OPTIMAL WEATHER STATION SELECTION FOR ELECTRIC LOAD FORECASTING

By

Yike Li

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2020

Approved by:

_____
Dr. Tao Hong

_____
Dr. Pu Wang

_____
Dr. Linquan Bai

ABSTRACT

YIKE LI. Optimal Weather Station Selection for Electric Load Forecasting.
(Under the direction of DR. TAO HONG)

Weather factors are playing key impacts on electricity load consumption. The proper selection of weather stations will contribute to the final electric load forecasting accuracy. How to efficiently select the stations among a good number of candidates within the territory of interest remains a pressing issue. This thesis proposes several comprehensive weather station selection (WSS) frameworks along with different statistical tests to determine the stations to be used for electric load forecasting. We demonstrate comprehensive implementation and effectiveness of these methods based on the Global Energy Forecasting Competition 2012 (GEFCom2012) data and Global Energy Forecasting Competition 2014 (GEFCom2014) data are evaluated by comparing to the selection results obtained from the WSS framework introduced in (Hong et al., 2015). We introduce theoretical optimum (TO) selection to unveil what the best WSS looks like given we have access to the future load and from which, we gain further insights on why some WSS frameworks outperform the others. Additionally, we extend our discussion on several practical data fitting issues on the WSS subject and suggest several actionable rules of thumb that load forecasting practitioners can follow. Our experimental results show that the forecasting accuracy can be significantly improved by several proposed selection frameworks. Meanwhile, several heuristic methods have been applied to cut down the computational cost.

DEDICATION

To My Family.

ACKNOWLEDGEMENTS

First of all, my sincere thanks go to Dr. Tao Hong. He is my first mentor to the forecasting field at my graduate school and my advisor during my master's study in the engineering management program at UNC Charlotte. He guided me onto a new area of study and encouraged me to take on challenges and succeed in this area. I would like to express my sincere gratitude to Dr. Hong for his trust and guidance through my master's program.

Secondly, I want to thank the other two committee members, Dr. Pu Wang and Dr. Linquan Bai for their generous offering of time and guidance to greatly improve the presentation of this thesis.

Thirdly, I also want to thank the group members at the Big Data Energy Analytics Laboratory, Deeksha Dharmapal, Shreyashi Shukla, Masoud Sobhani, Zehan Xu, and Saurabh Sangamwar. They provided generous help and support for my research.

Finally, I want to express my deepest love and appreciation to my parents: Yongsheng Li and Siqing Wang, and my girlfriend: Chunmei Chen. This research won’t be accomplished without their encouragement, tolerance, sacrifice, and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AMI	advanced metering infrastructure
BS	backward selection framework
BS-CV	backward selection framework with out-of-sample cross-validation test
BS-IS	backward selection framework with in-sample fit statistical test
BS-PS	backward selection framework with post-sample fit statistical test
CV	out-of-sample cross-validation test
ES	exhaustive search framework
ES-CV	exhaustive search framework with out-of-sample cross-validation test
ES-IS	exhaustive search framework with in-sample fit statistical test
ES-PS	exhaustive search framework with post-sample fit statistical test
FS	forward selection framework
FS-CV	forward selection framework with out-of-sample cross-validation test
FS-IS	forward selection framework with in-sample fit statistical test
FS-PS	forward selection framework with post-sample fit statistical test
GS	greedy selection framework
GS-CV	greedy selection framework with out-of-sample cross-validation test
GS-IS	greedy selection framework with in-sample fit statistical test
GS-PS	greedy selection framework with post-sample fit statistical test
GA	genetic algorithm
GEFCom2012	Global Energy Forecasting Competition 2012
GEFCom2014	Global Energy Forecasting Competition 2014
GEFCom2017	Global Energy Forecasting Competition 2017

IS	in-sample fit statistical test
MAPE	mean absolute percentage error
PS	post-sample fit statistical test
TO	theoretical optimum
Vanilla	Tao's Vanilla Benchmark Model
WSS	weather station selection
VSTLF	very short-term load forecasting
STLF	short-term load forecasting
MTLF	medium-term load forecasting
LTLF	long-term load forecasting

CHAPTER 1: INTRODUCTION

Electricity load forecasting processes have been integrated into the modern power systems among the utility industry during the past few decades. Unlike other utilities such as water and natural gas, electricity cannot be stored in bulk as efficiently and economically. Thus, actions need to be taken to ensure real-time balancing between the supply and demand, such that the stability and reliability of the grid can be maintained. Load forecasts help business stakeholders understand the electricity consumption behavior and make informed decisions on the generation, load balancing, grid planning, operations, energy trading strategies, revenue projection, and rate design. Load forecasts have been served as input to other industries, including regulatory commissions, industrial and big commercial companies, banks, trading firms, and insurance companies (Hong & Fan, 2016).

The calendar has been an important driving factor for load demand. Among all the calendar variables contributed to the load forecasting models, the day type, time of day, and month of the year are known to have rooted impacts on electricity usage. People can have disparate schedules on a weekday compared to a weekend day, therefore leads to different electricity usage patterns. Between the two different weekdays or weekend days, the load pattern could differ as well. For instance, people might stay up late on Fridays compared to the other weekdays; the wake-up time on Sunday may be different for people who need to get up earlier for church ceremonies. For a holiday, the load pattern can be unique no matter if it's falling under a weekday or a weekend day. Meanwhile, human being's activities vary at different time of a day, thus result in different load patterns – for a person who has a regular day job, on a workday, he/she will wake up in the morning and

go out for work, get home by the evening, and go to bed during the night. The impacts of human activities on the load patterns can also be realized by the month of a year. Months are sometimes grouped into seasons to present the annual seasonality of the load consumption (Hong, 2010).

Since human beings react to weather and climate events, both can have influential impacts on electricity usage. Weather refers to the short-term (hourly, daily, weekly) atmospheric variations within a region, which can be factors such as temperature, humidity, wind speed and direction, rainfall, etc. Climate describes the weather observations in a particular region over longer periods of time, which can cover insightful information such as normal temperature during winter, average seasonal precipitation, extreme weather days to expect, etc.

Temperature is known to have a strong correlation with electricity usage patterns. Since part of the load consumption is used to maintain ambient temperature and humidity to fulfill human's comfort needs, dry bulb temperature and adjusted temperature variables (e.g., wet bulb temperature, dew point temperature, etc.), along with their variants (polynomials, temperature of preceding hours, etc.), have been the used extensively in the load forecasting models.

Besides temperature, other weather variables such as humidity, wind speed, solar irradiance, and precipitation can impact electricity usage. Since the combination of heat and high humidity causes discomfort, humidity is usually included in the form of heat index (HI), discomfort index (Senjyu et al., 2005) or relative humidity with interaction to the coincident hour and temperature (Xie et al., 2018). The speed of wind increases the rate of evaporation of perspiration from human body and gives a cooling effect. Therefore during

a windy summer day, the load consumption could be lower since fewer cooling appliances will be used (Fahad & Arbab, 2014). Solar irradiance heats the surfaces of premises or buildings, leading to disruption of the inside temperature and in turn, plays a part in influencing the load for cooling or heating needs. Heavy rain or snow (precipitation) can make people stay home more which impacts their electricity usage, and meanwhile, it can cause the temperature to drop and thus lead to a positive or negative impact on load consumption (Fahad & Arbab, 2014).

Weather stations normally report these weather readings on an hourly or sub-hourly basis. The winning entries in GEFCom2012 and GEFcom2014 have shown that the load pattern under a certain region can better be represented and predicted by combining weather readings reported by multiple weather stations rather than a single station (Hong et al., 2014) (Xie & Hong, 2016). This is likely because of the electricity demand usually aggregates consumption from different geographical regions where there is typically available data from more than one weather station (Moreno-Carbonell et al., 2019). For instance, when the goal is to forecast the load of North Carolina, the weather readings reported at each city could be candidate input variables to the load forecasting model.

Although tens or even hundreds of weather stations are available for some US utilities within their service territories, only a small fraction of weather stations have been considered in their load forecasting systems (Hong et al., 2015). It is normally impractical to customize the weather station selection (WSS) and manually link the best weather station(s) to each zone based on investigations of their geographical location, topography information or atmospheric circulation condition. Even though the station(s) could be picked based on domain knowledge and experience, the forecasting performance with

under the station(s) could not be guaranteed, since the forecasting accuracy can be affected by the forecasting model we choose, the data history we use to train the model, and the data quality at the station(s).

Aiming at resolving this practical issue, two recent representative literature reported the endeavor of finding the best subset of weather stations for a territory of interest based on data-driven approaches. (Hong et al., 2015) proposed a WSS framework with simplicity and transparency based on a greedy search algorithm. This framework is currently being used by many power companies, such as North Carolina Electric Membership Corporation, which has been used as one of the case studies in the original paper (Y. Wang et al., 2019); (Moreno-Carbonell et al., 2019) proposed a GA-based (genetic algorithm) method to select the best set of stations which led to better accuracy. However, due to the stochastic nature of GA along with its complexity of setting up an adaptive GA framework for a specific WSS problem, without mentioning the significantly larger computational resources it would require given the size of its search space, the feasibility from a GA-based method may be limited in certain circumstances. With that being said, in the recent literature, there has not been any significant effort to both explore heuristic approaches with simplicity and transparency to improve the WSS and, in the meantime, unveil how close our selection is approaching the "optimal" selection (i.e., the selection by exhaustive search).

To bridge the aforementioned gap, in this thesis, we extend the case studies based on (Hong et al., 2015) using the load forecasting data from GEFCom2012 published by (Hong et al., 2014) and from GEFCom2014 published by (Hong et al., 2016). We propose several comprehensive WSS frameworks for electric load forecasting and present

transparent and reproducible implementations. In practice, unlike temperature, which can normally be forecasted quite accurately within four or five days ahead, other weather variables such as humidity, wind speed, and solar irradiance, are not as predictable. Although the Tao's Vanilla benchmark model (from now on, Vanilla) we used in this thesis includes temperature as the only weather variable, our proposed frameworks are not limited to the load forecasting models being used and other weather variables can be included upon availability. Note that the proposed frameworks are not intended to apply to weather-insensitive loads, i.e., industrial loads.

We introduce theoretical optimum (from now on, TO) selection by using two years of data before the test set for parameter estimation and locate the best subset of stations based on the forecasting performance on the test set. This not only unveils what the best subset looks like given we have access to the future load, but it also provides us insights on why some WSS frameworks outperform the others.

This thesis makes the following significant contributions to the WSS in the load forecasting literature: (1) this is the first time that the exhaustive search method has been explored and its effectiveness has been evaluated; (2) this is the first time that the theoretical optimum (TO) selection is introduced to unveil important insight on the effectiveness of WSS techniques; (3) this is the first time that a group of heuristic methods are implemented and compared on their selection behavior with transparency; (4) it covers the first formal comparison and extensive discussion among the model selection methods and leads to several actionable rules of thumb; (5) publicly available data in our case study and transparent implementations allow future researchers to reproduce our results.

The rest of this thesis is organized as follows: Chapter 2 introduces the literature review of this area; Chapter 3 introduces the background of the study, including the Vanilla model, the forecast evaluation metrics, and the high-level introduction of the proposed frameworks along with a few model selection methods; Chapter 4 introduces the detailed implementation steps of the proposed selection frameworks; Chapter 5 presents exploratory analysis on the case study data, illustrates experiment results and discusses the issues and findings. The thesis concludes in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

2.1. Electric Load Forecasting

The electric load forecasting literature gains a growing trend in recent decades. Industry researchers have been categorizing the load forecasting (LF) types based on the range of the forecast. There has not been a commonly accepted rule among the industries to define LF categories based on various forecasting horizons. Tao and Fan have classified LF into 4 different categories: very short term load forecasting (VSTLF), short term load forecasting (STLF), medium-term load forecasting (MTLF), and long term load forecasting (LTLF) with cut-off horizons as 1 day, 2 weeks, and 3 years, respectively (Hong & Fan, 2016).

VSTLF and STLF aim at bridging the gap between forecasting and decision making in demand response programs, hour-ahead scheduling, day-ahead scheduling, unit commitment, and day-ahead energy trading, whereas VSTLF is often viewed as a sub-problem of STLF since both can take weather forecasts as the inputs for the forecasting period (Luo et al., 2018). As of VSTLF, (Senjyu et al., 2002) leveraged similar day data as input to the neural network model (NN) for one-hour-ahead electricity load forecasting using the load data from Okinawa Electric Power Company in Japan. The NN outputs a forecasting correction instead of the forecasted load and thus reduced the NN structure and learning time. (Islam et al., 2017) conducted one-hour-ahead electricity load forecasting in a deregulated power grid. The paper proposed using a modified backpropagation neural network model (BPNN) while the initial parameters of the model were tuned by chaos-search genetic algorithm and optimized using a simulated annealing algorithm. The case study was performed using small load demand data from a small power utility and large-

sized grid data from New South Wales in Australia. The model was tested at four seasons to validate its adaptivity within the highly fluctuating situations. Since the lagged load was often used in the VSTLF models, whereas the load value collected in the most recent hour is very likely to be inaccurate, (Luo et al., 2018) presented a model-based anomaly detection method with adaptive threshold to cleanse the corrupted load data.

There is a rich literature in the field of STLF and it is worth mentioning that (Hong, 2010) set the foundation of applying multiple linear regression (MLR) models for one-week-ahead hourly load forecasting. This dissertation constructed a base model based on observing the relationship among the load consumption, temperature, calendar variables, and a linear trend. The extensions of the base model have been discussed to accommodate different scenarios (VSTLF with preceding hour load, MTLF/LTLF with econometric factors) and a few customizations including recency effect, weekend effect, holiday effect, and exponentially weighted least squares have been explored to further enhance the predicting power of the base model. Due to the transparency, computational simplicity and forecasting accuracy, the base model has been widely known as Tao's Vanilla Benchmark model and studied in many research papers. The model has also been used as a benchmark model in the recent global energy forecasting competitions, including GEFCom2012, 2014 and 2017.

