fitting a statistical model accurately so that the process can be applied to forecasting and monitoring. Hence, analyzing time series data requires unique tools and methods that investigate trends, seasonality, and noise.

Prophet Model: The Prophet model is a time series-based forecasting model developed by Facebook's data analytics team [47]. Prophet processes time series data with a decomposable model where non-linear trends fit with daily, weekly, and yearly seasonality along with the holiday effect. Recently, it has become popular because it works well with time series data, has strong seasonality, and has multiple seasons of historical data [48].

Tuning of Hyperparameters in the Prophet Model: The Prophet model allows automatic tuning of the four primary hyperparameters.

- The most influential parameter is the 'changepoint prior scale', which determines the scale of the change at the trend change point in the time series. It is a regularization penalty term referred to as L1-Lasso. When this parameter is very small, the model tends to underfit; when this value is too large, the model tends to overfit, with the trend changing significantly at the change point. Its default value is 0.05, and its recommended range is between 0.001 and 0.5.
- The second most impactful parameter is the 'seasonality prior scale', which controls the magnitude of the seasonality fluctuation. It can be considered a regularization penalty term referred to as L2-Ridge. Its default value is 10, which indicates that no regularization is applied. Its recommended range is between 0.01 and 10, where a smaller value corresponds to having a smaller seasonality.
- The 'holidays prior scale' is similar to the 'seasonality prior scale' and determines the scale of holiday effects. Its default value is 10, which means that no regularization is applied. Thus, its recommended range is between 0.01 and 10, where a smaller value corresponds to fewer holidays.
- The 'seasonality mode' has two main options: additive and multiplicative. The additive model considers trends, seasonality, and other effects by adding them when making predictions. This approach is appropriate for time series models with relatively constant seasonal variation over time. On the other hand, the multiplicative model considers multiplying the trend, seasonality, and other effects when making predictions.

Other existing hyperparameters include the changepoint range, growth, changepoints, yearly seasonality, weekly seasonality, daily seasonality, and holidays. These required manual tuning using a trial-and-error-based approach.

2.4.2. Machine Learning Approach

Machine Learning (ML) is a subset of Artificial Intelligence in which models learn from data, identify patterns, and assist in making decisions with fewer human interactions. The popular ML approaches used in STLF are Support Vector Regressors (SVRs), Neural Networks (NNs), Decision Trees (DTs), and Random Forest (RF). This section describes the RF-based machine learning approach used to develop the STLF model. Even though deep learning is a subset of machine learning, it requires more computations than traditional approaches do. Hence, this method is addressed separately in the following subsection.

Random Forest: DTs are supervised learning methods that predict discrete values of the parameters used to train a model. DTs start at the root and traverse slowly toward the nodes by partitioning the predictors using the divide and conquer strategy per predictor, with the greatest impact on uniformity in the result after a supervised learning technique is used to combine the multiple regressions per split. RF is a supervised learning algorithm that ensembles the outputs from various DTs (with bagging) or DTs with multiple subsets of training data. It helps improve the performance of function approximation or regressive prediction. The ensemble approach in RF uses various learning approaches and a majority of votes concept to obtain better predictive performance than that obtained from any constituent DT.

Tuning of Hyperparameters in the Random Forest Model: A grid of hyperparameter ranges can be defined using ScikitLearn's RandomizedSearchCV method, and performing random sampling from the grid with K-Fold cross-validation allows one to walk through each combination of grid values. The following are the parameters that need to be tuned in Random Forest.

- 'n estimators' refers to the number of trees considered in the forest;
- 'max features' refers to the maximum number of features considered for splitting a node;
- 'max depth' refers to the maximum number of levels in each decision tree;
- 'min samples split' refers to the minimum number of points placed in a node before it is split;
- 'min samples leaf' refers to the minimum number of points allowed in a leaf node;
- 'bootstrap' refers to an approach for sampling data points (with or without replacement).

2.4.3. Deep Learning Approach

Deep learning is a subset of machine learning. Deep learning models have a neural network architecture with multiple layers of processing units that have been applied successfully to a broad set of problems in different areas of image recognition, especially natural language processing. It is a type of artificial intelligence that imitates how humans gain certain kinds of knowledge as they gain information from a big data set. One of the key differences between machine learning and deep learning methods is that deep learning models require a high amount of data and more computational power to solve more complex problems, which cannot be solved with traditional machine learning methods.

Long Short-Term Memory (LSTM) is a widely used technique for deep learning based on recurrent neural networks with feedback connections. LSTM networks allow the capture of sequence pattern information in a time series manner. LSTMs utilize only the attributes provided in the training set to work with temporal correlations. LSTM, as its name states, tends to have both long-term and short-term memory. During the training phase, the weights and biases change during each training episode, comparable to how physiological changes in synaptic strength consider long-term memories; the activation patterns in the network that change once during each time step are similar to how the electric firing patterns change in the brain to consider short-term memories. The trained deep learning model consists of its architecture's LSTM, Dense, and Dropout layers.

Tuning of Hyperparameters in LSTM: Various hyperparameters in LSTM provide exceptional results when appropriately tuned.

- Number of hidden neurons (nodes) and layers: These two parameters are selected with a trial-and-error approach. Regularization techniques are generally used within a layer, which helps increase the model's accuracy.
- Number of units in a dense layer: A dense layer is a composite layer where each neuron receives input from all the neurons from the previous layers. Hence, it is known as 'densely connected'. Dense layers help improve the overall accuracy, and 5–10 units or nodes per layer are usually good input choices.
- Dropouts: Each LSTM layer, when accompanied by a dropout layer, helps to avoid overfitting in the training phase as it helps to bypass randomly selected neurons or

nodes. Thus, it is essential to note that dropout layers should not be used with output layers because they may hinder the model's output and cause calculation errors. A value of 20% is generally considered a good choice to balance between preventing model overfitting and maintaining model accuracy.

- Weight Initialization: Allocating different sets of weights gives the optimization process different starting points, thus providing different performance characteristics. Therefore, weights should be randomly initialized to small numbers to randomize the search process.
- Decay Rate: The Decay Rate helps update the nodes' weights after each training phase. These weights are multiplied by a factor slightly less than 1, eventually preventing them from growing too large.
- Activation Function: Activation functions help define a node's output as either the 'ON' or 'OFF' state. These functions are used to introduce non-linearity to models, allowing deep learning models to learn predictions from non-linear boundaries. The selection of the activation layer depends on the purpose of the application. 'SoftMax' is a popular activation function that can interpret output as probabilities.
- Learning Rate: This 'learning rate' hyperparameter helps define how quickly the network learns and updates its parameters. Setting a higher learning rate accelerates the learning process, but the model may not converge to global minima. Hence, a lower learning rate (between '0.0' and '0.1') is usually preferred, which slows down the learning process and helps the model converge to global minima.
- Number of Epochs: This hyperparameter helps to define the running number of iterations of the learning phase of the dataset. Theoretically, this number can be set to any integer value. However, it is generally set after gradually increasing until the validation accuracy decreases, considering that its training accuracy increases at that point.
- Batch Size: The 'batch size' defines the number of samples to be considered in the model before it updates the model's internal parameters. Compared with smaller sizes, larger sizes allow large gradient steps for the same number of samples considered. Therefore, a widely preferred value for a good batch size is 32.

2.5. Traditional Evaluation Metrics and Their Strengths and Limitations

The three main metrics used to validate the models' effectiveness and determine the models' fitness are the RMSE, MAE, and MAPE. These metrics have their own strengths and limitations, which are discussed below:

2.5.1. Root Mean Squared Error

The RMSE is a valuable metric for assessing forecasting accuracy because of its simplicity and sensitivity to large errors. It is a commonly used metric in forecasting to evaluate the accuracy of predictions by measuring the differences between the predicted values and actual observations. Therefore, it is essential to consider its limitations and use it with other metrics and diagnostic tools to understand the forecasting model performance comprehensively. Table 2 shows the advantages and disadvantages of using the RMSE metric to evaluate short-term forecasting methods.

Considering (x_i) as the actual value and (x'_i) as the modeled or predicted value for the 'n' number of observations, the RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - x'_i\right)^2} \quad (1)$$

Table 2. Traditional evaluation metrics' advantages and limitations.

Metric	Advantages	Limitations
RMSE	- Provides a straightforward measure of forecasting accuracy that is easy to interpret. It represents the average magnitude of errors between predicted and actual values, clearly indicating the model's performance. - Gives higher weight to large errors than metrics like MAE. This sensitivity can be beneficial in scenarios where large errors have significant consequences and must be minimized. - Allows for comparing forecasting accuracy between different models or approaches. Lower RMSE values indicate better predictive performance, making it useful for model selection and improvement.	- Can be sensitive to outliers, meaning that extreme values can disproportionately influence the overall error measure. This may lead to overemphasizing the importance of outliers in evaluating forecasting models. - Does not provide information about the directionality of errors (over or underestimation). As a result, it may not fully capture the performance of a forecasting model, especially if systematic biases exist. - Is sensitive to the scale of the data. Changes in the dependent variable's scale can affect the RMSE's magnitude, making comparisons between RMSE values across different datasets or time series with different scales challenging.
MAE	- Provides a simple and intuitive measure of forecasting accuracy. It represents the average magnitude of errors between predicted and actual values, making it easy to interpret and communicate to stakeholders. - Is scale-invariant, meaning that it is not affected by the scale of the data or the units in which the variables are measured. This property makes MAE suitable for comparing forecast accuracy across different datasets with varying magnitudes. - Is less sensitive to outliers than other error metrics like RMSE, as it does not square the errors. This robustness of outliers makes MAE a reliable measure of forecasting accuracy, especially in datasets with extreme values. - Treats all forecasting errors equally without considering their magnitude or directionality, bringing simplicity.	- Its simplicity in treating all forecasting errors equally, without considering their magnitude or directionality, may also be a limitation. Not differentiating between small and large errors could significantly impact decision-making and may not align with the priorities of certain forecasting applications where minimizing large errors is crucial. - Does not provide information about the variability or dispersion of forecast errors. In datasets with heterogeneous error magnitudes, MAE may not fully capture the distribution of errors or the consistency of forecasting performance. - Is not differentiable with respect to model parameters, making it unsuitable for optimization algorithms that require gradient-based optimization techniques. This limitation restricts the use of MAE in model tuning or parameter estimation tasks. - Does not distinguish between overestimation and underestimation errors, treating them equally. This lack of sensitivity to error directionality can be considered a drawback in certain forecasting applications, especially those where overestimation or underestimation has asymmetric costs or consequences.
MAPE	- Is scale-invariant, meaning that it is not affected by the scale of the data or the units in which the variables are measured. This property makes MAPE suitable for comparing forecasting accuracy across different datasets with varying magnitudes. - Gives equal weight to errors across different levels of actual values, reflecting forecast accuracy proportionally regardless of the magnitude of the observations. This property is particularly useful when forecasting across a wide range of values.	- Is sensitive to extreme values or outliers in the data, which means that large errors can disproportionately influence the overall percentage of errors. This sensitivity may distort the assessment of forecast accuracy, especially in the presence of outliers. Thus, MAPE should be used wisely with other metrics to evaluate forecast performance comprehensively.

2.5.2. Mean Absolute Error

The MAE is a simple and robust metric for evaluating forecasting accuracy, offering advantages such as ease of interpretation, scale-invariance, and robustness to outliers. However, it also has limitations, including its inability to differentiate between error magnitudes, lack of sensitivity to error directionality, and unsuitability for optimization tasks. The MAE should be used wisely in conjunction with other metrics to obtain a comprehensive assessment of forecast performance. Table 2 shows the advantages and limitations of this metric. The MAE measures the errors between paired observations and is calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_i'\right| \tag{2}$$

2.5.3. Mean Absolute Percentage Error

The MAPE is another commonly used metric in forecasting to evaluate the accuracy of predictions. It is a valuable metric for evaluating forecasting accuracy, offering a percentage-based measure that is intuitive and scale-invariant. It expresses the average magnitude of errors relative to the actual values, allowing for straightforward comparisons across different datasets and forecast horizons. However, one needs to be careful of its limitations, particularly regarding division by zero, sensitivity to extreme values, and bias towards underestimation, as Table 2 explains.

The MAPE is calculated as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - x_i'}{x_i}\right| \tag{3}$$

2.5.4. Summary of Traditional Evaluation Metrics Comparison

The RMSE, MAE, and MAPE each offer distinct advantages and limitations for evaluating forecasting accuracy. The RMSE is beneficial for highlighting large errors, providing a measure that gives higher weight to large deviations, which is useful in cases where minimizing significant errors is critical. However, it is sensitive to outliers and scale changes, making comparing datasets challenging. The MAE is simple, intuitive, and robust against outliers, as it treats all errors equally. Its scale-invariance makes it suitable for comparing forecasts across datasets of different magnitudes, though its equal treatment of errors can be a drawback in applications where larger errors are prioritized. Like the MAE, the MAPE is scale-invariant and expresses forecast accuracy as a percentage, which is especially useful for comparing forecasts across varying ranges. However, the MAPE is highly sensitive to outliers, as large errors can skew its interpretation, making it less reliable in datasets with extreme values. Figure 2 summarizes these key aspects.

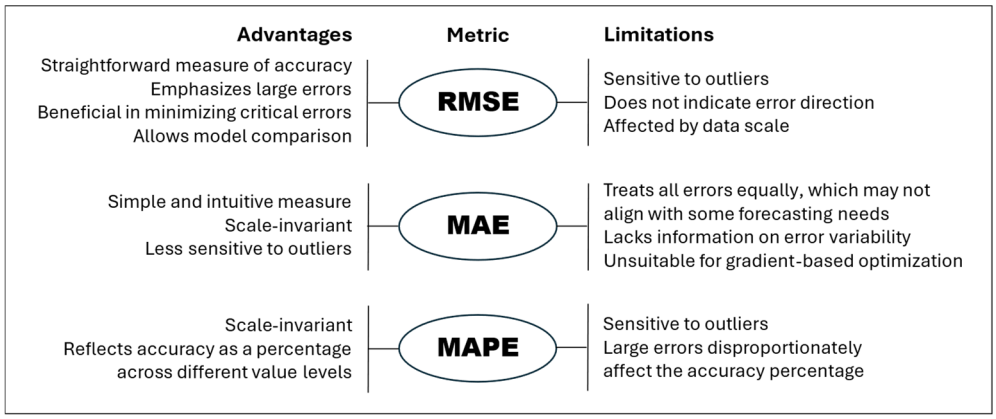

Figure 2. Traditional evaluation metrics: summary of advantages and limitations.

2.5.5. Bias

Bias metrics quantify a forecasting model's tendency to overestimate or underestimate actual values. This information is valuable for understanding the directionality of errors and identifying and addressing systematic biases. Bias metrics should be used with error metrics and other diagnostic tools to ensure a comprehensive assessment of forecasting performance and effectively inform decision-making.

- Bias metrics are often easy to interpret and communicate since they represent the average difference between the predicted and actual values. A positive bias indicates overestimation, whereas a negative bias indicates underestimation, providing clear insights into the forecasting model's performance.
- Bias metrics can help align forecasting performance with decision-making objectives by highlighting any systematic tendencies that may impact prediction accuracy. Decision-makers can use this information to adjust their expectations or take corrective actions to mitigate bias-related risks.
- Bias metrics complement traditional error metrics such as the MAE or RMSE by focusing specifically on the directionality of errors. By considering bias and error metrics together, stakeholders can gain a more comprehensive understanding of forecasting performance. However, bias metrics alone may not comprehensively evaluate forecasting accuracy, as they quantify only the directionality of errors without considering their magnitude or variability. The 'bias' is calculated as follows:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n} x_i - x_i' \tag{4}$$

2.6. Proposed Evaluation Metrics for Performance Quantification

Traditional evaluation metrics such as the RMSE, MAE, and MAPE are commonly used to assess forecasting models. These metrics offer valuable information related to the models' performance. However, they provide limited insights into biases, such as over- or under-forecasting. To address this need, the authors proposed new metrics within the EMPQ framework, designed to complement existing metrics by evaluating the model's bias tendencies and improving decision-making based on forecast demand accuracy.

The proposed EMPQ metrics classify predictions into three main categories: under-forecasts, over-forecasts, and overlapping values, which are instances where predictions closely match the actual values. For each forecast category, the metrics capture detailed aspects of prediction accuracy, including the percentage of forecasts near the actual values, intermediate range, and well above actual values (extreme deviations). These three categories enable an understanding of how close or far predictions lie from actual values and facilitate model adjustments for better alignment with real-world demand trends. This approach provides a layered analysis, helping identify patterns and systematic biases within forecasting models.

Each category corresponds to a different (bias) distribution defined based on calculated quartiles. The quartiles are four continuous intervals containing 25% of the data points, usually used to help understand the distribution and spread of data. Quartile 1 (Q1: 25th percentile) defines the border of the lower quarter of the data points. In the proposed EMPQ, any point of the under- and over-forecast falling in or below Q1 is considered part of the near-actual category. Similarly, Quartile 3 (Q3: 75th percentile) defines the border of the upper quarter of the data, and any point of the under- and over-forecast falling above this quartile is considered part of the well-above category. The third category, the intermediate range, contains all the data points higher than Q1 and lower than or equal to Q3. The three categories are illustrated in Figure 3.

The key calculations of the proposed EMPQ metrics include the minimum difference (mindiff), the maximum difference (maxdiff), and the average difference (avgdiff) between the actual and predicted values. Algorithm 1 provides the pseudocode for calculating these metrics, detailing the process for quantifying the forecast distribution around mindiff, maxdiff, and avgdiff. This clustering enables decision-makers to assess forecasting precision readily and make necessary adjustments for operational efficiency. For example, a high percentage of forecasts lying close to the actual value (within Q1) indicates high accuracy, whereas values exceeding Q3 indicate significant deviations.

The EMPQ framework is especially beneficial in scenarios with sudden demand shifts, such as ramp-up or ramp-down events in energy usage. By quantifying both over- and under-forecasting errors, the proposed metrics support continuous monitoring and real-time adjustments, enhancing the adaptability of the forecasting models to dynamic changes in demand.

Algorithm 1 Evaluation Metrics for Performance Quantification

Input: Observed and Forecasted values
Output: Performance Quantified metrics
for each modeled data do
Calculate the difference between modeled and actual data.
Differentiate data as under-forecast or over-forecast based on negative or positive difference values.
for the data with negative difference do
Generate mindiff, maxdiff, and avgdiff between observed and actual values.
Create three clusters C_{uf1}, C_{uf2}, C_{uf3} centered with these three difference values.
Apply clustering technique to rest of the differenced data considering minimum Euclidean distance (similar to k-means, where k = 3).
Report the population of each of the three clusters (as a percentage).
Report quartile values Q1, Q3 which represent mindiff (i.e., C_{uQ1}) and maxdiff, i.e., (C_{uQ3}) as a minimum and maximum deviation (or bias), respectively, from actual observed values.
Classify reported output as under-forecast.
end for
for the data with positive difference do
Generate mindiff, maxdiff, and avgdiff between observed and actual values.
Create three clusters C_{of1}, C_{of2}, C_{of3} centered with these three difference values.
Apply clustering technique to rest of the differenced data considering minimum Euclidean distance (similar to k-means, where k = 3).
Report the population of each of the three clusters (as a percentage).
Report quartile values Q1, Q3 which represent mindiff (i.e., C_{oQ1}) and maxdiff, i.e., (C_{oQ3}) as a minimum and maximum deviation (or bias), respectively, from actual observed values.
Classify reported output as over-forecast.
end for
end for