For power companies, MTLF and LTLF can be used to determine their long-term infrastructure commitment, system planning and forging energy policies with regulatory commissions. Rather than STLF that emphasizes fitting models to datasets and extrapolating from the past pattern, MTLF and LTLF tend to require more understanding towards how power system and electricity market work, as well as taking into account of

economy impacts (Khuntia et al., 2016). Some state of the art forecasting techniques for mid and long-term horizons in power systems were covered in (Khuntia et al., 2016). The paper stressed the importance of probabilistic load forecasting even though most utilities are still developing and making decisions using point forecasts. The MTLF and LTLF tend to generate multiple scenarios while each scenario would answer a what-if question in the future. (Ghedamsi et al., 2016) proposed a bottom-up model to forecast electricity load for Algeria residential buildings for the next 30 years. The authors leveraged GIS (Geographic Information System) to create and present the cartography of each climatic zone and then divided Algerian territory into 7 different zones based on the annual cost of energy consumption on cooling and heating per year responding to local climate, under residential category. The major energy consumers under each residential category were identified while HDD (heating degree days) and CDD (cooling degree days) were served as input to the degree-days method to produce the final forecast.

During recent years, many valuable forecasting techniques and models were being tested and implemented in the industries. These techniques can be roughly classified into three categories, namely the statistical models, the artificial intelligence (AI) based models, and the hybrid models which incorporate decomposition, clustering, optimization, and aggregation rules to further enhance the forecasting performance from the individual approaches (Ghalehkhondabi et al., 2017).

Within the family of statistical models, the univariate models, such as ARMA (auto-regressive and moving average) models (Weron, 2006) and exponential smoothing (Taylor, 2008) do not rely on explanatory variables, meaning it has lower data requirements than other widely used techniques such as MLR and artificial neural network (ANN) (Hong

& Fan, 2016). Due to their dependency on lagged load data, their forecasting performance can be good within a few steps ahead; while the forecast horizon gets longer, their forecasting accuracy could drop dramatically.

The regression model is established under the assumption that the model itself is a reasonable approximation to reality (Hyndman and Athanasopoulos, 2019). It can be powerful in providing accurate forecasts without the limitation of the forecast horizon, while the model itself as well as the parameter estimation process is more interpretable than some "black-box" models. One of the earliest implementation of a regression-based approach on STLF was proposed in (Papalexopoulos & Hesterberg, 1990), in which the authors employed a MLR model with temperature being modeled by heating and cooling degree functions, holidays being modeled by binary variables, and model parameters being estimated by weighted least squares. This approach was used to produce 24 hour ahead peak and hourly forecast on Pacific Gas and Electric Company's (PG&E) data and was concluded to make more accurate prediction than the existing approach at PG&E. More recently, (Hong, 2010) applied MLR models for one-week-ahead hourly load forecasting and the proposed base model has been used as a benchmark in many competitions and research papers. Aiming to further enhance the model forecasting performance, (P. Wang et al., 2016) explored lagged hourly temperature and moving average temperature variables in addition to the base model to form a family of "recency" models.

The AI-based models reach to another hype in recent years' load forecasting literature due to the rise of modern high-performance computing power on personal computers. Artificial Neural Network (ANN) models are the iconic AI-based models that learn tasks mimicking the behavior of the human brain. ANN models have the well-known

advantages of being able to approximate nonlinear functions and solve problems where the input-output relationship is neither well defined nor easily computable. (Hippert et al., 2001) gave a comprehensive review of ANN application to STLF. As one needs to define the appropriate model complexity and choose the right input variables, (Y. Chen et al., 2010) proposed a wavelet neural network model for 24-hour ahead electric load forecasting, while the similar day's load in the history were selected as input. The similar days were found by correlation analysis, which was to pick the same weekday index and similar weather to the forecast day and apply the day-of-a-year index within a neighborhood of that forecast day.

In the recent load forecasting literature, the hybrid models which incorporate decomposition, clustering, optimization, and aggregation rules were emerged to maximize the forecasting performance. (Zheng et al., 2017) proposed a long short-term memory (LSTM) neural networks to conduct 24-hour ahead load forecasting, where the empirical mode decomposition (EMD) was applied to decompose the singular values into intrinsic mode functions (IMF). Subsequently, each IMF was fed into an LSTM neural network and the outputs were combined to form the forecast after series reconstruction. As the model selection process is crucial before producing the final forecast, the evolutionary algorithms (EA) have been known as one of these heuristic-based approaches by mimicking the parameter selection process to a process of natural selection. (Li et al., 2014) proposed a hybrid quantized Elman neural network (HQENN) with the fewest quantized inputs to conduct 24-hour ahead load forecasting. The genetic algorithm (GA) was implemented to locate the optimal number of neurons in the quantum-map-layer and the hidden-layer in the HQENN to enhance the load forecasting performance.

Besides the load forecasting models mentioned above, other representative techniques among the literature include support vector machines (B. J. Chen et al., 2004); semi-parametric additive models (Fan & Hyndman, 2012); a gradient boosting model (Ben Taieb & Hyndman, 2014), and fuzzy regression (Hong & Wang, 2014).

In the load forecasting models, weather variables are frequently used to capture salient features of historical load curves. Temperature has been the most prevalent one, due to its strong correlation to the demand of weather-sensitive load and the fact that it can normally be forecasted quite accurately within four or five days ahead. To model the v-shaped load-temperature relationship (illustrated in FIGURE 9), (Fan et al., 2009) used piecewise linear functions with the cut-off temperature at 59 °F. (Ziel & Liu, 2016) used piecewise linear functions as well with the cut-off temperature at 50 °F and 60 °F. As the cut-off temperature of piecewise linear functions may not be the same at different service territories, (Hong, 2010) extended the use of 3rd ordered polynomials of the temperature in the model. While the load can be affected by the temperatures of the preceding hours, the lagged temperature and 24-hour average temperature were introduced into the model (P. Wang et al., 2016) to further enhance its forecasting accuracy.

Other weather variables such as humidity, wind speed, and precipitation, although not as predictable as temperature, also play a part in the load forecasting models per data availability. In load forecasting models, humidity can be included in the form of heat index (HI), temperature-humidity index (Senjyu et al., 2005), or relative humidity with interaction to coincident hour and temperature (Xie et al., 2018). Wind speed variables can be included in the form of the Wind Chill Index (WCI), wind speed adjusted temperature or wind speed with interaction to the coincident hour and temperature (Xie & Hong, 2017).

As precipitation could have a direct or indirect effect to load demand, (Senjyu et al., 2005) introduced binary values (rain: -1, no rain: 1) in the membership function for the precipitation variable to correct the neural network output before the next day forecasted load can be obtained.

Calendar has a rooted impact on electricity usage and introducing calendar variables helps explain seasonal behavior and some time-dependent information in the historical load curves. To model the combined effect from the day type and time of day, (Hong, 2010) applied an interaction effect in the MLR model between two class variables – weekday and hour. Since the 7 days of a week can be modeled by qualitative variable with multiple different classes (e.g., 2 classes: weekdays and weekends, 3 classes: weekdays and two separate weekend classes), (Xie et al., 2015) modified the days of year and combined Tuesday, Wednesday and Thursday together to reduce the degree of freedom of the model.

To model annual seasonality, (Hong, 2010) added a class variable – month, into the model to classify load patterns throughout the year. Some other literature also reported using season in the load forecasting model, where the season can be formed by combining pre-defined months (Charlton & Singleton, 2014) or days (K. Chen et al., 2019). Other than using the Gregorian calendar, (Xie & Hong, 2018) proposed using 24 solar terms from ancient China to classify the load patterns.

For a holiday, the load pattern can be unique. For the ten US federal holidays, six of them fall into fixed weekdays. In some cases, the holidays can be treated as a weekend day (Xie et al., 2015). Modeling load patterns on holidays can a challenging task due to the

lack of holiday data, the unique impact that a holiday may bring to the electricity usage, and its impact on the load pattern of the surrounding days (Hong, 2010).

Due to the high penetration of renewable energy resources (solar/wind farm, electric vehicles, etc.) and intensive market competitions, the probabilistic load forecasting (PLF) which captures the forecast uncertainty, has become a crucial input for decision making in the energy system planning and operation procedures. (Hong & Fan, 2016) provided a comprehensive review of PLF, including the notable techniques, evaluation methods, and common misunderstandings. Popular techniques to generate probabilistic load forecast include using quantile regression from point forecast (Liu et al., 2017)(Ben Taieb et al., 2016), residual simulation (Xie et al., 2017), Monte Carlo simulation (Laković et al., 2017), temperature scenario simulation (Gaillard et al., 2016) and hybrid methods combining the former approaches (Haben & Giasemidis, 2016). (Hong et al., 2016) summarized the recent research in this field and the techniques used by top entries on the PLF track of GEFCom2014. Meanwhile, the paper envisioned the development trend in the probabilistic forecasting field in the next 10 years.

As the era of smart grid has come since the 2000s and a massive number of smart meters being deployed in the field during the past decade, the meter readings give the power companies more insights towards their end-users. This huge amount of readings data has been leveraged to analyze customer usage patterns, which can lead to an extensive study on demand-side management (Andersen et al., 2017) and generates more study cases for an emerging subject called hierarchical load forecasting (HLF). HLF studies enable the industry to forecast electricity consumption at various levels at different forecasting

horizons, from the most granular level, a single household, to a major service territory across the country; from one-day ahead to several years ahead (Hong & Fan, 2016).

The literature on HLF is showing a growing trend in the recent decade. (Fan et al., 2009) established a multi-region load forecasting system to predict the aggregated electricity demand for a US Midwest utility company. The study introduced optimal region partitions of the sub-areas to minimize the forecast load at the aggregated level. (Lai & Hong, 2013) explored several methods of regional load grouping and averaging of weather stations under the hierarchical load forecasting settings based on the ISO New England data. (Zhang et al., 2015) proposed a 5-step hierarchical load forecasting framework using terabyte (TB) amount of AMI (advanced metering infrastructure) data. A hierarchical clustering technique based on normalized Euclidean distance is used to identify load patterns at bottom levels and prediction of load consumption under system level is the aggregation of individual load's forecasting results, as shown in FIGURE 1. The Global Energy Forecasting Competition 2012 (GEFCom2012) organized by the IEEE Working Group on Energy Forecasting (WGEF), with a hierarchical load forecasting track, was the first formal competition and attracted worldwide forecasters on solving a 2-level hierarchical load forecasting problem: to backcast and forecast hourly loads for a US utility at both zonal level (20 series) and at system level (sum of the 20 zonal level series) using temperature information from 11 weather stations. (Hong et al., 2014) summarized the methodologies used by 11 entries in the hierarchical load forecasting track and four of the winning teams have published their methods to the International Journal of Forecasting (Ben Taieb & Hyndman, 2014)(Charlton & Singleton, 2014)(Lloyd, 2014)(Nedellec et al., 2014).

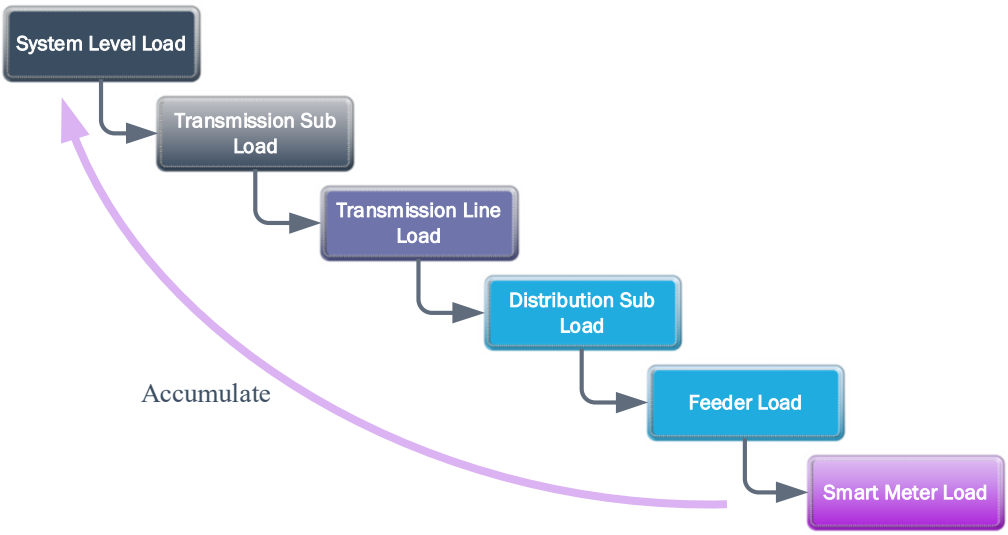

FIGURE 1: Load Forecasting Aggregation Topology

The Global Energy Forecasting Competition 2017 (GEFCom2017) brought together state-of-the-art techniques and methodologies under the hierarchical load forecasting theme by having a probabilistic load forecasting problem with multiple sizes of the hierarchy: to forecast the zonal and total loads of ISO New England in the qualifying match (3 levels), and forecast the load for hundreds of delivery point meters of a US utility (4 levels). (Hong et al., 2019) summarized the methodologies used by the top teams while found a modest usage of the hierarchy information among the winning teams. The authors attributed this cause to 3 reasons, namely the immaturity of hierarchical probabilistic forecasting in the literature, the intensity of the competition schedule, and the benefit of incorporating hierarchy information comparing to the extra complexity it brings.