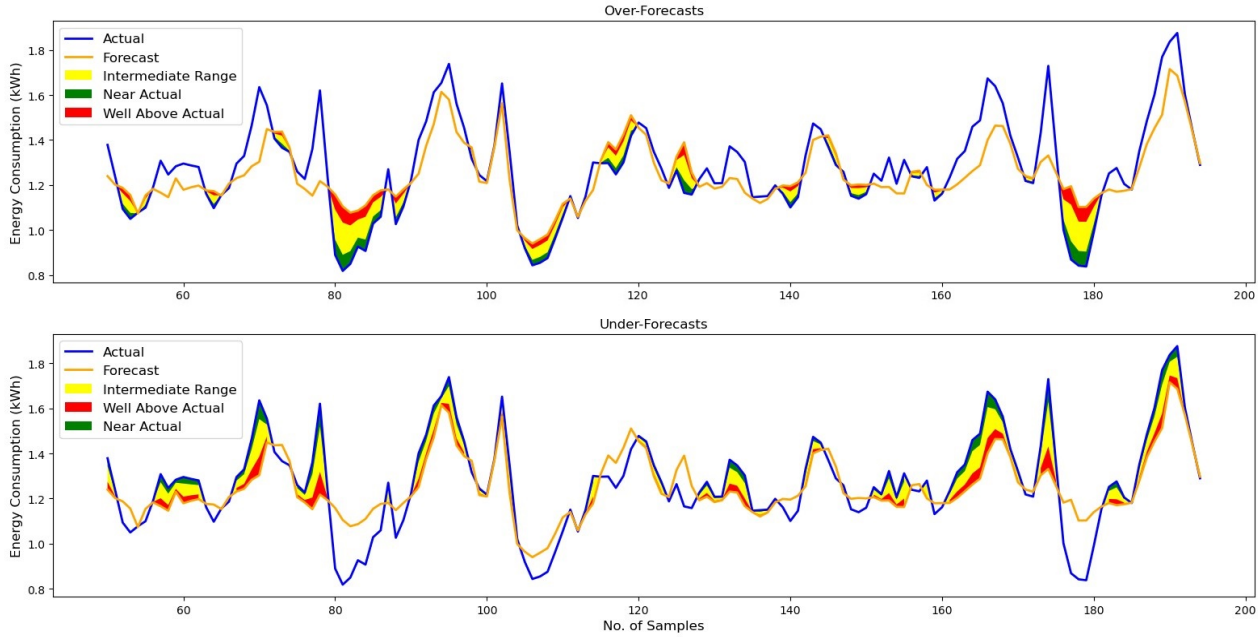

Figure 3. Illustration of the quantification of over-/under-forecasts.

Interpretation of the Proposed Metrics

The EMPQ metrics not only quantify the directionality of errors (under- vs. over-forecasting) but also evaluate how close the predictions are to the actual values. The metrics highlight values within Q1 as near actual values, suggesting high precision, whereas values beyond Q3 represent well above actual values. This classification provides practical insights for improving forecasting models by identifying consistent trends in under- or over-forecasting, which are critical for model selection and real-time adjustments in demand management.

For decision-makers, models with the highest percentage of predictions near the actual value in both the under- and over-forecasting categories (i.e., metrics C_{of1} and C_{uf1}) and the lowest maximum deviation percentages (C_{oQ3} and C_{uQ3}) are preferred. This balanced approach to evaluating forecast precision supports more informed resource allocation and operational adjustments, ensuring that models meet strategic goals in energy demand management. Table 3 presents the proposed evaluation metrics for performance quantification.

Table 3. Proposed evaluation metrics for performance quantification.

	Proposed Evaluation Metrics for Performance Quantification	Type
1	Percentage of instances having Over-Forecasts	
2	Percentage of instances having Under-Forecasts	
3	Percentage of Over-Forecasts near actual value (C_{of1})	
4	Percentage of Over-Forecasts intermediate range near Actual (C_{of3})	
5	Percentage of Over-Forecasts well-above actual value (C_{of2})	Over_forecasts
6	Max-Deviation Percentage from actual value	
7	Min-Deviation Percentage from actual value	
8	Percentage of Under-Forecasts near actual value (C_{uf1})	
9	Percentage of Under-Forecasts intermediate range near to actual (C_{uf3})	
10	Percentage of Under-Forecasts well-below actual value (C_{uf2})	Under-Forecasts
11	Max-Deviation Percentage from actual value	
12	Min-Deviation Percentage from actual value	

3. Results and Discussion

This section presents the results and discusses the use of the established and proposed metrics in the selected datasets and subsets in four subsections. The first covers the trends and

seasonal profiles of the main categories, 'AC-included' and 'AC-excluded'. The second includes the forecasting analysis for the two cooling categories at two different forecast horizons (one day ahead and one week ahead). Finally, the final two sections discuss the subsets' results based on the dwelling's built-up area and the number of bedrooms. Three advanced techniques were utilized for the two cooling categories in each domain of the time series: Prophet, Random Forest (machine learning), and LSTM (deep learning). However, for the dwelling subcategories (built-up areas and the number of bedrooms), only RF and LSTM were utilized since they yielded better performance metrics than Prophet did. All the metrics demonstrated in this section were calculated over a year for the specified forecast horizons (Tables 4–9).

Table 4. AC-included: Comparison of metrics across different forecast horizons for different approaches. The lowest RMSE achieved across the different models is highlighted by an arrow (→).

Category		**AC-Included**					
Forecast Horizon		**Day Ahead**			**Week Ahead**		
Metrics/Comparison of Different Algorithms		**Prophet**	**RF**	**LSTM**	**Prophet**	**RF**	**LSTM**
Overall	RMSE	80.02	62.78	→ **32.22**	90.21	75.88	→ **48.67**
	MAE	63.09	45.40	25.85	72.91	55.77	37.28
	MAPE	9.77	5.98	→ **3.92**	11.46	7.65	→ **5.04**
	Percentage of instances having Over-Forecasts	77.72	47.29	73.26	78.15	51.94	44.40
	Percentage of instances having Under-Forecasts	22.28	52.71	26.74	21.85	48.06	55.60
Over-Forecasts	Percentage of Over-Forecasts near Actual value	73.09	86.51	64.66	73.06	90.66	78.67
	Percentage of Over-Forecasts intermediate range near Actual	26.62	13.42	34.65	26.73	9.23	20.77
	Percentage of Over-Forecasts well-above Actual value	0.29	0.07	0.69	0.20	0.11	0.56
	Max-Deviation Percentage from Actual value	44.44	33.39	→ **10.68**	42.24	44.02	→ **19.41**
	Min-Deviation Percentage from Actual value	0.00	0.00	0.00	0.00	0.00	0.00
Under-Forecasts	Percentage of Under-Forecasts near Actual value	85.91	83.24	77.14	67.82	70.81	73.84
	Percentage of Under-Forecasts intermediate range near Actual	13.73	16.55	22.69	30.46	27.89	25.93
	Percentage of Under-Forecasts well-below Actual value	0.36	0.22	0.17	1.72	1.31	0.23
	Max-Deviation Percentage from Actual value	23.37	19.80	→ **15.98**	20.77	31.39	→ **20.32**
	Min-Deviation Percentage from Actual value	0.00	0.00	0.00	0.00	0.00	0.00

Bold numbers represent best results.

Table 5. AC-excluded: Comparison of metrics across different forecast horizons for approaches. The lowest RMSE achieved across the different models is highlighted by an arrow (→).

Category		**AC-Excluded**					
Forecast Horizon		**Day Ahead**			**Week Ahead**		
Metrics/Comparison of Different Algorithms		**Prophet**	**RF**	**LSTM**	**Prophet**	**RF**	**LSTM**
Overall	RMSE	5.61	5.50	→ **3.37**	6.27	5.65	→ **5.44**
	MAE	4.42	4.30	2.61	4.89	4.35	4.98
	MAPE	8.70	8.28	4.93	9.72	8.26	9.56
	Percentage of instances having Over-Forecasts	61.29	50.94	45.00	65.94	47.60	1.64
	Percentage of instances having Under-Forecasts	38.71	49.06	55.00	34.06	52.40	98.36
Over-Forecasts	Percentage of Over-Forecasts near Actual value	75.21	67.97	64.83	66.52	68.85	62.94
	Percentage of Over-Forecasts intermediate range near Actual	24.70	31.65	33.96	33.14	30.82	32.87
	Percentage of Over-Forecasts well-above Actual value	0.09	0.38	1.21	0.35	0.34	4.20
	Max-Deviation Percentage from Actual value	75.96	54.12	→ **21.06**	99.72	45.60	→ **11.46**
	Min-Deviation Percentage from Actual value	0.00	0.00	0.00	0.00	0.00	0.02
Under-Forecasts	Percentage of Under-Forecasts near Actual value	75.55	78.41	75.94	84.38	81.87	27.16
	Percentage of Under-Forecasts intermediate range near Actual	23.98	21.45	23.65	15.42	17.95	72.67
	Percentage of Under-Forecasts well-below Actual value	0.47	0.14	0.41	0.20	0.17	0.17
	Max-Deviation Percentage from Actual value	31.13	35.71	→ **26.08**	35.55	37.35	→ **32.05**
	Min-Deviation Percentage from Actual value	0.00	0.01	0.00	0.00	0.00	0.01

Bold numbers represent best results.

Table 6. AC-included Built-up area subcategory: Comparison of day-ahead forecast metrics across day-ahead forecasts between the LSTM and RF approaches. The lowest RMSE achieved across the different models is highlighted by an arrow (→).