Although forecasting loads at the aggregated level has been a relatively matured area, due to the increasing need for highly accurate load forecasting, (Y. Wang et al., 2019) summarized four rising topics that are gaining attention from academic researchers and industry practitioners. The first one is to integrate the forecast errors obtained from a pool

of major forecast evaluation methods into the decision-making process. The second one is to leverage smart meter information and data-driven approaches to detect load transfers. The third one is to leverage the emerging data sources and study the impacts on the traditional load profiles from the disruption of distributed energy resources, energy storage, and smart home devices. The fourth topic relates to what this thesis focuses on, which is to fully utilize zonal, regional load and local weather data through WSS to improve load forecast accuracy under the context of HLF.

2.2. Weather Station Selection

Since weather factors have been playing key impacts to the electricity load consumption at different hierarchies, weather readings such as temperature, humidity, wind speed, and cloud cover, are widely used as input variables in load forecasting models. As it is impractical to customize the WSS and manually link the best weather station(s) to each zone based on the investigation of their geographical location, topography information or atmospheric circulation condition, a few literature in the field were published to find the best subset of weather stations for a territory of interest.

Among the 4 winning entries in GEFCom2012, (Charlton & Singleton, 2014) drew a baseline model selecting one weather station per calendar group based on the goodness-of-fit of a multiple linear regression model using singular value decomposition (SVD) for computation of the coefficients, and the final forecast was combined from 5-best fitted weather stations. (Nedellec et al., 2014) selected one station for each zone using a step-wise procedure based on minimizing a V-fold cross-validation criterion. (Ben Taieb & Hyndman, 2014) selected one station for each zone based on the forecasting performance

on a test week. (Lloyd, 2014) uses temperature series from all weather stations. The benchmark method selected one station for each zone based on the best-fit model.

As the decision-making processes in the energy industry rely further on the probabilistic load forecasts to better capture the uncertainties in the modern grid, the Global Energy Forecasting Competition 2014 (GEFCom2014), under the theme of probabilistic load forecasting, was held including a track on electric load forecasting, with a WSS problem similar to the setup of the hierarchical load forecasting track in GEFCom2012. The competition problem under the electric load forecasting track was to forecast the quantiles of hourly loads for a US utility using temperature information from 25 weather stations. (Hong et al., 2016) summarized the methodologies used by 7 teams in the electric load forecasting track. Six of the winning teams published their methods to the International Journal of Forecasting (Gaillard et al., 2016)(Ziel & Liu, 2016)(Xie & Hong, 2016)(Dordonnat et al., 2016)(Haben & Giasemidis, 2016)(Mangalova & Shesterneva, 2016). Among the 7 winning teams in GEFCom2014, team Tololo ranked in the first place and fitted each weather station to a generalized additive model (Gaillard et al., 2016). Four stations were selected based on their performance under the generalized cross-validation (GCV) criterion. Team Adada who ranked in the second, selected seven stations based on GCV initially and refined the selection using an exponentially weighted average (EWA) algorithm (Dordonnat et al., 2016). The method ended up selecting three stations. Team Jingrui Xie ranked in the third and selected 11 weather stations using the selection framework from (Hong et al., 2015) (Xie & Hong, 2016). Team Ziel Florian who took the second place in GEFCom2014-L, selected two weather stations based on the goodness of

fit to a cubic regression of the load against the temperature (Ziel & Liu, 2016). Other reported winning teams used the average of temperature series from 25 weather stations.

In the GEFCom2017, 28 anonymous weather stations data including temperature and relative humidity were provided. (Hong et al., 2019) reported that only two out of eight top entries reported their WSS methods: team GeertScholma used goodness-of-fit of a polynomial temperature model to select one weather station for each meter; team QUINKAN first applied hierarchical clustering to categorize 28 stations into 11 clusters, then assigned the best cluster to each meter series based on the performance of a GBM (generalized boosting method) (Kanda & Veguillas, 2019).

Some of the aforementioned methods followed a practice of first subjectively selecting a fixed number of weather stations for each zone, then identifying the best weather station(s) under the constraint of the fixed number(s). (Hong et al., 2015) suggested this was a counter-intuitive process due to the aspects of geographic diversity, demographic diversity, and the end-use diversity for each zone. Other methods subjectively selected a number of weather stations based on the goodness of fit to a pre-defined forecasting model. The process of determining how many stations to be included relied heavily on the user's empirical experience to the forecasting model itself and the results were not convincing enough since they have not been compared with other alternatives under the same method.

Given the aforementioned situation related to WSS and to solve the issue of how many stations should be used and which station(s) to be used, (Hong et al., 2015) proposed a WSS framework by removing the constraint on the fixed number of weather stations. The framework was based on a greedy approach – a heuristic approach to reach a good result

within an affordable amount of computation time. (Moreno-Carbonell et al., 2019) pointed out that this approach, despite its simplicity and reproducibility, has a major drawback of not allowing to remove useless variables with its incremental nature and thus, can lead to an over-parameterized combination of stations. They instead proposed a WSS approach based on GA (genetic algorithm), which allowed selecting the best set of weather stations for each value of K (K=number of stations being selected). The optimal K was chosen based on the cross-validation error rate of each trial and the one-standard-error rule was used to locate the most parsimonious combination of the stations. However, due to the stochastic nature of GA such as the randomness of initial search point, along with its complexity to tune the parameters and form up an appropriate GA framework which is adaptive to a specific WSS problem, without mentioning the considerably larger computing resources it would require given the size of the search space, the feasibility and business value from the GA-based method may be limited in certain circumstances.

In terms of ways to combine temperature series from the chosen weather stations, the simple average has been the most common way. A study in (Lai & Hong, 2013) has shown that the simple average can outperform the weighted average based on the ISO New England data set, in which the weights were driven by load size and economic output under each region. (Sobhani et al., 2019) tested out seven combination methods with simple averaging as the benchmark using data from GEFCom2012. Results have shown that averaging the forecasts from the seven methods could outperform the majority of the individual methods. (Moreno-Carbonell et al., 2019) leveraged Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm to update weights of chosen weather stations based

on training results of the Vanilla model. Results have shown that the proposed combination method outperformed the simple average method.

CHAPTER 3: THEORETICAL BACKGROUND

3.1. Multiple Linear Regression

Multiple linear regression is a classical statistical technique, attempting to model the relationship between the response variable and two or more independent variables. Each independent variable, sometimes called predictor or regressor, is associated with a coefficient, which measures its marginal effect among the independent variables to the response variable. Comparing to other black box approaches like neural networks, the model is easy to interpret. The model can have the following form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} + \varepsilon_i$$

(1) – Linear Regression Model

where y is the response variable, namely the variable we want to forecast. $X_1, \dots, X_k$ are the k predictors and $\beta_1, \dots, \beta_k$ are the associated coefficients. β_0 is called the intercept and measures the mean of response variable when all the independent variables are 0. The response function can be written as:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

(2) – Response Function of the Regression Model

The multiple linear regression resides on the following key assumptions:

- The model is a reasonable approximation to reality – meaning the relationship between the response and independent variables satisfies the linear equation (Hyndman, R.J. and Athanasopoulos, 2018).
- The error terms, $\varepsilon_1, \dots \varepsilon_i$, are assumed to be caused by independent measurements and describe the deviance between the data samples and the true values along the

regression line. The error term ε_i is assumed to be independent, normally distributed variable $N(0, \sigma^2)$ with constant variance.

- In the ideal situation, the error terms should be unrelated to the predictors.
- The independent variables shouldn't be highly correlated with each other.

The coefficients can be estimated by fulfilling the least sum of error square condition, which can be defined as follows:

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) = \underset{(\beta_0, \beta_1, \dots, \beta_k)}{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_{1,i} - \cdots - \beta_k X_{k,i})^2$$

(3) – Parameter Estimation Problem Formulation

where N is the number of observations. This equation can have closed-form solutions through the matrix inverse:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(4) – Parameter Estimation in Matrix Form

For large datasets which involves a huge amount of computation, or the inverse of $X^T X$ does not exist, the coefficients can be estimated by gradient descent approaches.

A predictor can either be quantitative or qualitative. To predict a weather-responsive load, a quantitative predictor can be the outside temperature in Fahrenheit. An example qualitative predictor can be the month of the year with 12 classes (January, February, …, December), indicating various load levels at different months. Under which scenario, this qualitative predictor will be represented by 11 indicator variables, sometimes called the dummy variables (with values 0 and 1) as follows:

$$\begin{cases} X_1 = 1, if\ the\ month\ is\ January \\ X_1 = 0, otherwise \\ X_2 = 1, if\ the\ month\ is\ February \\ X_2 = 0, otherwise \\ \qquad \qquad \dots \\ X_{11} = 1, if\ the\ month\ is\ November \\ X_{11} = 0, otherwise \end{cases}$$

(5) – Dummy Coding of Month Variables

And the response function of the regression model with the month of the year as the predictor variable is:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{11} X_{11}$$

(6) – Response Function of Regression Model with Month Variables

When the effects of one predictor variable depend on the level(s) of some other predictor variable(s), interaction effects can be included in the regression model (Hong et al., 2011). To predict a weather-responsive load, the temperature (quantitative variable) can interact with the month (qualitative variable) since the coincident temperature is not independent of the month of the year.

3.2. Base Model

While it is worth noting that the use of one load forecasting model or the other can lead to different "fitness" of each station and we cannot conclude which set of weather station is the best one in general terms (Moreno-Carbonell et al., 2019), our goal is to find the best WSS based on the Vanilla model. This model was first proposed by (Hong, 2010) and then used in GEFCom2012 as the benchmark for the hierarchical load forecasting track (Hong et al., 2014). This model produces relatively good forecasts with computational simplicity. The model is specified as follows:

$$Load = \beta_0 + \beta_1 \cdot Trend + \beta_2 \cdot Hour + \beta_3 \cdot Weekday + \beta_4 \cdot Month + \beta_5 \cdot T + \beta_6 \cdot T^2 + \beta_7 \cdot T^3 + \beta_8 \cdot Hour \cdot Weekday + \beta_9 \cdot T \cdot Hour + \beta_{10} \cdot T^2 \cdot Hour + \beta_{11} \cdot T^3 \cdot Hour + \beta_{12} \cdot T \cdot Month + \beta_{13} \cdot T^2 \cdot Month + \beta_{14} \cdot T^3 \cdot Month$$

(7) – Vanilla Model

Load is the coincident response variable that will be forecasted by the predictor variables on the right side. The standalone predictors includes a Trend variable (an increasing natural number) to represent a linear trend of the load along with the data history; a 24 classes qualitative variable, Hour, to represent load level shift at 24 hours of a day; a 7 classes qualitative variable, Weekday, to represent load level shift at seven days of a week; a 12 classes qualitative variable, Month to represent load level shift at 12 months of a year; quantitative variables, T, T^2 and T^3, to represent the relationship between the temperature (combined from the weather station(s)) polynomials and the load. The temperature polynomials are interacting with the hour of the day and the month of the year. The hour of the day is interacting with the day of the week. Further details about this benchmark model can be found in (Hong et al., 2011).

For the weather station combination, (Sobhani et al., 2019) has shown many ways can be used to combine temperature series. To make a direct comparison to the results in (Hong et al., 2015), we use simple averages to combine temperature series and create virtual weather stations in this thesis.

3.3. The Model Evaluation Metrics

There are various model evaluation metrics in the load forecasting field to inform the forecasting accuracy of a model. Since there are not many loads close to zero in the data history and to establish a direct comparison to the results in (Hong et al., 2015), we

use MAPE (mean absolute percentage error) as the error measure to evaluate the forecast performance:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}$$

(8) – The Error Measure, MAPE

where N is the number of observations in the test period, y_i and $\hat{y}_i$ are the actual values and predicted values, respectively.

3.4. WSS Frameworks

In this section, we will give out some high-level introduction on four major WSS frameworks we are going to present in the later chapters, namely the exhaustive search framework and three heuristic frameworks – forward selection framework, backward selection framework, and greedy selection framework. Three types of statistical tests are considered under each framework: the in-sample fit, post-sample fit, and out-of-sample cross validation test. These tests vary on the way of error estimation and thus lead to different WSSs. We will cover more details about these three statistical tests in Section 3.5. To avoid verbose illustration, for the four major frameworks, we denote the exhaustive search framework as ES, the forward selection framework as FS, the backward selection framework as BS and the greedy selection framework as GS. We denote the in-sample fit methods as IS, the post-sample methods as PS and the out-of-sample cross validation methods as CV.

In this thesis, we denote N to be the number of weather station candidates to choose from, where $N = 11$ for the GEFCom2012 case study and $N = 25$ for the GEFCom2014 case study.

The exhaustive search method, sometimes called the best subset selection method, is a type of searching method to locate the global optimum by going through a finite number of possibilities sequentially. In the context of weather station selection for load forecasting, it is to go through and test out every possible WSS and use the best-performed selection from the history to forecast the future load. We label each weather station as i , where $i = 1, \dots, N$. The set of available weather stations, S, is denoted as:

$$S = \{1, \dots, N\}$$

(9) – The Set of Weather Stations

It's power set, denoted as $P(S)$ or 2^S, represents the set containing all possible subsets of S as its elements:

$$P(S) = \{\emptyset, \{1\}, \dots, \{N\}, \{1,2\}, \dots, \{N-1, N\}, \dots\}$$

(10) – The Power Set of Weather Station Subsets

The cardinality of the power set $P(S)$ is equal to 2^N, which includes the empty set $\emptyset$. Since at least one weather station will be selected, the number of weather station subset T, is equal to $2^N - 1$. In which case, for ES-IS and ES-PS, the WSS process requires $2^N - 1$ iterations to evaluate the Vanilla model. For ES-CV, it requires $3(2^N - 1)$. The number of possible WSSs for the two case studies are listed in *TABLE 1* and *TABLE 2*. The implementation details of the ES framework are elaborated in Section 4.1.