Category		**AC-Included**											
Approach		**LSTM (Day-Ahead Forecasting)**						**RF (Day-Ahead Forecasting)**					
Metrics/Areas (m^2)		**0–50**	**51–100**	**100–150**	**151–200**	**201–300**	**>300**	**0–50**	**51–100**	**100–150**	**151–200**	**201–300**	**>300**
Overall	RMSE	→ **4.70**	→ **12.16**	→ **37.50**	→ **16.09**	→ **5.85**	→ **1.87**	12.34	64.16	74.99	37.02	16.53	11.86
	MAE	4.36	9.88	30.11	13.51	4.56	1.75	7.88	36.52	42.21	22.03	10.65	7.98
	MAPE	17.23	7.15	11.87	13.69	9.65	8.89	30.61	18.40	17.78	19.77	24.20	36.15
	% of instances having Over-Forecasts	3.56	53.15	19.31	27.22	76.55	4.51	54.26	56.45	54.50	56.69	58.33	63.92
	% of instances having Under-Forecasts	96.44	46.85	80.69	72.78	23.45	95.49	45.74	43.55	45.50	43.31	41.67	36.08
Over-Forecasts	% of Over-Forecasts near Actual value	51.45	57.92	55.54	49.33	72.61	46.45	84.39	89.50	86.82	90.84	85.91	87.03
	% of Over-Forecasts intermediate range near Actual	45.34	40.14	41.73	48.02	27.35	43.40	15.27	10.27	12.53	9.02	13.82	12.81
	% of Over-Forecasts well-above Actual value	3.22	1.94	2.73	2.65	0.04	10.15	0.34	0.22	0.65	0.14	0.27	0.16
	Max-Deviation % from Actual value	→ **113.30**	→ **81.47**	→ **45.58**	→ **61.60**	→ **11.63**	→ **72.66**	175.44	129.84	132.63	158.77	134.42	351.31
	Min-Deviation % from Actual value	0.11	0.00	0.00	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.04
Under-Forecasts	% of Under-Forecasts near Actual value	12.25	56.73	41.67	35.57	78.72	18.47	92.36	92.98	92.55	93.67	93.15	89.31
	% of Under-Forecasts intermediate range near Actual	83.74	42.34	56.28	61.09	21.13	81.17	7.44	6.89	7.23	6.17	6.74	10.47
	% of Under-Forecasts well-below Actual value	4.01	0.93	2.04	3.33	0.15	0.36	0.20	0.13	0.23	0.16	0.11	0.22
	Max-Deviation % from Actual value	→ **9.45**	→ **10.96**	→ **17.63**	→ **11.87**	→ **14.23**	→ **4.17**	74.00	68.07	57.34	53.57	58.12	55.70
	Min-Deviation % from Actual value	0.00	0.00	0.02	0.00	0.00	0.03	0.03	0.03	0.00	0.01	0.00	0.01

Bold numbers represent best results.

Table 7. AC-excluded Built-up area subcategory: Comparison of day-ahead forecast metrics across day-ahead forecasts between the LSTM and RF approaches. The lowest RMSE achieved across the different models is highlighted by an arrow (→).

Category		**AC-Excluded**											
Approach		**LSTM (Day-Ahead Forecasting)**						**RF (Day-Ahead Forecasting)**					
Metrics/Areas (in m^2)		**0–50**	**51–100**	**100–150**	**151–200**	**201–300**	**>300**	**0–50**	**51–100**	**100–150**	**151–200**	**201–300**	**>300**
Overall	RMSE	→ **0.14**	→ **0.18**	→ **0.85**	→ **0.21**	→ **0.56**	→ **0.08**	6.02	4.32	7.72	3.56	1.61	1.58
	MAE	0.13	0.13	0.84	0.20	0.55	0.06	3.71	2.68	4.73	2.38	1.12	1.09
	MAPE	2.31	1.59	4.08	2.50	17.38	2.12	22.79	25.44	20.13	26.13	38.08	38.99
	% of instances having Over-Forecasts	55.49	26.97	0.35	99.79	0.85	11.19	58.68	60.89	56.47	61.85	61.20	70.29
	% of instances having Under-Forecasts	44.51	73.03	99.65	0.21	99.15	88.81	41.32	39.11	43.53	38.15	38.80	29.71
Over-Forecasts	% of Over-Forecasts near Actual value	77.91	94.86	48.39	69.51	14.86	86.30	86.05	87.53	84.56	88.83	76.31	97.58
	% of Over-Forecasts intermediate range near Actual	21.92	4.58	41.94	30.45	82.43	13.39	13.68	12.19	15.16	11.13	23.35	2.40
	% of Over-Forecasts well-above Actual value	0.17	0.55	9.68	0.03	2.70	0.31	0.28	0.28	0.28	0.04	0.34	0.02
	Max-Deviation Percentage from Actual value	→ **22.33**	→ **2.69**	→ **1.88**	→ **2.07**	→ **97.93**	→ **42.59**	253.42	287.10	219.73	196.32	315.93	290.16
	Min-Deviation Percentage from Actual value	0.01	0.00	0.02	0.01	0.13	0.00	0.01	0.00	0.01	0.01	0.00	0.00
Under-Forecasts	% of Under-Forecasts near Actual value	59.59	63.08	1.21	38.89	0.42	80.29	92.34	93.67	91.00	94.23	93.12	95.43
	% of Under-Forecasts intermediate range near Actual	34.83	36.81	97.67	50.00	87.07	19.65	7.42	6.13	8.71	5.66	6.65	4.42
	% of Under-Forecasts well-below Actual value	5.58	0.11	1.13	11.11	12.51	0.05	0.25	0.20	0.29	0.12	0.24	0.15
	Max-Deviation Percentage from Actual value	→ **1.70**	→ **6.26**	→ **3.06**	→ **0.27**	→ **24.49**	→ **3.13**	75.61	76.26	74.95	79.51	82.29	76.83
	Min-Deviation Percentage from Actual value	0.00	0.00	0.01	0.00	3.56	0.00	0.01	0.00	0.01	0.00	0.01	0.01

Bold numbers represent best results.

3.1. Trend and Seasonal Profiles

Figure 4 shows the trends and daily, weekly, and yearly seasonal component profiles for the 'AC-included' and 'AC-excluded' categories. The 'AC-included' category shows a rising quadratic trend. In contrast, the 'AC-excluded' category shows a close to linear trend, as shown in Figure 4 top-left and top-right, respectively. However, their weekly profiles are similar but at different scales, where consumption is lower during weekends than on weekdays.

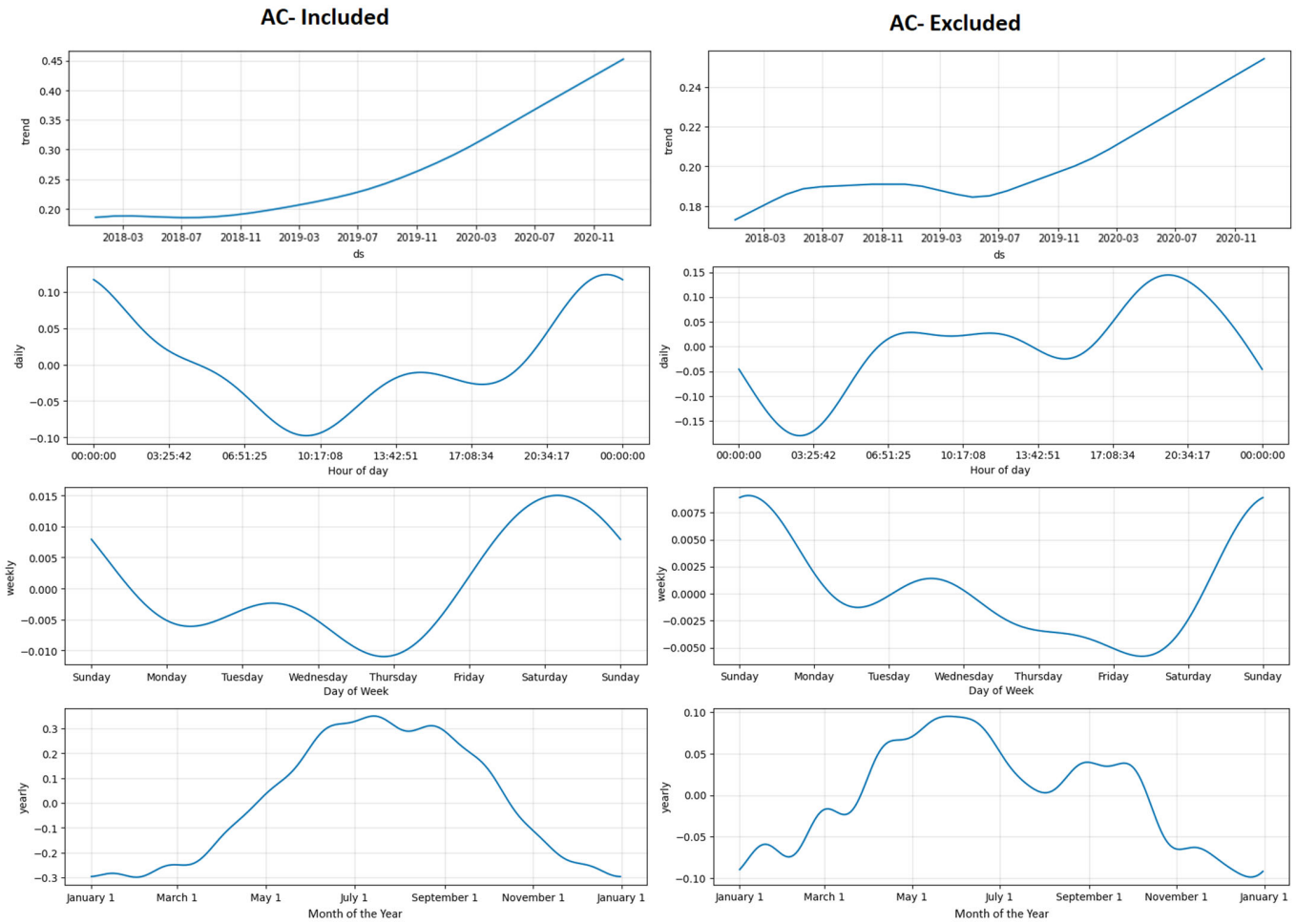

Figure 4. Trend and seasonal profiles of 'AC-included' and 'AC-excluded' category data (note that the vertical scales in 'AC-included' and 'AC-excluded' are different from each other).

The daily profile correlates with the temperature and occupancy in the 'AC-included' category. Since people tend to be out during the daytime, consumption during that period is lower.

As the temperature increases, consumption gradually increases, and, when people return home in the evening, consumption peaks. With respect to the yearly profile, the effect of temperature can be observed starting in April and peaking in June.

The temperature usually peaks in July and August in Dubai, but people (most of whom are expats) also tend to go on vacation during this time, thus reducing the energy demand. Therefore, consumption decreased slightly in those months.

Figure 4 also shows a similar dip in the 'AC-excluded' category. The overall results reinforce the initial idea of studying these two categories separately since, as the results show, they have distinct profiles and very different consumption levels. It is also important to note that the negative values in the seasonal component plot do not necessarily signify

negative consumption. Negative values indicate periods where the observed values are lower than the overall trend or the seasonal average.