TABLE 1: No. of Possible WSSs under Each Group Size (GEFCom2012)

No. of stations selected	No. of possibilities
1	11
2	55
3	165
4	330
5	462
6	462
7	330
8	165
9	55
10	11
11	1
Total:	2047

TABLE 2: No. of Possible WSSs under Each Group Size (GEFCom2014)

No. of stations selected	No. of possibilities
1	25
2	300
3	2300
4	12650
5	53130
6	177100
7	480700
8	1081575
9	2042975
10	3268760
11	4457400
12	5200300
13	5200300
14	4457400
15	3268760
16	2042975
17	1081575
18	480700
19	177100
20	53130
21	12650
22	2300
23	300
24	25
25	1
Total:	33,554,431

Forward and backward selections both belong to the family of stepwise regression methods and have been widely used as the heuristic methods in the other fields, aiming at achieving good results with a less computational cost. FS begins with an empty set and adds in variables one at a time to the incumbent group (a group containing all selected variable(s)). In the context of WSS for load forecasting, instead of going through every possible WSS, forward steps begin with an empty group with no weather stations being selected and add one station at a time that gives the largest error measure improvement. This process will stop at a group of weather station(s) when no further improvement can be achieved by adding one more weather station to the incumbent group, or there's no more weather stations we can add.

BS begins with a full house of variables and the selection process is to eliminate one variable at a time from the incumbent group. In the context of WSS for load forecasting, the backward steps begin with all the weather stations being selected, eliminate weather station one at a time while recording down the selection which gives the largest error measure improvement. The selection process stops at a group of weather station(s) when no further improvement can be achieved by eliminating one more weather station from the incumbent group.

The general implementation processes of FS and BS are visualized in *FIGURE 2* and *FIGURE 3*. The implementation details of FS and BS for WSS are elaborated in Section 4.2 and Section 4.3. The FS and BS implementations with IS or PS require at most $N + (N-1) + \cdots + 1 = \frac{N(N+1)}{2}$ iterations to evaluate the Vanilla model and locate the WSS. The FS and BS implementations with CV require at most $3(N + (N-1) + \cdots + 1) = \frac{3N(N+1)}{2}$ iterations given the use of 3-fold CV. Under the GEFCom2012 case study,

FS and BS along with IS or PS require at most 55 iterations when $N = 11$. When $N = 25$ for the GEFCom2014 case study, FS and BS along with IS and PS require at most 300 iterations.

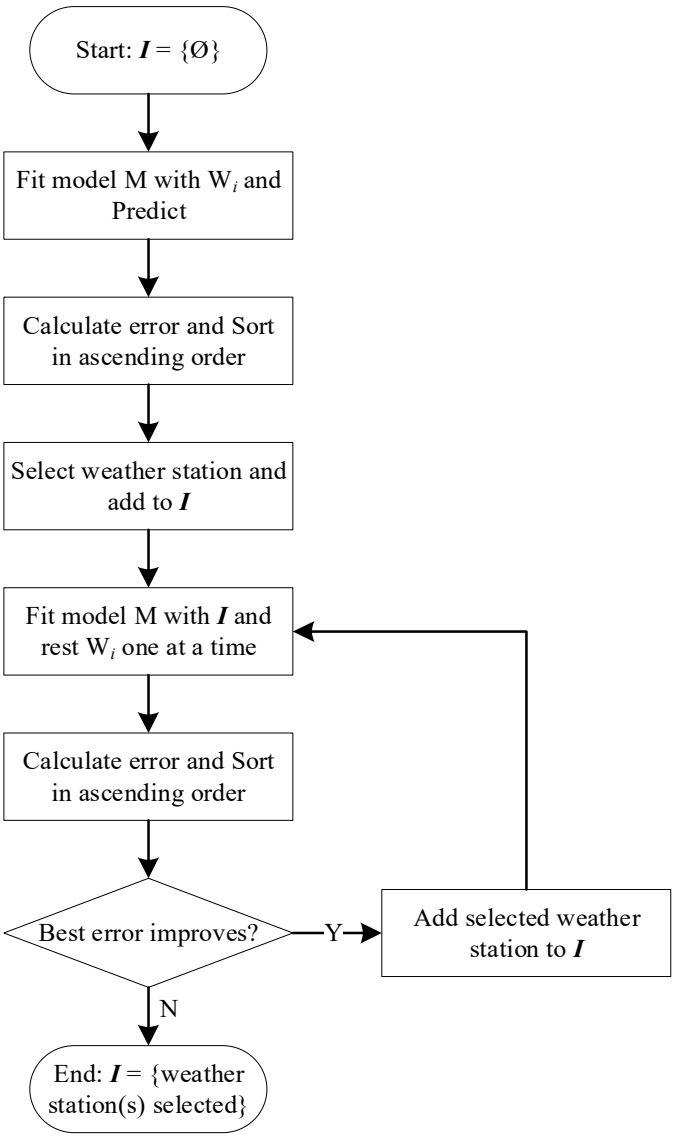

FIGURE 2: FS Implementation Process

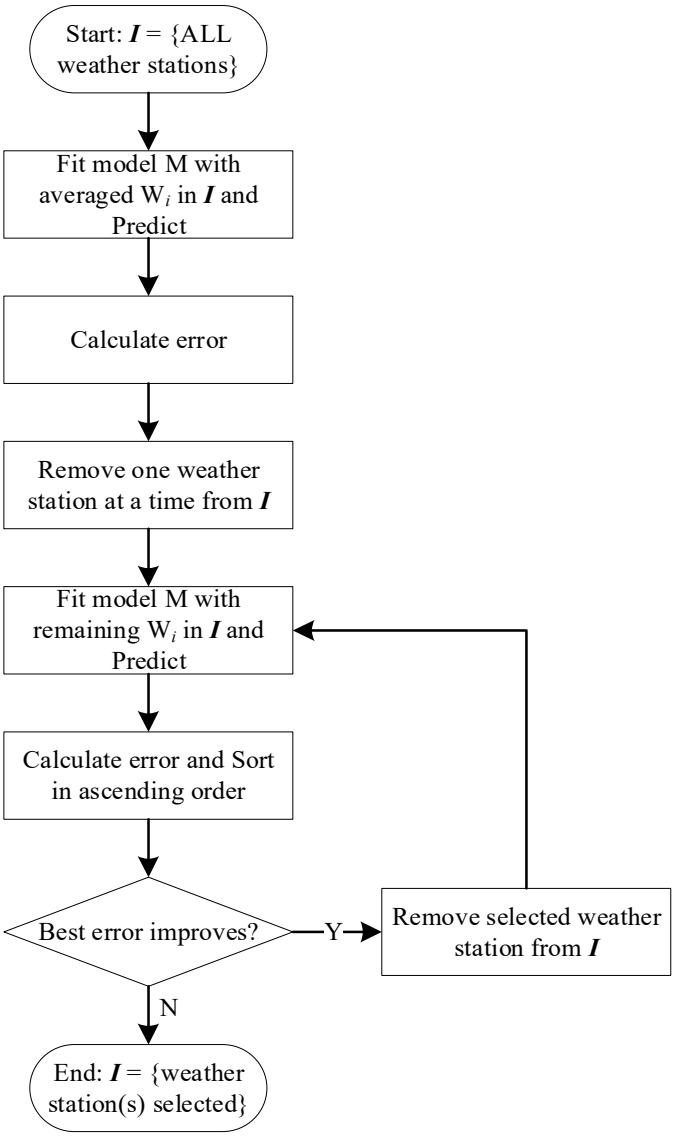

FIGURE 3: BS Implementation Process

A greedy selection method with the post-sample fit (GS-PS) has been illustrated in (Hong et al., 2015). In this thesis, we implement the greedy selection framework for WSS based on the in-sample fit and cross-validation errors. The general implementation process for the greedy selection methods is visualized in *FIGURE 4*. The implementation details of GS-PS and GS-CV for WSS are elaborated in Section 4.4. The benchmark and GS-PS both require $N + N = 2N$ iterations to evaluate the Vanilla model and locate the WSS. GS-CV requires $N + 3N = 4N$ iterations given the 3-fold CV implementation. Under the

GEFCom2012 case study, the benchmark and GS-PS require exact 22 iterations when $N = 11$. While $N = 25$ for the GEFCom2014 case study, the benchmark and GS-PS require exact 50 iterations.

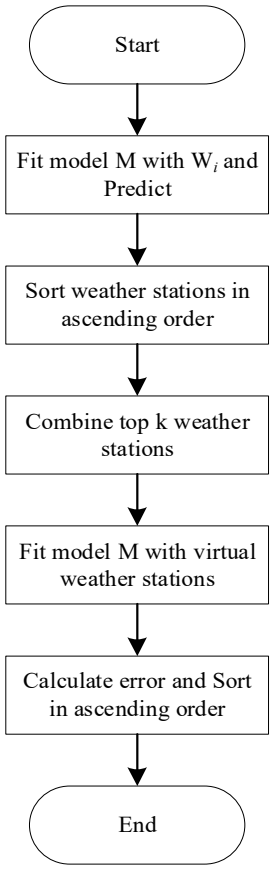

FIGURE 4: GS Implementation Process

3.5. Statistical Tests for Model Selection

Various statistical tests can be used to evaluate the forecasting performance of a model. Each framework we've introduced in Section 3.4 comes along with three types of statistical tests, namely the in-sample (IS) fit, post-sample (PS) fit, and out-of-sample cross validation (CV) test.

From the data history, when we use the same group of samples to estimate model parameters and evaluate the forecasting performance of the model, we call it the IS fit

method. When we evaluate the forecasting performance using a group of samples that were collected after the date on which the model parameters were estimated, we call it the PS fit method. *FIGURE 5* gives an example of the in-sample fit and post-sample fit method implemented within the GEFCom2012 case study.

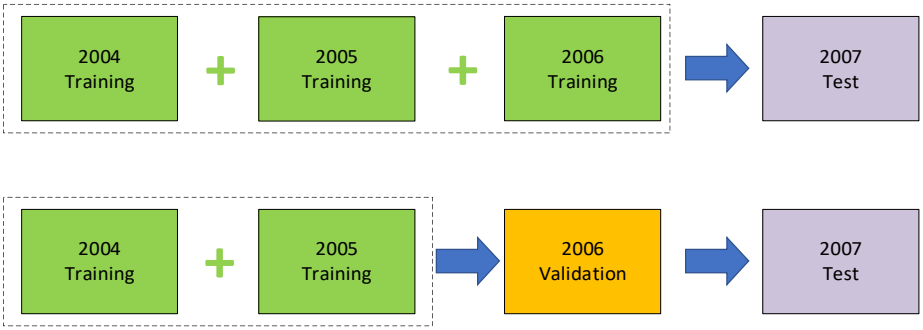

FIGURE 5: Example of In-Sample Fit (up) and Post-Sample Fit (down) Method in GEFCom2012 Case Study

CV, also called the V-fold cross validation (VFCV), is another type of model evaluation technique, where we divide the data into V subsets and each time we hold out one subset to evaluate forecasting performance, while the remaining subsets are used for model parameter estimation. This process will repeat V times then the average error across V trials is computed. *FIGURE 6* gives an example of CV implemented within the GEFCom2012 case study.

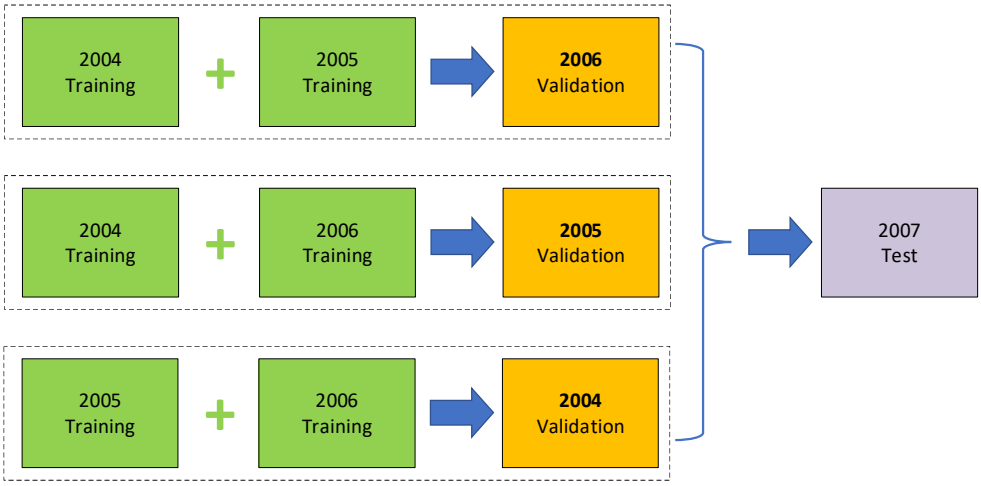

FIGURE 6: Example of CV in GEFCom2012 Case Study

In this thesis, we denote the number of years in the training and validation period by L_{train} and L_{val} respectively. We let $L_{train} = 2$ and $L_{val} = 1$ for post-sample fit methods and we conduct WSS based on the forecasting error on the validation period. For in-sample fit methods, we let $L_{train} = 3$ and $L_{val} = 0$ since we do not need validation data, while we conduct WSS based on the in-sample fit error. The CV methods will follow the 3-fold cross validation process, namely for each trial, one-year worth of data will be held out from the 3 years data history for validation and the WSS will be conducted based on the average error across the three validation periods.

CHAPTER 4: PROPOSED METHODOLOGY

4.1. Exhaustive Search Framework (ES)

In this section, we elaborate on the implementation details of ES. The ES comes along with three statistical tests and we denote them as ES-IS, ES-PS, and ES-CV. The exhaustive search framework with the in-sample fit (ES-IS) is implemented in the following steps:

(1) Denote the temperature series from each weather station as $W_i, i = 1, \dots, N$. Combine the temperature series of the weather stations to create a new temperature series $CW_k, k = 1, \dots, T$, and $T = 2^N - 1$. Following (Hong et al., 2015), a simple combination with averaging temperature of the selected stations has been used.

(2) Fit a predefined model M with the load data and the temperature data CW_k in the training period. We let $L_{train} = 3$ and $L_{val} = 0$ to conduct the in-sample fit. Calculate the in-sample fit error, $MAPE_k$, under CW_k.

(3) Sort the resulting error measures in ascending order and record the combined temperature series that leads to the smallest error measure.

(4) Use the selected combined temperature series over the last two years in the training period to fit the model M and forecast the load of the test period. In the GEFCom2012 case study, the training period for this step is 2005-2006 and the test period is 2007; in the GEFCom2014 case study, the training periods for this step are 2006-2007, 2007-2008 and 2008-2009 for the test years 2008, 2009 and 2010, respectively.

Like the implementation of ES-IS, we conduct ES-PS by letting $L_{train} = 2$ and $L_{val} = 1$. For the GEFCom2012 case study, we use the first two years of data (2004-2005)

to estimate the parameters of the model and predict the load in the third year (2006). The WSS resulting in the smallest error measure in the third year (2006) will be chosen to forecast the load of the test period (2007). For the GEFCom2014 case study, we use the first two years of data (2005-2006) to estimate the parameters of the model and predict the load in the third year (2007). The WSS leading to the smallest error measure in the third year (2007) will be chosen to forecast the load of the test period (2008). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

Like the implementation of ES-PS, for the GEFCom2012 case study, we implement the ES-CV by conducting a 3-fold cross validation technique in the training period (2004-2006). We divide the data into three segments nearly equally based on the calendar year. One of the three segments is used as validation data and the rest two segments are used for training. The performance of a WSS is evaluated based on its average error measure across all three segments. The WSS leading to the smallest error measure will be used to forecast the load of the test period (2007) with parameter estimation based on the year of 2005-2006. For the GEFCom2014 case study, we implement the ES-CV by conducting a 3-fold cross validation technique in the training period (2005-2007). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

4.2. Forward Selection Framework (FS)

In this section, we elaborate on the implementation details of FS. The FS comes along with three statistical tests and we denote them as FS-IS, FS-PS, and FS-CV. The

implementation steps of the forward selection framework with the in-sample fit (FS-IS) are listed as below:

(1) Denote the set of selected weather stations as I and at the starting point, $I = \{\emptyset\}$.

(2) For the temperature series from each weather station, $W_i, i = 1, \dots, N$, fit a predefined model M with the load data and the temperature data W_i in the training period. We let $L_{train} = 3$ and $L_{val} = 0$ for the in-sample fit method.

(3) Calculate the in-sample fit error, $MAPE_{(1),i}$, under W_i. The subscript 1 indicates only one weather station is considered to the combined temperature series at this time.

(4) Sort the resulting error measures in ascending order. Record the temperature series resulting in the smallest error measure and add it to I.

(5) Combine each temperature series of the remaining weather stations with the station(s) in I to create a new temperature series.

(6) Fit the model M using the selected combined temperature series over the training period and calculate the in-sample fit error.

(7) Sort the resulting error measures in ascending order, record the combined temperature series resulting in the smallest error measure and compare the error measure with the selection in I.

(8) If the error measure obtained in (7) is improved, repeat (5) - (7) until there is no weather station remains to be added or no further improvement can be achieved by adding one more weather station to the incumbent selection I. In other words, if the record obtained in (7) has a larger error measure than the selection in I, the station(s) in I will be selected.

(9) Use the selected combined temperature series over the last two years in the training period to fit the model M and forecast the load of the test period. Under the GEFCom2012 case study, the training period is 2005-2006 for this step and the test period is 2007; Under the GEFCom2014 case study, the training periods are 2006-2007, 2007-2008 and 2008-2009 for the test periods 2008, 2009 and 2010, respectively.

Like the steps of implementing FS-IS, we implement the FS-PS by letting $L_{train} = 2$ and $L_{val} = 1$. For the GEFCom2012 case study, we use the first two years of data (2004-2005) to estimate the parameters of the model and predict the load in the third year (2006). The error measure will be evaluated based on the third year (2006) for each combined temperature series in order to determine whether one more weather station will be added and which to be added to the incumbent group. The WSS leading to the smallest error measure in the third year (2006) will be used to forecast the load of the test period (2007). For the GEFCom2014 case study, we use the first two years of data (2005-2006) to estimate the parameters of the model and predict the load in the third year (2007). The error measure will be evaluated based on the third year (2007) for each combined temperature series in order to determine whether one more weather station will be added and which to be added to the incumbent group. The WSS leading to the smallest error measure in the third year (2007) will be used to forecast the load of the test period (2008). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

Like the implementation of FS-PS, for the GEFCom2012 case study, we implement the FS-CV by conducting a 3-fold cross validation technique in the training period (2004-

2006). We divide the data into three segments nearly equally based on the calendar year. One of the three segments is used as validation data and the rest two segments are used for training. The performance of a WSS is evaluated based on its average error measure across all three segments. If adding a station can result in a lower cross-validation error, the selection process will continue; otherwise, the selection process will stop and the incumbent selection will be used to forecast the load of the test period (2007) with parameter estimation on the year of 2005-2006. For the GEFCom2014 case study, we implement the FS-CV by conducting a 3-fold cross validation technique in the training period (2005-2007). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

4.3. Backward Selection Framework (BS)

In this section, we elaborate on the implementation details of BS. The BS will come along with three statistical tests and we denote them as BS-IS, BS-PS, and BS-CV. The implementation steps of the backward selection framework with the in-sample fit (BS-IS) are listed as below:

(1) Denote the set of selected weather stations as I, and at the starting point, $I = \{1, \dots, N\}$.

(2) Combine the temperature series from the stations in I, denoted as $CW_{(N)}$. The subscript N denotes there are N weather stations included in the combined temperature series.

(3) Fit a predefined model M with the load data and the temperature series $CW_{(N)}$ in the training period. We let $L_{train} = 3$ and $L_{val} = 0$ for the in-sample fit methods.

(4) Calculate the in-sample fit error, $MAPE_{(N)}$, under $CW_{(N)}$.

(5) Eliminate one station from I and combine temperature series of the remaining $N-1$ weather stations to create a new temperature series, $CW_{(N-1),k}, k = 1, \dots, N-1$. The subscript $N-1$ denotes there are $N-1$ weather stations included in the combined temperature series.

(6) Fit the model M with the combined temperature series $CW_{(N-1),k}$ over the training period.

(7) Calculate the in-sample fit error measure, $MAPE_{(N-1),k}$ under $CW_{(N-1),k}$ and sort the resulting error measures in ascending order. Record the combined temperature series resulting in the smallest error measure and compare the error measure with the $MAPE_{(N)}$ in (4).

(8) If the error measure is improved, eliminate the weather station selected in (7) from I and repeat (5)-(7) until there is no weather station remains can be eliminated, or there is no further improvement by eliminating one more weather station from the incumbent selection I. In other words, if the record obtained in (7) has a larger error measure than the selection in I, the station(s) in I will be selected.

(9) Use the selected combined temperature series over the last two years in the training period to fit the model M and forecast the load of the test period. Under the GEFCom2012 case study, the training period is 2005-2006 for this step and the test period is 2007; in the GEFCom2014 case study, the training periods are 2006-2007, 2007-2008 and 2008-2009 for the test periods 2008, 2009 and 2010, respectively.

Like the BS-IS implementation, we conduct BS-PS by letting $L_{train} = 2$ and $L_{val} = 1$. For the GEFCom2012 case study, we use the first two years of data (2004-2005) to estimate the parameters of the model and predict the load in the third year (2006). The

error measure will be evaluated based on the third year (2006) for each combined temperature series in order to determine whether one more weather station will be eliminated and which to be removed from the incumbent group. The WSS leading to the smallest error measure in the third year (2006) will be used to forecast the load of the test period (2007). For the GEFCom2014 case study, we use the first two years of data (2005-2006) to estimate the parameters of the model and predict the load in the third year (2007). The error measure will be evaluated based on the third year (2007) for each combined temperature series in order to determine whether one more weather station will be eliminated and which to be removed from the incumbent group. The WSS leading to the smallest error measure in the third year (2007) will be used to forecast the load of the test period (2008). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

Like the implementation of BS-PS, for the GEFCom2012 case study, we implement the BS-CV by conducting a 3-fold cross validation technique in the training period (2004-2006). We divide the data into three segments nearly equally based on the calendar year. One of the three segments is used as validation data and the rest two segments are used for training. The performance of a WSS is evaluated based on its average error measure across all three segments. If removing a station can result in a lower cross-validation error, the selection process will continue; otherwise, the selection process will stop and the incumbent selection will be used to forecast the load of the test period (2007) with parameter estimation on the year of 2005-2006. For the GEFCom2014 case study, we implement the BS-CV by conducting a 3-fold cross validation technique in the training

period (2005-2007). We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

4.4. Greedy Selection Framework (GS)

In this section, we elaborate on the implementation details of GS. The GS will come along with three statistical tests and we denote them as GS-IS, GS-PS, and GS-CV. The GS-PS is the benchmark method and the implementation details can be found in (Hong et al., 2015). The implementation steps of the greedy selection framework with the in-sample fit (GS-IS) are listed as below:

(1) Denote the temperature series from each weather station as $W_i, i = 1, \dots, N$. Fit a predefined model M with the load data and the temperature data W_k in the training period. We let $L_{train} = 3$ and $L_{val} = 0$ for the in-sample fit methods.

(2) Calculate the in-sample fit error, $MAPE_i$ under W_i and sort the resulting error measures in ascending order.

(3) Combine the temperature series of the top k weather stations to create a new temperature series $CW_k, k = 1, \dots, N$. Here, the average temperature of the top k weather stations is calculated.

(4) Use the selected combined temperature series over the last two years in the training period to fit the model M and forecast the load of the test period. Under our GEFCom2012 case study, the training period is 2005-2006 in this step and the test period is 2007; under our GEFCom2014 case study, the training periods are 2006-2007, 2007-2008 and 2008-2009 for the test periods of 2008, 2009 and 2010, respectively.

Like the implementation of the benchmark method (GS-PS), we implement the GS-CV by conducting a 3-fold cross validation technique in the training period (2004-2006) for the GEFCom2012 case study. Each weather station will be sorted based on the in-sample fit error in all three training years (2004-2006). We then divide the data into three segments nearly equally based on the calendar year. One of the three segments is used as validation data and the rest two segments are used for training. The top k weather stations will be used to create virtue stations and the performance of each is evaluated based on the average error across all three validation segments. The virtue station resulting in the lowest error will be used to forecast the load of the test period (2007) with parameter estimation based on the year of 2005-2006. For the GEFCom2014 case study, we implement the GS-CV by conducting a 3-fold cross validation technique in the training period (2005-2007) before forecasting year 2008. We repeat the same implementation and apply the rolling process until we complete the forecasts for all three years (2008-2010).

CHAPTER 5: CASE STUDY & DISCUSSION

5.1. GEFCom2012 Case Study

To make a fair comparison with the results in (Hong et al., 2015), we extend the case study with the data from the hierarchical load forecasting track of GEFCom2012. Using this public dataset allows our results to be reproduced and conveniently compared for future studies by other researchers.

The entire data set consists of hourly load data of 20 zones from the 1st hour of 1/1/2004 to the 6th hour of 6/30/2008, with hourly temperature history from 11 anonymous weather stations ($W_1 - W_{11}$).

To model the daily seasonality of hourly load, the coincident hour has been used to categorize the load patterns throughout the day. *FIGURE 7* presents the scatterplots of load – temperature relationship at each hour from the aggregated zone and the average temperature from the 11 stations.

To model the annual seasonality, the categorical variables – month, have been included to categorize load patterns throughout the year. *FIGURE 8* gives the scatterplots of load – temperature relationship at each month of a year from the aggregated zone and the average temperature from the 11 stations in this case study. The correlation between the load and temperature during the winter months (Month=12, 1, 2, 3) and summer months (Month=6, 7, 8, 9) is stronger than the remaining months, largely due to during these 6 months, load tends to increase when the weather (temperature here) gets extreme. In the remaining 4 months (Month=4, 5, 10, 11), the relationship between the two is weak.