3.2. Cooling Categories

The authors named the cooling categories 'AC-included' and 'AC-excluded'. The 'AC-included' demand usually correlates positively with the ambient temperature and occupancy. In the case of 'AC-excluded', there is a tendency for more irregularity since the load is affected mainly by occupancy.

Figure 5 presents day-ahead forecast comparison plots for both categories using Prophet, RF, and LSTM. In the AC-included case, the RF and Prophet models followed the general trend of the observed data, although they missed finer variations in demand.

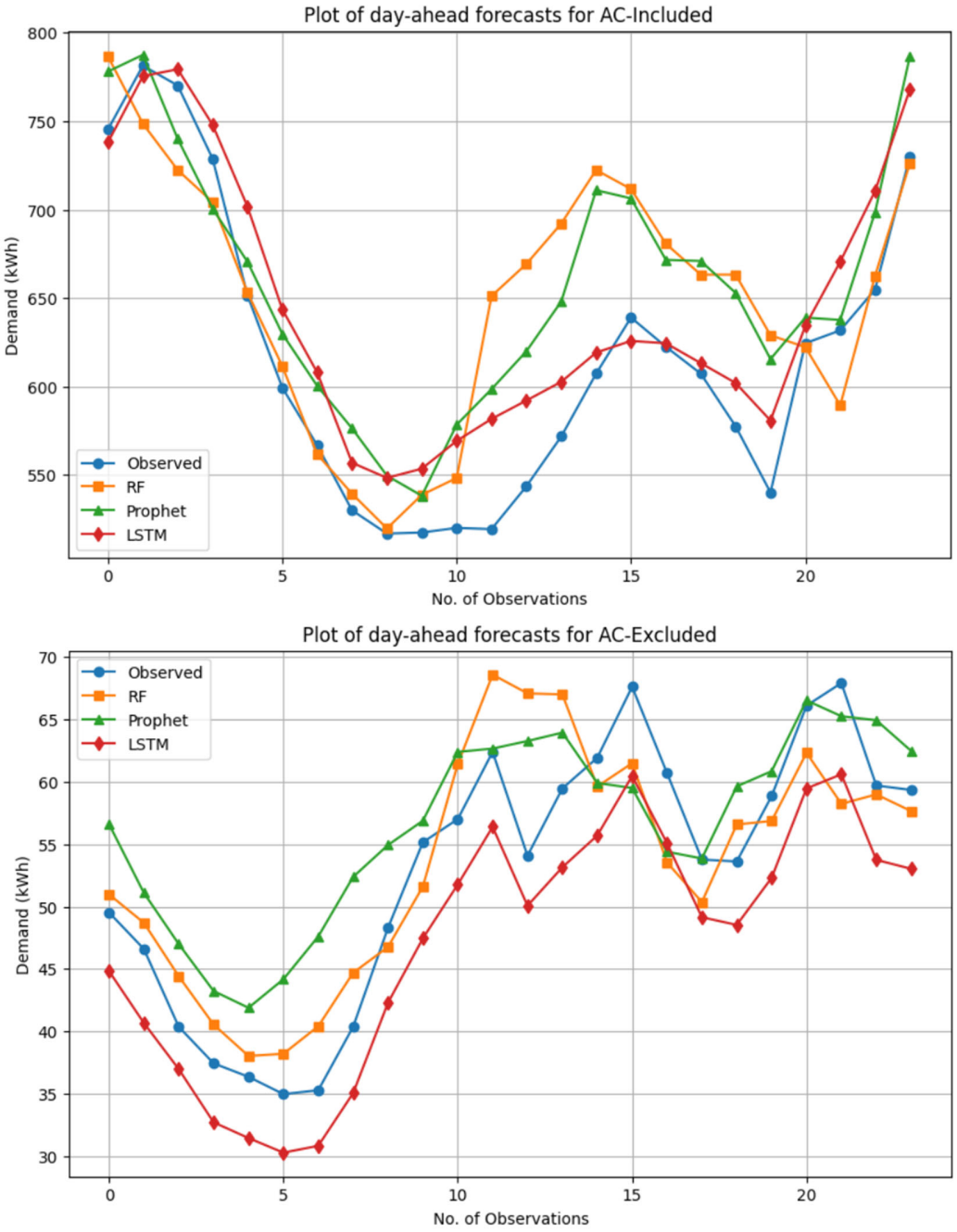

Figure 5. Comparison of Prophet, RF, and LSTM algorithms for 1-day-ahead forecasts for AC-included (**top**) and AC-excluded (**bottom**) categories.

The LSTM model captured short-term fluctuations more closely than the other models did, particularly during peak periods, suggesting that it may be better suited for capturing short-term dependencies in these data. All the models show some lag or underprediction during peaks, which could indicate a challenge in accurately predicting sudden demand spikes in the AC-Included category.

For the AC-included category, the overall demand was lower, and the fluctuations were less intense than those in the AC-included category. The RF model followed the trend closely and had lower deviations than the Prophet model. The LSTM model exhibited some degree of underprediction during demand peaks but performed slightly better than Prophet in capturing the structure of the demand curve. Prophet is overestimated or underestimated in certain sections, indicating that it may not be as effective as RF and LSTM for this category.

Figure 6 shows a week-ahead forecast plot for both AC categories. The LSTM model again demonstrated a close fit to the actual data, capturing demand fluctuations more accurately than Prophet and the RF. The LSTM model appeared to adapt well to the complex temporal dependencies in the AC-included category and smoother patterns in the AC-excluded category, making it suitable for capturing short- and medium-term patterns over the week. RF also performed relatively well, especially in week-ahead forecasts for the AC-included category. It generally captured the trend and seasonal components but underpredicted the peaks. Overall, Figures 5 and 6 highlighted the suitability of each model for different forecasting tasks, with LSTM generally excelling in scenarios requiring more complex temporal modeling, such as day-ahead and week-ahead load forecasting.

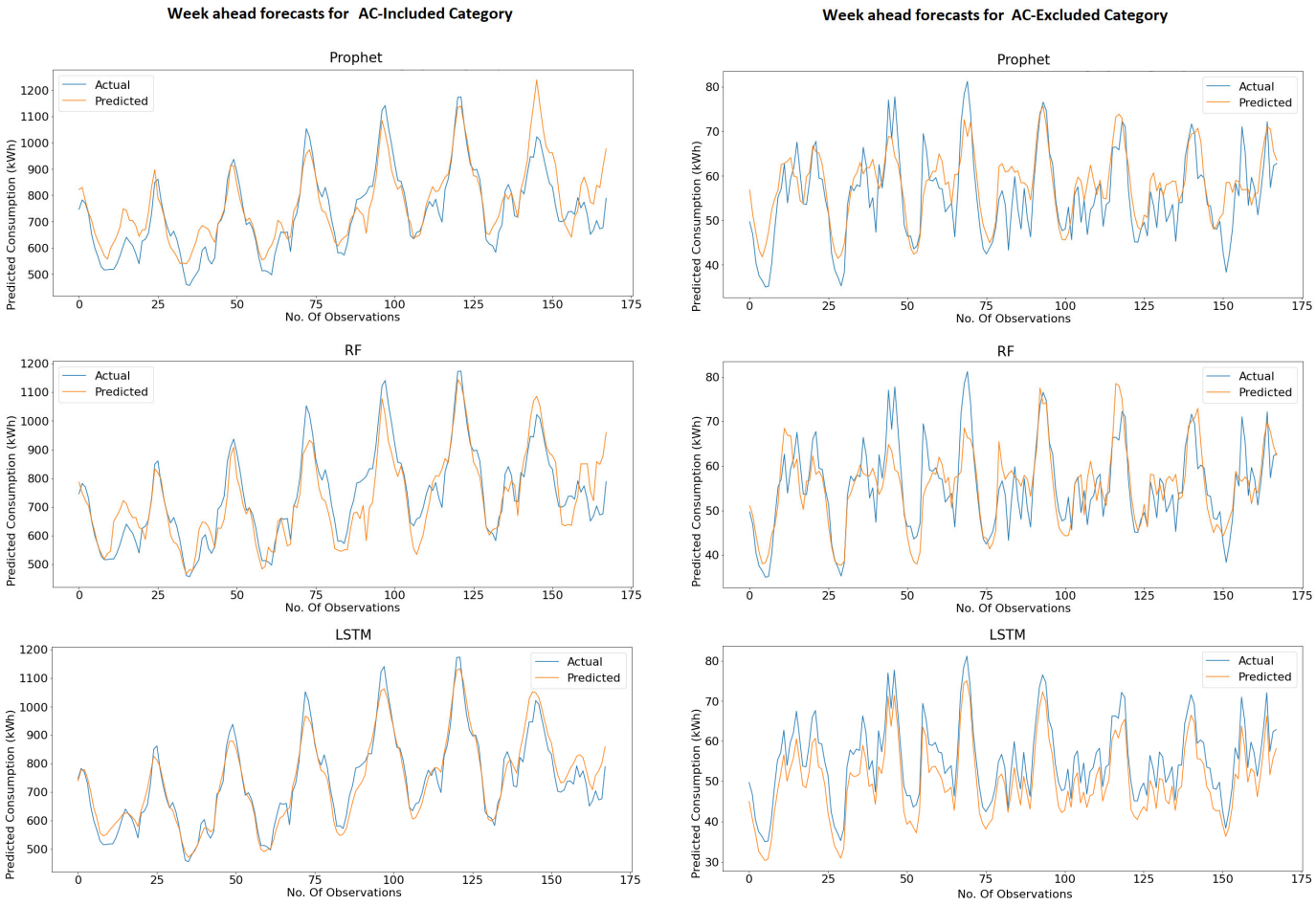

Figure 6. Week-ahead forecasts for the 'AC-included' (**left**) and 'AC-excluded' (**right**) categories using Prophet (**top**), RF (**middle**), and LSTM (**bottom**) algorithms.

Prophet is a time series approach, so the modeling of the daily seasonal component is visible in its week-ahead forecasting plot. However, since RF and LSTM are more data-centric black-box approaches, such seasonal components are not visible in their forecasting plots.