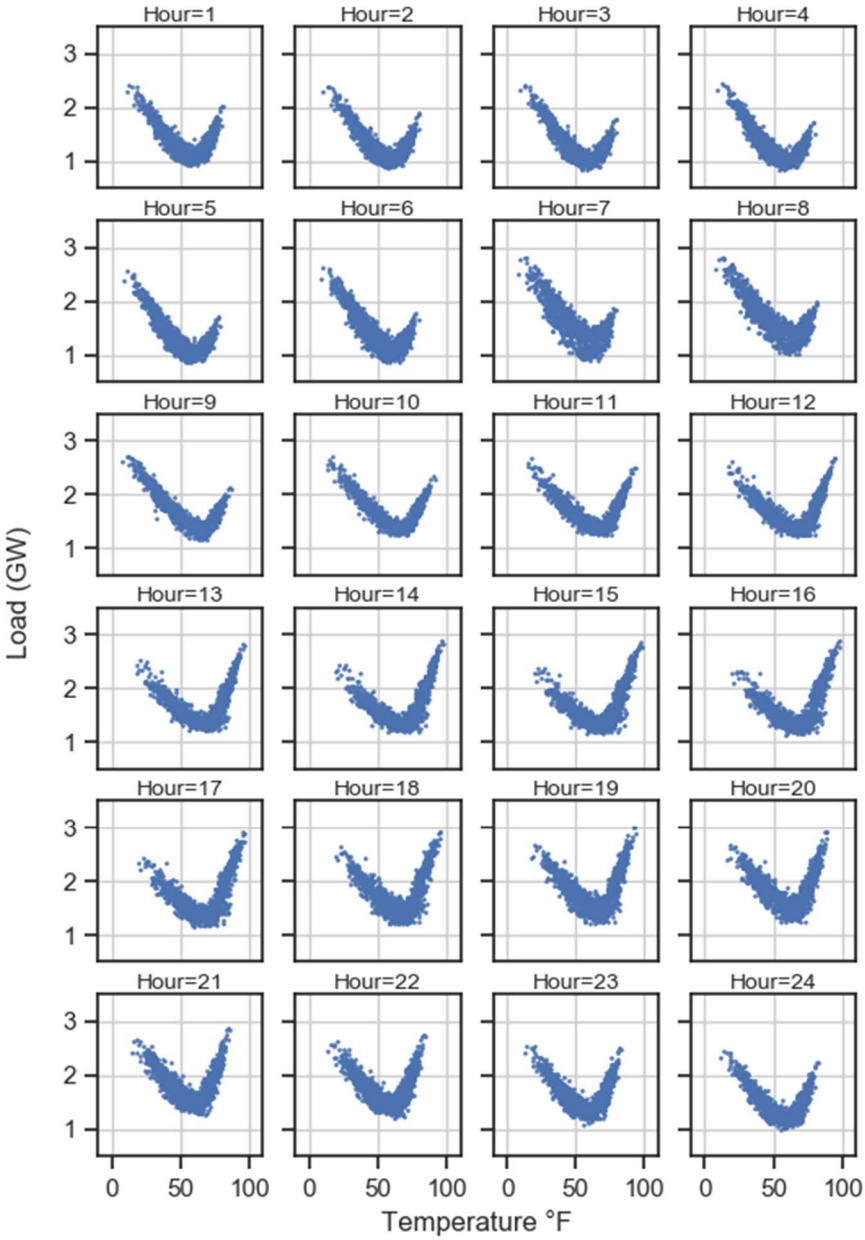

FIGURE 7: Load-temperature Scatterplots for 24 Hours (GEFCom2012)

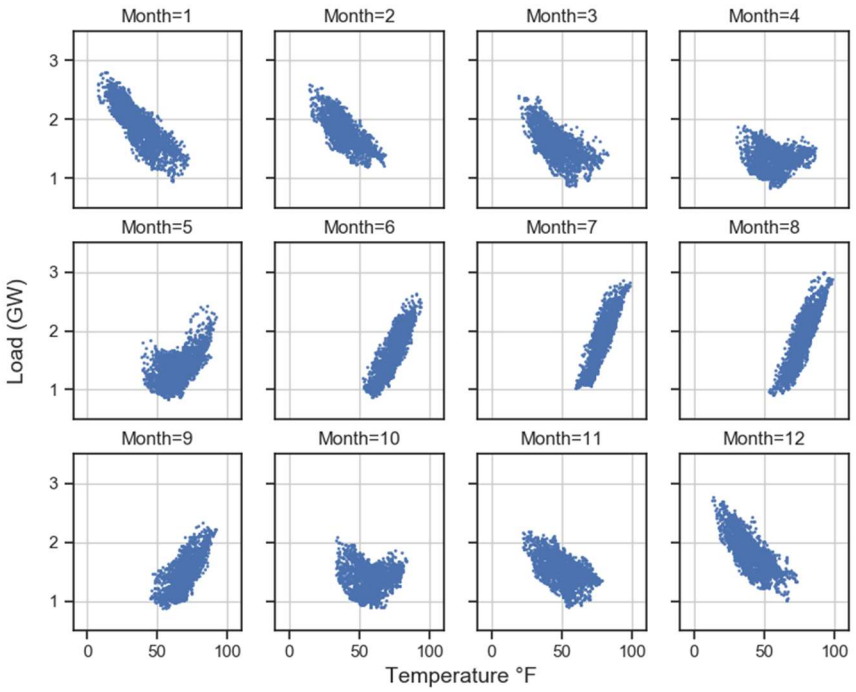

FIGURE 8: Load-temperature Scatterplots for 12 Months (GEFCom2012)

Temperature is known to have a strong correlation with electricity usage patterns. *FIGURE 9* shows the scatterplot of load – temperature relationship using three years of data (2004-2006) from the aggregated zone and the average temperature from the 11 stations. The graph shows a strong correlation (the typical "hockey stick" shape) between the load and temperature. On the left arm, the load goes up for heating needs during the winter when the temperature drops below a certain point. On the right arm, the load goes up for cooling needs during the summer when the temperature increases. This scatterplot shows that there is a cutoff point at around 60 Fahrenheit.

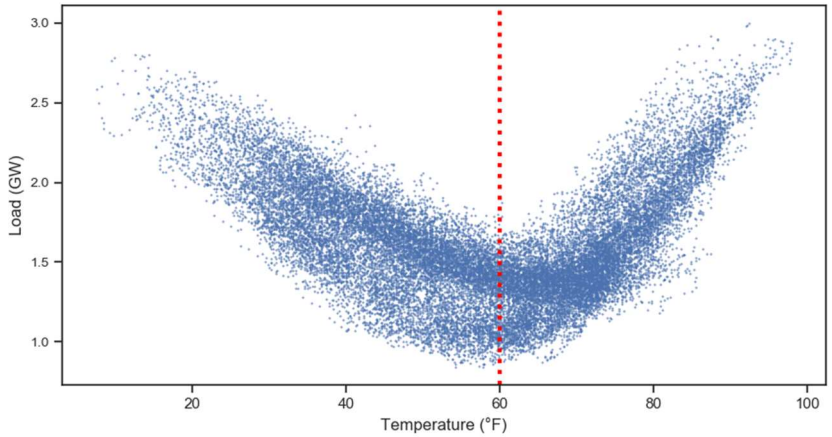

FIGURE 9: Load-temperature Scatterplot (GEFCom2012)

Four years (2004-2007) of hourly load and temperature data from all 11 weather stations are used in this case study, with three years being partitioned as the training period (2004-2006) and one year being held out as the test period (2007).

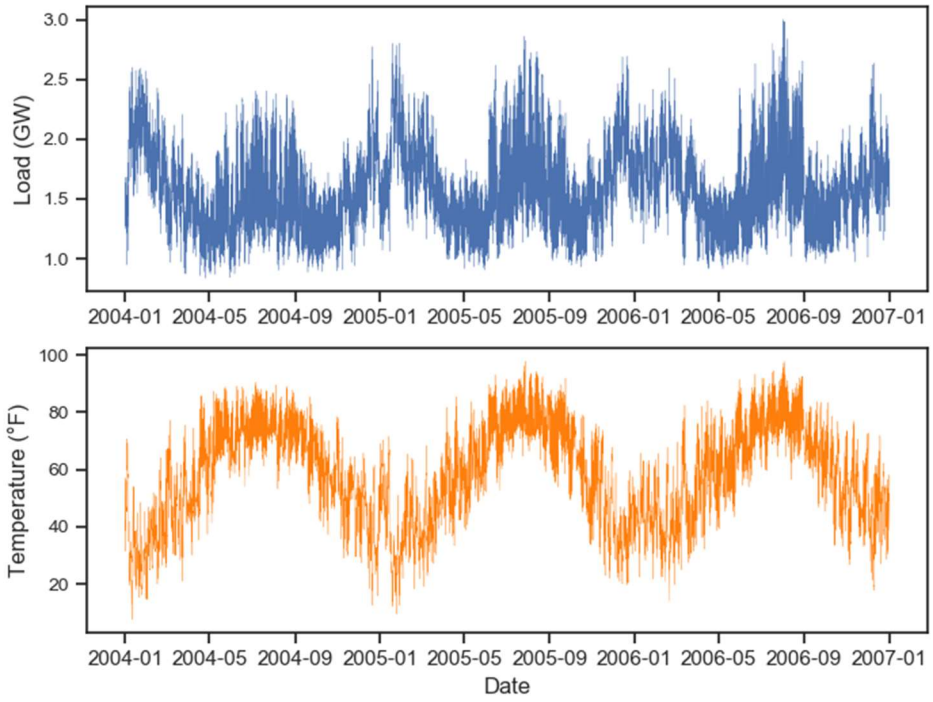

FIGURE 10: Time Series Plot of Hourly Load and Temperature (Zone 21, GEFCOM2012, 2004-2006)

Three years (2004-2006) of the hourly load plot at the aggregated load zone - Z_{21} and hourly temperature plot obtained from the average of 11 temperature series, are given

in *FIGURE 10*. For GEFCom2012 data, we have 11 weather stations data available and *TABLE 3* shows the statistics (mean, standard deviation, minimum and maximum) of temperature recorded at each station. It can be seen the temperature values vary a wide range, particularly for the minimum temperature. A random day – 1/20/2005 – is also picked out to demonstrate the variance of hourly temperature reported at different weather stations. As shown in *FIGURE 11* boxplots, the recorded temperature at 4 AM could range from ~20 to ~35 Fahrenheit among the weather stations.

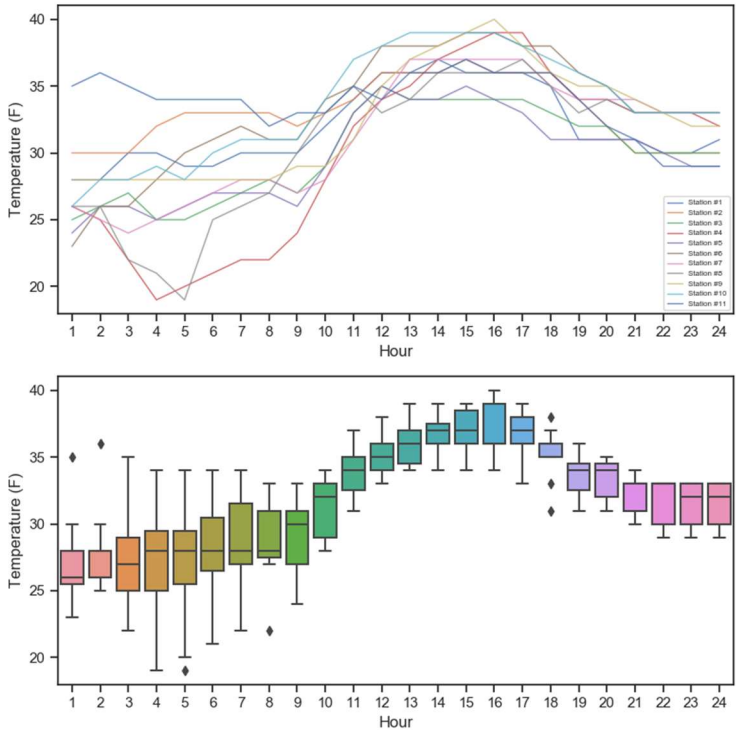

FIGURE 11: Line Plot (up) and Boxplots (down) of Reported Temperature at Each Weather Station for Zone 21 at 1/20/2005 (GEFCom2012)

TABLE 3: Statistics of Temperature (in Fahrenheit) Reported at Each Weather Station (GEFCom2012)

Station	Mean	Std.	Minimum	Maximum
1	59.78	16.79	12	103
2	54.94	17.44	0	95
3	56.20	17.65	8	98
4	60.25	17.42	11	103
5	56.87	17.65	6	99
6	58.85	17.09	7	103
7	58.98	18.04	8	104
8	59.33	17.78	6	103
9	56.81	18.00	5	100
10	58.93	17.43	5	102
11	55.11	17.97	0	98
Range	54.94 - 60.25	16.79 - 18.04	0 - 12	95 - 104

The proposed framework in (Hong et al., 2015), namely the GS-PS in this thesis, is used as the benchmark to justify the effectiveness of other approaches, and results are listed in *TABLE 4* through *TABLE 7*. The green highlighted MAPE values represent proposed approaches resulting in superior performance in specific zones compared to the benchmark method. The grey highlighted MAPE values represent proposed approaches with worse performance in specific zones compared to the benchmark. The buckets neither highlighted in green nor gray represent the proposed approaches leading to the same selection as the benchmark method.

A "Theoretical Optimum" column has been introduced, by going through each possible WSS and locate the selection resulting in the smallest error measure on the test year (2007) using two years (2005-2006) of data to estimate the parameters of the model. The "Dist. (%)" column is created and serves as another error metrics besides MAPE, which demonstrates the percentage difference between the MAPE of a proposed selection approach and the TO selection. At each zone, a forecast leading to a shorter distance (i.e., a smaller percentage difference) to the TO selection, is a selection closer to the TO. A

forecast resulting in zero distance to the TO selection indicates the method achieves the TO. For instance, the GS-IS shown in *TABLE 4* achieves the TO under zone 1 and 2.

The "No." column gives the number of stations selected under each load zone. The "Average (regular zones)" row gives the average number of stations being selected, the average MAPE (%) and the average Dist. (%) among the regular zones.

Among the 20 zones at the bottom level, Z_4 experienced a major outage and Z_9 is an industrial customer. In order for a meaningful comparison to (Hong et al., 2015), the forecasting MAPE results of Z_4 and Z_9 have been moved to the bottom of the tables, within the section of "Excluded Zones", and they are not in our scope of comparison.