3.2.1. 'AC-Included' Category

Table 4 compares the established and proposed performance metrics for the Prophet, RF, and LSTM results, specifically for the "AC-included" category in every forecast horizon. The established metrics, such as the RMSE and MAPE, indicate that Prophet performed lower than RF and LSTM did. These metrics also suggest that the LSTM outperformed the other two approaches mentioned.

The model was selected based on the least RMSE. Then, the proposed metrics were extracted. The lowest RMSE achieved across the different models is highlighted in bold in the following tables. Similarly, the least maximum deviation among several approaches, which indicates the preciseness of the models, is also highlighted in the following tables. The proposed metrics show over-/under-forecasts by the selected approach and maximum and minimum deviation percentages from the actual values. The higher the percentage of forecasts (i.e., under-/over-forecasts) near the actual value is, the more precise the estimates.

For the 'AC-included' category, the maximum deviation of the prophet algorithm is approximately 44% for the day-ahead and 42% for the week-ahead forecast horizons. Additionally, most of the time, the Prophet algorithm is over-forecasting with an approximately 80% distribution in all three horizons. The distribution of RF is approximately 50% at all three forecast horizons, whereas LSTM tends to over-forecast with approximately 75% for day-ahead forecasts. On the other hand, the LSTM under-forecast for a week-ahead forecast with approximately 55%. However, the maximum deviation for the RF is approximately 33% for day-ahead forecasts and approximately 44% for week-ahead forecasts. Moreover, the maximum deviation for LSTM is approximately 10% for day-ahead forecasts and 20% for week-ahead forecasts. Therefore, LSTM performed better in this scenario.

3.2.2. 'AC-Excluded' Category

The information in Table 5 is similar to that in Table 4 but in relation to the 'AC-excluded' category. Prophet again showed the least performance in terms of the RMSE and MAPE, and LSTM outperformed the Prophet and RF approaches.

For the 'AC-excluded' category, the maximum deviation of the prophet algorithm is approximately 75% for day-ahead horizons and approximately 100% for week-ahead horizons. Additionally, most of the time, the prophet algorithm is over-forecasting with an approximately 60% distribution on its side in all three horizons. The distribution of the RF is approximately 50% in the two forecast horizons, whereas the LSTM tends to under-forecast with approximately 55% for day-ahead forecasts. LSTM significantly under-forecasts in its week-ahead forecasts with values of approximately 98%. However, the percentage distribution of LSTM near the maximum deviation is low, with values of approximately 0.17% and 0.67%. Additionally, the maximum deviation for the RF is approximately 54% for day-ahead forecasts, and approximately 45% for week-ahead forecasts. Moreover, the maximum deviation for the LSTM is approximately 26% for day-ahead forecasts, and 32% for week-ahead forecasts.

3.3. Built-Up Area

Since the RF and LSTM models outperformed the Prophet method in the main cooling categories, only the performances of RF and LSTM models are discussed here for the built-up area's subcategory for day-ahead forecasting. Tables 6 and 7 show that LSTM outperformed the RF based on the lowest RMSE and MAPE for each built-up area size for the 'AC-included' and 'AC-excluded' categories. LSTM outperforms RF by having an RMSE that is less than three times lower and achieving a MAPE of less than half in most built-up area subcategories. Table 6 shows that the LSTM model outperformed the RF model for each area-type apartment in the 'AC-excluded' categories. LSTM outperforms RF since its RMSE and MAPE values are ten times lower in most built-up area subcategories.

3.4. Bedroom Number

Although the LSTM outperformed the other approaches in the previous evaluation comparison (Sections 3.2.1, 3.2.2 and 3.3), the RF outperformed LSTM in the 'bedroom-number' subcategory for both the 'AC-included' and 'AC-excluded' categories concerning the RMSE and MAPE. Table 7 shows that the MAPE value for the LSTM model is almost twice that of the RF model. The resulting error in forecasting this subcategory is relatively large due to the high variance in the 'bedrooms-number' category data. Table 7 also shows that the modeled LSTM is only slightly inclined towards under-forecasts for the' AC-included' category. The percentage of maximum deviation for over-forecasts is relatively lower, approximately 70% in most cases for the LSTM approach. In contrast, the modeled RF seems inclined towards over-forecasts, with most instances, approximately 90%, having values close to the actual values. RF obtained the best evaluation since the total percentage of instances near the actual value for both under-forecasts and over-forecasts was higher than that of the LSTM. Table 8 shows that the modeled LSTM is inclined towards over-forecasts where the maximum deviation is close to or greater than 200%. At the same time, the modeled RF (unlike the AC category) is also inclined towards over-forecasts, with a maximum deviation slightly less than that of the LSTM. Since the total percentage of instances near the actual value for both under-forecasts and over-forecasts in the case of the RF is better, providing the RF with better evaluation. Thus, the RF performed better than the LSTM did in both AC categories.

This proposed experimental work shows that on the occasions in which the lowest RMSE is observed across various comparable approaches, the maximum deviation percentage from the actual value (i.e., Q3 of the differenced bias) from the proposed metrics is also the lowest for both over-/under-forecasting. Thus, the experimental results show that the best model selected based on RMSE tends to have the best fit for the predicted values as the percentage of highly deviated values (also referred to as predicted outliers) is lower.

Like the 'AC-included' category, Table 9 shows that even for the 'AC-excluded' category, the MAPE is almost twice for LSTM compared with that of the RF. The percentage of instances for over-forecasts is approximately 95% for LSTM modeling apartments with two and three bedrooms. In contrast, the percentage of over-forecast instances for RF modeling is close to 60% for the same-mentioned 'bedroom-number', which is slightly greater.

Table 8. AC-included Bedroom-number subcategory: Comparison of day-ahead forecast metrics between the LSTM and RF approaches.

Category		**AC-Included**							
Approach		**LSTM (Day-Ahead Forecasting)**				**RF (Day-Ahead Forecasting)**			
Metrics/Bedroom-Type		**0 Studio**	**1 Bedroom**	**2 Bedrooms**	**3 Bedrooms**	**0 Studio**	**1 Bedroom**	**2 Bedrooms**	**3 Bedrooms**
Overall	RMSE	68.33	186.53	41.88	15.83	→ **57.05**	→ **179.09**	→ **36.94**	→ **8.15**
	MAE	43.50	109.37	27.82	10.28	26.97	82.95	18.01	4.25
	MAPE	23.14	16.71	22.78	31.13	13.02	12.46	12.94	14.83
	% of instances having Over-Forecasts	43.41	42.10	72.44	19.69	73.44	71.96	74.54	80.31
	% of instances having under-Forecasts	56.59	57.90	27.56	80.31	26.56	28.04	25.46	19.69
Over-Forecasts	% of Over-Forecasts near Actual value	83.87	80.51	79.82	84.30	→ **91.42**	→ **90.42**	→ **90.52**	→ **93.76**
	% of Over-Forecasts intermediate range near Actual	15.53	18.73	19.89	14.94	8.05	9.31	9.04	6.11
	% of Over-Forecasts well-above Actual value	0.61	0.76	0.28	0.76	0.53	0.27	0.44	0.13
	Max-Deviation % from Actual value	131.23	139.72	96.44	97.98	→ **98.94**	→ **91.96**	→ **87.59**	→ **76.96**
	Min-Deviation % from Actual value	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Under-Forecasts	% of Under-Forecasts near Actual value	90.51	93.40	88.87	91.58	88.14	88.93	84.93	79.01
	% of Under-Forecasts intermediate range near Actual	9.32	6.50	10.59	8.37	11.65	10.79	14.84	20.70
	% of Under-Forecasts well-below Actual value	0.16	0.10	0.54	0.06	0.21	0.29	0.22	0.29
	Max-Deviation % from Actual value	72.75	69.52	→ **45.95**	73.44	→ **65.61**	→ **66.22**	51.92	→ **46.43**
	Min-Deviation % from Actual value	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00

Bold numbers represent best results.

Table 9. AC-excluded Bedroom-number subcategory: comparison of day-ahead forecast metrics between the LSTM and RF approaches.

Category		**AC-Excluded**							
Approach		**LSTM (Day-Ahead Forecasting)**				**RF (Day-Ahead Forecasting)**			
Metrics/Bedroom-Type		**0 Studio**	**1 Bedroom**	**2 Bedrooms**	**3 Bedrooms**	**0 Studio**	**1 Bedroom**	**2 Bedrooms**	**3 Bedrooms**
Overall	RMSE	**6.38**	**11.45**	**7.86**	**2.37**	→ **4.57**	→ **10.86**	→ **5.29**	→ **1.61**
	MAE	4.60	7.65	6.07	2.03	3.02	6.89	3.46	1.12
	MAPE	42.71	20.45	52.86	65.80	26.07	20.27	26.49	37.90
	% of instances having Over-Forecasts	78.28	17.10	93.51	95.64	66.34	50.24	65.81	61.00
	% of instances having Under-Forecasts	21.72	82.90	6.49	4.36	33.66	49.76	34.19	39.00
Over-Forecasts	% of Over-Forecasts near Actual value	65.36	81.93	87.29	60.29	80.24	79.41	84.79	75.62
	% of Over-Forecasts interm. range near Actual	34.22	17.14	12.59	39.47	19.38	20.06	14.85	23.88
	% of Over-Forecasts well-above Actual value	0.42	0.94	0.11	0.24	0.38	0.52	0.36	0.51
	Max-Deviation % from Actual value	205.08	→ **78.47**	265.17	318.93	→ **140.24**	158.34	→ **188.43**	→ **308.73**
	Min-Deviation % from Actual value	0.01	0.01	0.06	0.18	0.01	0.00	0.01	0.00
Under-Forecasts	% of Under-Forecasts near Actual value	90.46	88.91	75.49	75.07	90.23	89.33	93.42	93.18
	% of Under-Forecasts interm. range near Actual	8.96	10.85	23.28	23.88	9.29	10.30	6.51	6.62
	% of Under-Forecasts well-below Actual value	0.58	0.23	1.23	1.05	0.47	0.37	0.07	0.20
	Max-Deviation % from Actual value	→ **49.18**	→ **65.00**	→ **48.90**	→ **49.89**	69.17	67.52	75.87	82.41
	Min-Deviation % from Actual value	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01

Bold numbers represent best results.