TABLE 4: Experimental Results of TO and GS Methods (GEFCom2012)

		Theoretical Optimum		GS-PS (Benchmark)			GS-CV			GS-IS		
	Zone	No.	MAPE (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
	21	3	4.854	11	5.221	7.559	11	5.221	7.562	8	5.029	3.612
Regular Zones	1	2	6.947	3	7.008	0.880	3	7.008	0.879	3	7.008	0.879
	2	3	5.453	6	5.616	2.985	6	5.616	2.994	6	5.616	2.994
	3	3	5.453	6	5.616	2.985	6	5.616	2.994	6	5.616	2.994
	5	1	9.006	3	9.881	9.720	8	10.334	14.753	3	9.881	9.722
	6	3	5.398	7	5.553	2.873	6	5.559	2.989	6	5.559	2.989
	7	3	5.453	6	5.616	2.985	6	5.616	2.994	6	5.616	2.994
	8	3	7.235	4	7.500	3.660	4	7.500	3.664	3	7.478	3.357
	10	2	6.371	6	6.696	5.108	4	6.509	2.178	4	6.509	2.178
	11	4	7.596	4	7.699	1.358	4	7.699	1.356	3	7.813	2.856
	12	1	6.739	4	6.781	0.621	4	6.781	0.620	3	6.741	0.035
	13	1	7.279	4	7.391	1.543	3	7.294	0.211	3	7.294	0.211
	14	2	9.242	5	9.381	1.502	5	9.381	1.505	2	9.242	0.000
	15	1	7.425	2	7.438	0.173	2	7.438	0.177	1	7.425	0.000
	16	4	7.961	7	8.124	2.047	7	8.124	2.049	7	8.124	2.049
	17	8	5.233	6	5.262	0.551	6	5.262	0.553	6	5.262	0.553
	18	4	6.389	1	6.724	5.245	1	6.724	5.247	3	6.674	4.467
	19	4	7.873	3	7.900	0.340	3	7.900	0.346	3	7.900	0.346
	20	3	5.495	6	5.745	4.549	5	5.697	3.667	4	5.643	2.700
	Average (regular zones)	2.9	6.808	4.6	6.996	2.729	4.6	7.003	2.732	4.0	6.967	2.296
Excluded Zones	4	2	15.559	2	16.075	3.316	2.000	16.075	3.316	2	16.075	3.316
	9	2	136.649	9	139.133	1.818	9.000	139.133	1.817	7	138.384	1.269

TABLE 5: Experimental Results of ES Methods (GEFCom2012)

	Zone	ES-PS			ES-CV			ES-IS		
		No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
Regular Zones	21	3	5.166	6.425	5	5.027	3.556	4	4.968	2.357
	1	2	6.947	0.000	2	6.947	0.000	3	7.008	0.879
	2	4	5.688	4.308	5	5.611	2.890	5	5.498	0.814
	3	4	5.688	4.309	5	5.611	2.890	5	5.498	0.814
	5	3	11.156	23.874	5	10.501	16.609	5	10.170	12.928
	6	3	5.739	6.311	5	5.590	3.550	5	5.436	0.699
	7	4	5.688	4.309	5	5.611	2.890	5	5.498	0.814
	8	3	7.385	2.073	3	7.385	2.073	3	7.478	3.357
	10	3	6.535	2.582	3	6.451	1.267	5	6.621	3.934
	11	3	7.791	2.569	4	7.699	1.356	3	7.813	2.856
	12	3	6.896	2.322	4	6.781	0.620	3	6.741	0.035
	13	5	7.595	4.347	4	7.392	1.554	3	7.294	0.211
	14	3	9.319	0.833	3	9.319	0.833	3	9.319	0.833
	15	3	7.551	1.695	3	7.551	1.695	1	7.425	0.000
	16	2	8.366	5.084	4	8.060	1.248	5	7.995	0.424
	17	4	5.291	1.112	5	5.241	0.157	3	5.299	1.256
	18	4	6.744	5.558	4	6.783	6.171	3	6.700	4.865
	19	4	8.112	3.036	3	7.900	0.346	3	7.900	0.346
	20	5	5.640	2.634	4	5.622	2.306	4	5.622	2.306
	Average (regular zones)	3.4	7.118	4.275	3.9	7.003	2.692	3.7	6.962	2.076
Excluded Zones	4	4	16.153	3.820	4	16.153	3.820	4	15.928	2.373
	9	3	139.087	1.784	4	140.373	2.725	4	138.975	1.702

TABLE 6: Experimental Results of FS Methods (GEFCom2012)

	Zone	FS-PS			FS-CV			FS-IS		
		No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
Regular Zones	21	3	5.166	6.425	5	5.072	4.495	6	4.975	2.493
	1	2	6.947	0.000	2	6.947	0.000	3	7.008	0.879
	2	4	5.688	4.308	5	5.611	2.890	5	5.498	0.814
	3	4	5.688	4.309	5	5.611	2.890	5	5.498	0.814
	5	3	11.156	23.874	4	10.346	14.881	5	10.170	12.928
	6	4	5.660	4.856	5	5.590	3.550	5	5.436	0.699
	7	4	5.688	4.309	5	5.611	2.890	5	5.498	0.814
	8	3	7.385	2.073	3	7.385	2.073	3	7.478	3.357
	10	3	6.451	1.267	3	6.451	1.267	5	6.621	3.934
	11	3	7.791	2.569	4	7.699	1.356	3	7.813	2.856
	12	3	6.896	2.322	4	6.781	0.620	3	6.741	0.035
	13	3	7.562	3.888	4	7.392	1.554	3	7.294	0.211
	14	3	9.319	0.833	3	9.319	0.833	3	9.319	0.833
	15	3	7.551	1.695	3	7.551	1.695	1	7.425	0.000
	16	2	8.366	5.084	2	8.366	5.084	3	8.330	4.633
	17	4	5.291	1.112	5	5.241	0.157	3	5.299	1.256
	18	3	6.902	8.030	2	6.886	7.781	3	6.700	4.865
	19	4	8.112	3.036	3	7.900	0.346	3	7.900	0.346
	20	5	5.640	2.634	4	5.622	2.306	4	5.622	2.306
	Average (regular zones)	3.3	7.116	4.233	3.7	7.017	2.899	3.6	6.980	2.310
Excluded Zones	4	4	16.153	3.820	4	16.153	3.820	2	16.075	3.316
	9	3	139.087	1.784	4	140.373	2.725	4	139.450	2.050

TABLE 7: Experimental Results of BS Methods (GEFCom2012)

	Zone	BS-PS			BS-CV			BS-IS		
		No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
	21	6	4.924	1.434	6	4.924	1.434	7	4.945	1.871
Regular Zones	1	7	7.473	7.578	7	7.473	7.578	4	7.245	4.293
	2	10	5.880	7.824	10	5.880	7.824	5	5.497	0.812
	3	10	5.880	7.824	10	5.880	7.824	5	5.497	0.812
	5	6	11.024	22.410	6	11.024	22.410	5	10.243	13.740
	6	10	5.856	8.485	10	5.856	8.485	5	5.517	2.210
	7	10	5.880	7.824	10	5.880	7.824	5	5.497	0.812
	8	7	7.601	5.051	10	8.068	11.506	5	7.687	6.249
	10	9	7.139	12.066	10	7.275	14.191	5	6.912	8.500
	11	10	8.792	15.753	8	8.774	15.515	5	8.151	7.309
	12	7	7.915	17.454	9	7.750	14.996	7	7.667	13.770
	13	8	8.023	10.221	8	7.802	7.196	6	7.476	2.711
	14	9	9.972	7.899	10	9.968	7.851	9	9.689	4.839
	15	9	7.919	6.648	10	7.984	7.526	9	7.773	4.690
	16	8	8.802	10.563	10	8.521	7.033	10	8.723	9.566
	17	7	5.374	2.687	10	5.396	3.120	5	5.410	3.372
	18	6	6.841	7.080	10	6.728	5.307	5	6.732	5.368
	19	6	8.424	6.996	10	8.294	5.348	7	8.105	2.950
	20	10	5.968	8.615	11	5.929	7.892	7	5.709	3.894
	Average (regular zones)	8.3	7.487	9.610	9.4	7.471	9.413	6.1	7.196	5.328
Excluded Zones	4	5	16.483	5.938	5	16.483	5.938	10	15.666	0.685
	9	8	138.921	1.663	9	139.496	2.083	9	135.648	-0.733

The ranking of MAPE (%) along with the number of selected stations is illustrated in *TABLE 8*. The performance of the benchmark are highlighted in dark green and the performance of TO selection are highlighted in pink. In general, the GS-IS, ES-IS, and FS-IS lead to better MAPE (%) compared to the benchmark method at both the top level (Z_{21}) and the bottom level (average of the 18 regular zones). Among the 12 methods including the benchmark, the ES-IS results in the lowest average MAPE among the 18 regular zones and ranks 3rd at the top level (Z_{21}); the GS-IS ranks 2nd on the average of the 18 regular zones and 6th at the top level (Z_{21}); the FS-IS ranks 3rd on the average of the 18 regular zones and 4th at the top level (Z_{21}).

At the top level (Z_{21}), except for GS-CV which owns the same WSS as the benchmark, all the other proposed approaches select fewer weather stations and lead to lower MAPEs compared to the benchmark.

At the bottom level, *TABLE 8* shows a trend that the IS methods lead to superior results than the CV methods and PS methods lead to the worst. We make an extensive discussion on this topic in Section 5.3 – Section 5.5.

Regarding the number of stations being selected, the TO selections suggest fewer stations can lead to better results. We see the trend that the BS methods tend to select the most stations. The FS methods tend to select the fewest while the GS methods tend to lay in the middle. We make an extensive discussion on their WSS behavior in Section 5.6.

TABLE 8: Ranking under Average Regular Zones and Aggregated Zone (GEFCom2012)

	Avg of 18 Regular Zones			Aggregated Zone (Z_{21})		
Rank	Avg MAPE (%)		Avg No.	MAPE (%)		No.
	TO	6.808	2.9	TO	4.854	3
1	ES-IS	6.962	3.7	BS-CV	4.924	6
2	GS-IS	6.967	4.0	BS-PS	4.924	6
3	FS-IS	6.980	3.6	BS-IS	4.945	7
4	GS-PS	6.996	4.6	ES-IS	4.968	4
5	ES-CV	7.003	3.9	FS-IS	4.975	6
6	GS-CV	7.003	4.6	ES-CV	5.027	5
7	FS-CV	7.017	3.7	GS-IS	5.029	8
8	FS-PS	7.116	3.3	FS-CV	5.072	5
9	ES-PS	7.118	3.4	ES-PS	5.166	3
10	BS-IS	7.196	6.1	FS-PS	5.166	3
11	BS-CV	7.471	9.4	GS-CV	5.221	11
12	BS-PS	7.487	8.3	GS-PS	5.221	11

5.2. GEFCom2014 Case Study

In order to examine the generalization capability of the proposed approaches, we conduct another case study using the data from the load forecasting track of GEFCom2014. Using this public dataset allows our results to be reproduced and conveniently compared for future studies by other researchers.

The probabilistic load forecasting track in GEFCom2014 involved a total of six years (2005-2010) of hourly load data and 10 years (2001-2010) of weather data. The weather data was formed by temperature series from 25 anonymous weather stations ($W_1 - W_{25}$). Six years (2005-2010) of load and temperature data are used in this case study. *TABLE 9* gives the statistics (mean, standard deviation, minimum and maximum) of temperature recorded at each station. It can be seen the temperature values vary a wide range, particularly for the minimum temperature.

Sliding simulation is a widely used forecast evaluation technique while it mimics the forecasting operations in the real world. Specifically, the sliding simulation technique applies a rolling forecast using data from a pre-defined length of the historical window (e.g., three years) to predict a period (e.g., one year) (Tashman, 2000). Since we have six full years (2005-2010) in the load data history, we first use the data from 2005 to 2007 as the training data to forecast the year of 2008. Then we advance the forecast origin by one year to forecast the year 2009 using the data from 2006 to 2008 as the training data. We repeat the rolling process until we complete the forecasts for all three years (2008-2010). *FIGURE 12* presents the sliding simulation in this case study.