3.5. Significance of the Proposed EMPQ Metrics

Quantifying under- and over-forecasting errors in electricity load forecasting has significant implications for forecasting accuracy, the applicability of different forecasting models, and the potential for improving demand predictions in electricity systems. Given that electricity demand is highly variable, influenced by numerous factors (e.g., temperature, holidays, and economic activity), and subject to predictable and unpredictable events, understanding and addressing the impacts of under- and over-forecasting is crucial for effective grid management, resource planning, and operational efficiency. In time series load forecasting, metrics such as EMPQ show several significant benefits by considering details about under-forecasting and over-forecasting.

The proposed EMPQ provide valuable insights that were not fully captured by traditional metrics like MAPE and RMSE. In Table 4, while LSTM achieved the lowest RMSE and MAPE for both day-ahead and week-ahead forecasts in the "AC-included" category, the EMPQ metrics revealed high percentages of over- and under-forecasts, indicating deviations from actual values that traditional metrics overlooked. Similarly, in Table 5, where RF and LSTM had similar MAPE and RMSE values in the "AC-excluded" week-ahead forecasting, EMPQ metrics identified LSTM as the better model due to its lower maximum deviation percentage. Furthermore, Table 8 shows a case where the RMSE and MAPE favored the RF in the "bedroom-number" subcategory under "AC-included" conditions, yet the EMPQ metrics revealed that the LSTM model had a higher percentage of forecasts close to the actual demand values, suggesting that it was better at capturing demand fluctuations. These cases highlight the importance of the proposed EMPQ framework in providing a more comprehensive and informative evaluation of model performance.

3.5.1. Balanced Performance Assessment

EMPQ metrics help to distinguish between under-forecasting and over-forecasting, which enables the identification of systematic biases in the forecasting model. For example, in Table 7 (AC-included category), the LSTM model exhibited favorable RMSE and MAPE values. However, EMPQ metrics revealed a significant bias, with 73.26% over-forecasts for day-ahead predictions, suggesting that while LSTM minimized overall error, it consistently overestimated demand. This insight highlights potential systematic biases in LSTM predictions, which would be overlooked by the RMSE and MAPE alone. Such detailed metrics allow forecasters to adjust models and create a more balanced, reliable forecast by specifically addressing these biases.

3.5.2. Model Improvement and Selection

EMPQ metrics also serve as a diagnostic tool for model improvement and selection, clarifying model performance beyond traditional error measures. In Table 8 (AC-excluded category, week-ahead forecast), both the RF and LSTM models had similar MAPE and RMSE values, but the EMPQ metrics showed LSTM's superior performance with a much lower maximum deviation percentage (11.46% compared to RF's 45.60%). This result indicates that LSTM's forecasts were closer to the actual values despite similar traditional metrics, which helps identify the LSTM as the more accurate model under specific conditions. Additionally, Table 8 (bedroom-number subcategory in AC-included) showed that traditional metrics favored RF over LSTM, yet EMPQ metrics revealed the LSTM's higher percentage of forecasts near actual values, indicating better demand fluctuation captured by LSTM. This further underscores the value of the EMPQ in nuanced model selection and tuning.

3.5.3. Enhanced Decision Making and Risk Mitigation

With respect to proper resource allocation in businesses, particularly those in energy or utilities, knowing whether forecasts tend to be over or under can significantly impact resource planning. For example, in Table 5 (AC-excluded category, week-ahead forecast), Random Forest showed high over-forecast percentages (up to 45.60%) compared with the

LSTM. This tendency toward over-forecasting by RF can lead to inefficient planning and operation due to over-provisioning and increased costs. On the other hand, LSTM demonstrated a more balanced approach, making it suitable for situations where minimizing the forecast deviation is crucial. Thus, understanding the direction of forecast errors helps in risk management.

3.5.4. Economic Impact

Different forecasting errors have different cost implications. Under-forecasting might lead to emergency purchases or ramp-up costs, whereas over-forecasting results in higher operational costs. For example, Table 4 highlights that LSTM had a 10% maximum deviation in day-ahead forecasts under the AC-included category, whereas Prophet showed deviations as high as 44% in some cases. The lower deviation of the LSTM suggests it could minimize the economic impact by reducing the potential penalties associated with under-forecasting, which can lead to emergency ramp-up costs. Similarly, understanding Prophet's high deviation rate allows for preemptive measures to mitigate these costs. Thus, quantifying these errors helps calculate potential economic impacts more accurately and makes cost-effective decisions.

Additionally, in many regulated industries, there are penalties associated with forecasting inaccuracies. Distinguishing between under- and over-forecasting can aid in minimizing these penalties by addressing the specific causes of forecast deviations.

3.5.5. Operational Efficiency

The proposed EMPQ metrics facilitate continuous monitoring and improvement of forecasting models by identifying patterns of under- and over-forecasting, which supports real-time adjustments and enhances operational efficiency. For example, in Table 4, the LSTM model for day-ahead forecasts in the AC-included category showed lower overall RMSE but exhibited a tendency toward over-forecasting (73.26%) compared to other models. This information allows organizations to implement corrective measures if they observe a consistent trend toward over-forecasting, minimizing unnecessary energy allocation and resource costs. Conversely, Table 5 (week-ahead forecast for AC-excluded) shows Random Forest's higher maximum deviation (45.60%) in over-forecasting, highlighting the need for adjustment if under-utilized resources become a recurring issue.

These EMPQ metrics can also support a feedback loop for continuous learning and improvement. As organizations track and analyze these error patterns, they can iteratively refine models to capture load dynamics better and adjust to changing conditions. This approach ensures that forecasting models evolve in response to observed errors, leading to improved demand prediction over time.

Quantifying under- and over-forecasting errors in electricity load forecasting has significant implications for forecasting accuracy, model applicability, and efforts to optimize demand predictions. By understanding the asymmetry of forecasting errors, organizations can do the following:

- Improve model accuracy through tailored metrics that better capture the operational impacts of forecast deviations.
- Select and adapt models that accommodate system constraints and respond dynamically to shifts in electricity demand.
- Implement adaptive, real-time forecasting techniques and integrate external factors to strengthen predictive capabilities, ultimately enhancing the resilience, efficiency, and cost-effectiveness of the electricity grid.

4. Conclusions

This study presents a novel approach to evaluating forecasting models for residential load prediction by introducing the EMPQ, new metrics that provide a deeper understanding of model performance beyond the established ones. The EMPQ framework reveals

crucial insights into over- and under-forecasting tendencies, which are often overlooked in conventional assessments such as RMSE or MAPE.

To demonstrate the validity of the proposed evaluation metrics, the authors applied them to evaluate the performance of three forecasting models for short-term electricity demand forecasting among residential customers of the Dubai Electricity and Water Authority. The residential customers were grouped according to whether their cooling demand was captured by electricity smart meters. The analysis considered other variables, including temperature, dwelling size, and number of bedrooms.

The assessment of the new performance metrics revealed that they were crucial in determining that the LSTM model outperforms both the Prophet and Random Forest models in capturing demand fluctuations. LSTM exhibited lower maximum deviation percentages, and its forecasts were more aligned with actual demand, establishing it as the most accurate model in this study. This finding highlights the importance of adopting a more comprehensive evaluation strategy, such as the EMPQ algorithm, for comparing forecasting models, especially in contexts such as smart grid operations where accuracy and bias control are paramount.

While the proposed new metrics offer valuable performance insights, they also introduce increased complexity in model comparison and may challenge interpretability in certain contexts. Future research could focus on extending the EMPQ framework to longer-term forecasts and diverse geographical conditions, as well as refining these metrics to enhance usability and interpretability. Additionally, future research could explore various forecasting models developed for various forecast horizons and temporal resolutions and evaluate them using the proposed metrics. By applying these aspects, EMPQ could become a more universally applicable tool for evaluating load forecasts in smart grids, ultimately guiding more informed decision-making in energy management and for professionals responsible for demand-related initiatives.

Author Contributions: Conceptualization, P.M., H.R. and E.R.-U.; methodology, P.M. and H.R.; formal analysis, P.M.; investigation, P.M. and H.R.; resources, P.M. and H.R.; data curation, P.M.; writing—original draft preparation, P.M., H.R. and E.R.-U.; writing—review and editing, E.R.-U. and T.P.; visualization, P.M. and E.R.-U.; supervision, E.R.-U.; project administration, E.R.-U.; funding acquisition, E.R.-U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because of privacy restrictions. The data is property of Dubai Electricity and Water Authority (DEWA).

Acknowledgments: The Dubai Electricity & Water Authority (DEWA) supported this work under the project "Decision Support Tool for Developing Energy Strategies Based on Data-Driven Models".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Judge, M.A.; Khan, A.; Manzoor, A.; Khattak, H.A. Overview of Smart Grid Implementation: Frameworks, Impact, Performance and Challenges. *J. Energy Storage* **2022**, *49*, 104056. [CrossRef]
2. Potdar, V.; Chandan, A.; Batool, S.; Patel, N. Big Energy Data Management for Smart Grids—Issues, Challenges and Recent Developments. In *Smart Cities*; Springer: Cham, Switzerland, 2018; pp. 177–205.
3. Liu, Y.; Wang, G.; Guo, W.; Zhang, Y.; Dong, W.; Guo, W.; Wang, Y.; Zeng, Z.X. Power Data Mining in Smart Grid Environment. *J. Intell. Fuzzy Syst.* **2021**, *40*, 3169–3175. [CrossRef]
4. Kaytez, F.; Taplamacioglu, M.C.; Cam, E.; Hardalac, F. Forecasting Electricity Consumption: A Comparison of Regression Analysis, Neural Networks and Least Squares Support Vector Machines. *Int. J. Electr. Power Energy Syst.* **2015**, *67*, 431–438. [CrossRef]
5. Hammad, M.A.; Jereb, B.; Rosi, B.; Dragan, D. Methods and Models for Electric Load Forecasting: A Comprehensive Review. *Logist. Sustain. Transp.* **2020**, *11*, 51–76. [CrossRef]
6. Yu, S.; Wang, K.; Wei, Y.-M. A Hybrid Self-Adaptive Particle Swarm Optimization–Genetic Algorithm–Radial Basis Function Model for Annual Electricity Demand Prediction. *Energy Convers Manag.* **2015**, *91*, 176–185. [CrossRef]
7. Bedi, J.; Toshniwal, D. Deep Learning Framework to Forecast Electricity Demand. *Appl. Energy* **2019**, *238*, 1312–1326. [CrossRef]

8. Moral-Carcedo, J.; Pérez-García, J. Temperature Effects on Firms' Electricity Demand: An Analysis of Sectorial Differences in Spain. *Appl. Energy* **2015**, *142*, 407–425. [CrossRef]
9. McCulloch, J.; Ignatieva, K. Intra-Day Electricity Demand and Temperature. *Energy J.* **2020**, *41*, 161–182. [CrossRef]
10. Momani, M.A. Factors Affecting Electricity Demand in Jordan. *Energy Power Eng.* **2013**, *05*, 50–58. [CrossRef]
11. Hossein Motlagh, N.; Mohammadrezaei, M.; Hunt, J.; Zakeri, B. Internet of Things (IoT) and the Energy Sector. *Energies* **2020**, *13*, 494. [CrossRef]
12. Csoknyai, T.; Legardeur, J.; Akle, A.A.; Horváth, M. Analysis of Energy Consumption Profiles in Residential Buildings and Impact Assessment of a Serious Game on Occupants' Behavior. *Energy Build.* **2019**, *196*, 1–20. [CrossRef]
13. Serrallés, R.J. Electric Energy Restructuring in the European Union: Integration, Subsidiarity and the Challenge of Harmonization. *Energy Policy* **2006**, *34*, 2542–2551. [CrossRef]
14. Apostoleris, H.; Sgouridis, S.; Stefancich, M.; Chiesa, M. Utility Solar Prices Will Continue to Drop All over the World Even without Subsidies. *Nat. Energy* **2019**, *4*, 833–834. [CrossRef]
15. Manandhar, P.; Rafiq, H.; Rodriguez-Ubinas, E. Current Status, Challenges, and Prospects of Data-Driven Urban Energy Modeling: A Review of Machine Learning Methods. *Energy Rep.* **2023**, *9*, 2757–2776. [CrossRef]
16. Fallah, S.N.; Ganjkhani, M.; Shamshirband, S.; Chau, K.-w. Computational Intelligence on Short-Term Load Forecasting: A Methodological Overview. *Energies* **2019**, *12*, 393. [CrossRef]
17. Hippert, H.S.; Pedreira, C.E.; Souza, R.C. Neural Networks for Short-Term Load Forecasting: A Review and Evaluation. *IEEE Trans. Power Syst.* **2001**, *16*, 44–55. [CrossRef]
18. He, F.; Zhou, J.; Feng, Z.; Liu, G.; Yang, Y. A Hybrid Short-Term Load Forecasting Model Based on Variational Mode Decomposition and Long Short-Term Memory Networks Considering Relevant Factors with Bayesian Optimization Algorithm. *Appl. Energy* **2019**, *237*, 103–116. [CrossRef]
19. Bu, X.; Wu, Q.; Zhou, B.; Li, C. Hybrid Short-Term Load Forecasting Using CGAN with CNN and Semi-Supervised Regression. *Appl. Energy* **2023**, 338. [CrossRef]
20. Chen, J.L.; Li, G.; Wu, D.C.; Shen, S. Forecasting Seasonal Tourism Demand Using a Multiseries Structural Time Series Method. *J. Travel Res.* **2019**, *58*, 92–103. [CrossRef]
21. Khan, I.A.; Akber, A.; Xu, Y. Sliding Window Regression Based Short-Term Load Forecasting of a Multi-Area Power System. In Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering, Edmonton, AB, Canada, 5–8 May 2019.
22. Madhukumar, M.; Sebastian, A.; Liang, X.; Jamil, M.; Shabbir, M.N.S.K. Regression Model-Based Short-Term Load Forecasting for University Campus Load. *IEEE Access* **2022**, *10*, 8891–8905. [CrossRef]
23. Behm, C.; Nolting, L.; Praktiknjo, A. How to Model European Electricity Load Profiles Using Artificial Neural Networks. *Appl. Energy* **2020**, *277*, 115564. [CrossRef]
24. Shi, H.; Xu, M.; Li, R. Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN. *IEEE Trans. Smart Grid* **2017**, *9*, 5271–5280. [CrossRef]
25. Lazzari, F.; Mor, G.; Cipriano, J.; Gabaldon, E.; Grillone, B.; Chemisana, D.; Solsona, F. User Behaviour Models to Forecast Electricity Consumption of Residential Customers Based on Smart Metering Data. *Energy Rep.* **2022**, *8*, 3680–3691. [CrossRef]
26. He, Y.; Deng, J.; Li, H. Short-Term Power Load Forecasting with Deep Belief Network and Copula Models. In Proceedings of the Proceedings-9th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2017.
27. Ouyang, T.; He, Y.; Li, H.; Sun, Z.; Baek, S. Modeling and Forecasting Short-Term Power Load With Copula Model and Deep Belief Network. *IEEE Trans. Emerg. Top. Comput. Intell.* **2019**, *3*, 127–146. [CrossRef]
28. Deng, X.; Ye, A.; Zhong, J.; Xu, D.; Yang, W.; Song, Z.; Zhang, Z.; Guo, J.; Wang, T.; Tian, Y.; et al. Bagging–XGBoost Algorithm Based Extreme Weather Identification and Short-Term Load Forecasting Model. *Energy Rep.* **2022**, *8*, 8661–8674. [CrossRef]
29. Jurado, M.; Samper, M.; Rosés, R. An Improved Encoder-Decoder-Based CNN Model for Probabilistic Short-Term Load and PV Forecasting. *Electr. Power Syst. Res.* **2023**, *217*, 109153. [CrossRef]
30. Wang, S.; Deng, X.; Chen, H.; Shi, Q.; Xu, D. A Bottom-up Short-Term Residential Load Forecasting Approach Based on Appliance Characteristic Analysis and Multi-Task Learning. *Electr. Power Syst. Res.* **2021**, *196*, 107233. [CrossRef]
31. Langevin, A.; Cheriet, M.; Gagnon, G. Efficient Deep Generative Model for Short-Term Household Load Forecasting Using Non-Intrusive Load Monitoring. *Sustain. Energy Grids Netw.* **2023**, *34*, 101006. [CrossRef]
32. Wan, A.; Chang, Q.; AL-Bukhaiti, K.; He, J. Short-Term Power Load Forecasting for Combined Heat and Power Using CNN-LSTM Enhanced by Attention Mechanism. *Energy* **2023**, *282*, 128274. [CrossRef]
33. Wei, N.; Yin, L.; Li, C.; Wang, W.; Qiao, W.; Li, C.; Zeng, F.; Fu, L. Short-Term Load Forecasting Using Detrend Singular Spectrum Fluctuation Analysis. *Energy* **2022**, *256*, 124722. [CrossRef]
34. Li, T.; Qian, Z.; He, T. Short-Term Load Forecasting with Improved CEEMDAN and GWO-Based Multiple Kernel ELM. *Complexity* **2020**, *2020*, 1209547. [CrossRef]
35. Kim, T.Y.; Cho, S.B. Predicting Residential Energy Consumption Using CNN-LSTM Neural Networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
36. Sekhar, C.; Dahiya, R. Robust Framework Based on Hybrid Deep Learning Approach for Short Term Load Forecasting of Building Electricity Demand. *Energy* **2023**, *268*, 126660. [CrossRef]

37. Sadaei, H.J.; de Lima e Silva, P.C.; Guimarães, F.G.; Lee, M.H. Short-Term Load Forecasting by Using a Combined Method of Convolutional Neural Networks and Fuzzy Time Series. *Energy* **2019**, *175*, 365–377. [CrossRef]
38. Ran, P.; Dong, K.; Liu, X.; Wang, J. Short-Term Load Forecasting Based on CEEMDAN and Transformer. *Electr. Power Syst. Res.* **2023**, *214*, 10885. [CrossRef]
39. Tong, C.; Zhang, L.; Li, H.; Ding, Y. Attention-Based Temporal–Spatial Convolutional Network for Ultra-Short-Term Load Forecasting. *Electr. Power Syst. Res.* **2023**, *220*, 109329. [CrossRef]
40. Yang, D.; Guo, J.; Li, Y.; Sun, S.; Wang, S. Short-Term Load Forecasting with an Improved Dynamic Decomposition-Reconstruction-Ensemble Approach. *Energy* **2023**, *263*, 125609. [CrossRef]
41. Robeson, S.M.; Willmott, C.J. Decomposition of the Mean Absolute Error (MAE) into Systematic and Unsystematic Components. *PLoS ONE* **2023**, *18*, e0279774. [CrossRef]
42. Jadon, A.; Patil, A.; Jadon, S. A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting. In *Data Management, Analytics and Innovation*; Springer: Singapore, 2024; pp. 117–147.
43. Rafiq, H.; Manandhar, P.; Rodriguez-Ubinas, E.; Barbosa, J.D.; Qureshi, O.A. Analysis of Residential Electricity Consumption Patterns Utilizing Smart-Meter Data: Dubai as a Case Study. *Energy Build.* **2023**, *291*, 113103. [CrossRef]
44. United Arab Emirates Ministry of Energy and Infrastructure. *UAE State of Energy Report 2019*; Ministry of Energy: Abu Dhabi, United Arab Emirates, 2020; p. 53.
45. Meteostat Dubai International Airport Weather Data. Available online: https://meteostat.net/en/station/41194?t=2018-01-01/2021-12-31 (accessed on 14 February 2024).
46. Manandhar, P.; Rafiq, H.; Rodriguez-Ubinas, E.; Barbosa, J.D.; Qureshi, O.A.; Tarek, M.; Sgouridis, S. Understanding Energy Behavioral Changes Due to COVID-19 in the Residents of Dubai Using Electricity Consumption Data and Their Impacts. *Energies* **2022**, *16*, 285. [CrossRef]
47. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
48. Bendiek, P.; Taha, A.; Abbasi, Q.H.; Barakat, B. Solar Irradiance Forecasting Using a Data-Driven Algorithm and Contextual Optimisation. *Appl. Sci.* **2021**, *12*, 134. [CrossRef]