TABLE 9: Statistics of temperature (in Fahrenheit) reported at each weather station (GEFCom2014)

Station	Mean	Std.	Minimum	Maximum
1	61.35	17.58	9	104
2	61.25	17.24	9	104
3	55.70	16.60	4	93
4	61.32	16.80	11	103
5	63.27	15.98	17	102
6	61.39	16.53	15	101
7	62.38	16.26	15	101
8	63.21	16.43	12	104
9	54.02	16.41	0	93
10	62.45	16.94	12	104
11	60.05	17.13	11	100
12	60.93	16.98	9	102
13	59.49	16.95	5	104
14	63.34	14.60	21	94
15	60.43	17.28	9	103
16	63.77	15.81	14	101
17	59.88	17.39	9	102
18	63.73	14.94	16	98
19	58.63	16.85	7	100
20	62.78	16.11	12	100
21	61.12	16.61	16	104
22	62.02	17.06	9	102
23	62.23	17.54	10	104
24	61.43	17.39	11	104
25	60.50	17.37	9	104
Range	54.02 - 63.77	14.60 - 17.58	0 - 21	93 - 104

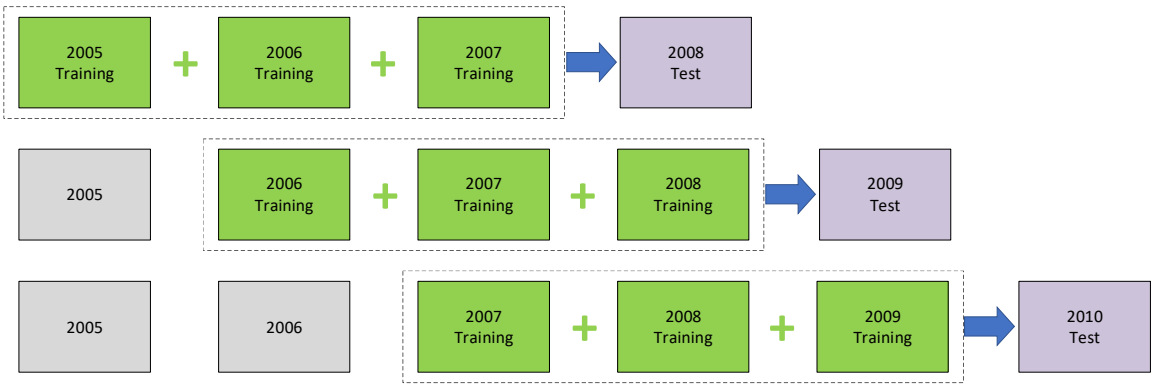

FIGURE 12: Demonstration of sliding simulation in GEFCom2014 case study

Like the GEFCom2012 case study, the WSS framework in (Hong et al., 2015) is used as the benchmark to justify the effectiveness of our proposed approaches and results are listed in *TABLE 10* through *TABLE 13*. The green highlighted MAPE values represent

proposed approaches with better forecasting performance at a specific test year comparing to the benchmark method. The grayed out MAPE buckets indicate proposed approaches that are inferior to the benchmark method.

Same to the GEFCom2012 case study, the "Theoretical Optimum" column has been introduced, by going through each possible WSS and locate the selection resulting in the smallest MAPE error on the test year (2008, 2009, 2010) using the corresponding two years (2006-2007, 2007-2008, 2008-2009) of data to estimate the parameters of the model. The "Dist. (%)" column is created and serves as another error metrics besides MAPE, which demonstrates the percentage difference between the MAPE of a proposed selection approach and the TO selection. At each specific test year under the proposed approaches, a forecast leading to a shorter distance (i.e., a smaller percentage difference) to the TO selection, is WSS closer to the TO.

TABLE 10: Experimental Results of TO and GS Methods (GEFCom2014)

Test Year	*Theoretical Optimum*		GS-PS (Benchmark)			GS-CV			GS-IS		
	No.	MAPE (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
2008	9	5.199	12	5.455	4.935	14	5.435	4.544	14	5.435	4.544
2009	6	5.919	17	6.579	11.138	14	6.628	11.972	12	6.838	15.521
2010	8	5.247	11	5.798	10.512	10	5.790	10.360	10	5.790	10.360
Average	7.7	5.455	13.3	5.944	8.862	12.7	5.951	10.379	12.0	6.021	10.141

TABLE 11: Experimental Results of ES Methods (GEFCom2014)

Test Year	ES-PS			ES-CV			ES-IS		
	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
2008	11	5.307	2.086	9	5.244	0.869	10	5.338	2.677
2009	9	6.302	6.465	10	6.412	8.320	11	6.292	6.301
2010	6	5.646	7.622	7	5.584	6.436	10	5.401	2.942
Average		5.752	6.586		5.747	5.208		5.677	3.973

TABLE 12: Experimental Results of FS Methods (GEFCom2014)

Test Year	FS-PS			FS-CV			FS-IS		
	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
2008	11	5.309	2.120	10	5.233	0.663	10	5.338	2.677
2009	9	6.302	6.465	8	6.836	15.492	11	6.292	6.301
2010	6	5.646	7.622	7	5.584	6.431	10	5.401	2.942
Average	8.7	5.752	5.402	8.3	5.884	5.613	10.3	5.677	3.973

TABLE 13: Experimental Results of BS Methods (GEFCom2014)

Test Year	BS-PS			BS-CV			BS-IS		
	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
2008	18	5.689	9.430	12	5.528	6.338	14	5.656	8.807
2009	16	6.796	14.814	25	7.021	18.611	15	6.583	11.209
2010	17	6.613	26.047	18	6.373	21.470	17	6.170	17.603
Average	17.0	6.366	16.764	18.3	6.307	15.629	15.3	6.136	12.540

Based on the average MAPE among the three test years, all the ES methods and FS methods perform better than the benchmark. The ranking of MAPE (%) and the number of stations being selected are listed in *TABLE 14*, with the benchmark selection highlighted in dark green and TO selection highlighted in pink.

Regarding the number of stations being selected, the TO selection suggests fewer stations can lead to better results. Like the GEFCom2012 case study, we see a strong trend that the BS methods tend to select the most stations. The FS methods tend to select the fewest while the GS methods tend to lay in the middle. We make an extensive discussion on their WSS behavior in Section 5.6. The ES-PS and FS-PS give very close results with ES-PS being marginally better (5.7518% vs. 5.7524%). Interestingly, the ES-IS and FS-IS in this case study result in the same WSSs among all three test years. Comparing to the benchmark method, they improve the MAPE by 4.49% (i.e., from 5.944% to 5.677%).

TABLE 14: Ranking under Average Test Year (GEFCom2014)

Rank	Method	MAPE (%)	Avg No.
	TO	5.455	7.7
1	ES-IS	5.677	10.3
1	FS-IS	5.677	10.3
3	ES-CV	5.747	8.7
4	ES-PS	5.752	8.7
5	FS-PS	5.752	8.7
6	FS-CV	5.884	8.3
7	GS-PS	5.944	13.3
8	GS-CV	5.951	12.7
9	GS-IS	6.021	12.0
10	BS-IS	6.136	15.3
11	BS-CV	6.307	18.3
12	BS-PS	6.366	17.0

5.3. An Overfitting Issue

Training an algorithm and evaluate its statistical performance on the same data set could result in the data to be over-trained and cause the overfitting issue (Arlot & Celisse, 2010). We first analyze the overfitting issue on the GEFCom2012 case study, in which we compare the results of ES-PS to the benchmark method (GS-PS). The ES-PS searches through each possible WSS, records the selection resulting in the smallest error measure in the validation period and uses it to forecast in the test year. Under the aggregated zone as well as the regular zones, it achieves WSSs leading to lower MAPEs than the benchmark method in the validation period. Nevertheless, the superior forecasting performance on the validation period does not guarantee the same advantage in the test period. As shown in *TABLE 15*, the ES-PS results in worse forecasting performance in 13 out of 18 regular zones under the test period. The average MAPE of the 18 regular zones under ES-PS is also higher than the one under GS-PS.

The above illustrates a common overfitting issue on the validation data. The cross-validation (CV) technique has been known and widely used to remediate this issue and improve the forecasting performance. In our GEFCom2012 case study, we adopt the V-fold cross validation (VFCV) technique to the ES and compare the results to the ES-PS method. In *TABLE 16*, the bold MAPE values indicate a method is leading to superior forecasting performance than the other. Among the 18 regular zones, the ES-CV results in lower MAPEs in 13 zones. Under the ES and comparing to the traditional post-sample fit, the cross-validation technique reduces the forecasting error (MAPE) by 1.6% (i.e., from 7.118% to 7.003%) on the average of 18 zones, and 2.7% (i.e., from 5.116% to 5.027%) under the aggregated zone (Z_{21}).

Note that from *TABLE 8*, the ES-CV outperforms the benchmark (GS-PS) at the aggregated zone but not on the average of regular zones, which may be due to some overfitting issue still exists after CV is implemented.

TABLE 15: Results Comparison Between GS-PS and ES-PS (GEFCom2012)

		GS-PS (Benchmark)				ES-PS			
	Zone	No.	Test MAPE (%)	Dist. (%)	Validation MAPE (%)	No.	Test MAPE (%)	Dist. (%)	Validation MAPE (%)
	21	11	5.221	7.559	4.912	3	5.166	6.425	**4.589**
Regular Zones	1	3	7.008	0.880	6.662	2	6.947	0.000	**6.661**
	2	6	5.616	2.985	5.025	4	5.688	4.308	**4.806**
	3	6	5.616	2.985	5.025	4	5.688	4.309	**4.806**
	5	3	9.881	9.720	9.549	3	11.156	23.874	**8.961**
	6	7	5.553	2.873	5.153	3	5.739	6.311	**4.883**
	7	6	5.616	2.985	5.025	4	5.688	4.309	**4.806**
	8	4	7.500	3.660	6.115	3	7.385	2.073	**5.984**
	10	6	6.696	5.108	5.991	3	6.535	2.582	**5.876**
	11	4	7.699	1.358	6.244	3	7.791	2.569	**6.211**
	12	4	6.781	0.621	7.345	3	6.896	2.322	**7.173**
	13	4	7.391	1.543	7.659	5	7.595	4.347	**7.596**
	14	5	9.381	1.502	10.176	3	9.319	0.833	**9.921**
	15	2	7.438	0.173	7.977	3	7.551	1.695	**7.756**
	16	7	8.124	2.047	9.967	2	8.366	5.084	**9.269**
	17	6	5.262	0.551	5.153	4	5.291	1.112	**5.091**
	18	1	6.724	5.245	7.500	4	6.744	5.558	**7.216**
	19	3	7.900	0.340	8.389	4	8.112	3.036	**8.325**
	20	6	5.745	4.549	5.262	5	5.640	2.634	**5.123**
	Average (regular zones)		6.996	2.729	6.901		7.118	4.275	**6.692**

TABLE 16: Results Comparison Between ES-PS and ES-CV (GEFCom2012)

	Zone	ES-PS			ES-CV		
		No.	Test MAPE (%)	Dist. (%)	No.	Test MAPE (%)	Dist. (%)
Regular Zones	21	3	5.166	6.425	5	**5.027**	3.556
	1	2	6.947	0.000	2	6.947	0.000
	2	4	5.688	4.308	5	**5.611**	2.890
	3	4	5.688	4.309	5	**5.611**	2.890
	5	3	11.156	23.874	5	**10.501**	16.609
	6	3	5.739	6.311	5	**5.590**	3.550
	7	4	5.688	4.309	5	**5.611**	2.890
	8	3	7.385	2.073	3	7.385	2.073
	10	3	6.535	2.582	3	**6.451**	1.267
	11	3	7.791	2.569	4	**7.699**	1.356
	12	3	6.896	2.322	4	**6.781**	0.620
	13	5	7.595	4.347	4	**7.392**	1.554
	14	3	9.319	0.833	3	9.319	0.833
	15	3	7.551	1.695	3	7.551	1.695
	16	2	8.366	5.084	4	**8.060**	1.248
	17	4	5.291	1.112	5	**5.241**	0.157
	18	4	**6.744**	5.558	4	6.783	6.171
	19	4	8.112	3.036	3	**7.900**	0.346
	20	5	5.640	2.634	4	**5.622**	2.306
	Average (regular zones)		7.118	4.275		**7.003**	2.692

We extend the same experiment to the GEFCom2014 data (*TABLE 17*) and we have not noticed significant overfitting issue among the test years.

TABLE 17: Results Comparison Between GS-PS and ES-PS (GEFCom2014)

Test Year	GS-PS (Benchmark)				ES-PS			
	No.	MAPE (%)	Dist. (%)	Validation MAPE (%)	No.	MAPE (%)	Dist. (%)	Validation MAPE (%)
2008	12	5.455	4.935	5.658	11	5.307	2.086	**5.361**
2009	17	6.579	11.138	5.439	9	6.302	6.465	**5.199**
2010	11	5.798	10.512	6.510	6	5.646	7.622	**5.919**
Average		5.944	8.862	5.869		5.752	6.586	**5.493**

We further employ the V-fold cross validation (VFCV) technique to ES and compare the results to the ES-PS method under the GEFCom2014 case study. In *TABLE*

18, the green highlights indicate the superior forecasting performance comparing to the benchmark, and the bold MAPE values represent a method (between ES-PS and ES-CV) with superior forecasting performance comparing to the other. Among the three test years, the ES-CV results in lower MAPEs in the years of 2008 and 2010. In this case, the average MAPE shows the cross-validation technique gives marginal forecasting error improvement by 0.08% (i.e., from 5.752% to 5.747%) compared to the traditional post-sample fit under the ES.

TABLE 18: Results Comparison Between ES-PS and ES-CV (GEFCom2014)

Test Year	GS-PS (Benchmark)			ES-PS			ES-CV		
	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)	No.	MAPE (%)	Dist. (%)
2008	12	5.455	4.935	11	5.307	2.086	9	**5.244**	0.869
2009	17	6.579	11.138	9	**6.302**	6.465	10	6.412	8.320
2010	11	5.798	10.512	6	5.646	7.622	7	**5.584**	6.436
Average		5.944	8.862		5.752	6.586		**5.747**	5.208

FIGURE 13 gives a visual inspection of MAPE (%) comparison between the CV and PS based methods. More detail level comparisons are presented in *TABLE 19* and *TABLE 20*, while *TABLE 19* includes the aggregated zone. In the majority of the load zones (GEFCom2012) and test years (GEFCom2014), the exceeding performance under the CV based methods provide proof that the CV methods could result in a more robust WSS than the PS methods. Nevertheless, one needs to be aware that CV methods can increase the computational cost substantially depending on the size of the holdout sample we choose. Besides, the CV methods show more advantages when the training process is highly likely to result in overfitting issues on a single validation period, like ES-PS. When we implement CV and PS methods under the heuristic frameworks (GS, FS, and BS), the superiority of

the CV methods is not that evident, especially in the GEFCom2014 case study (*FIGURE 13*, right).

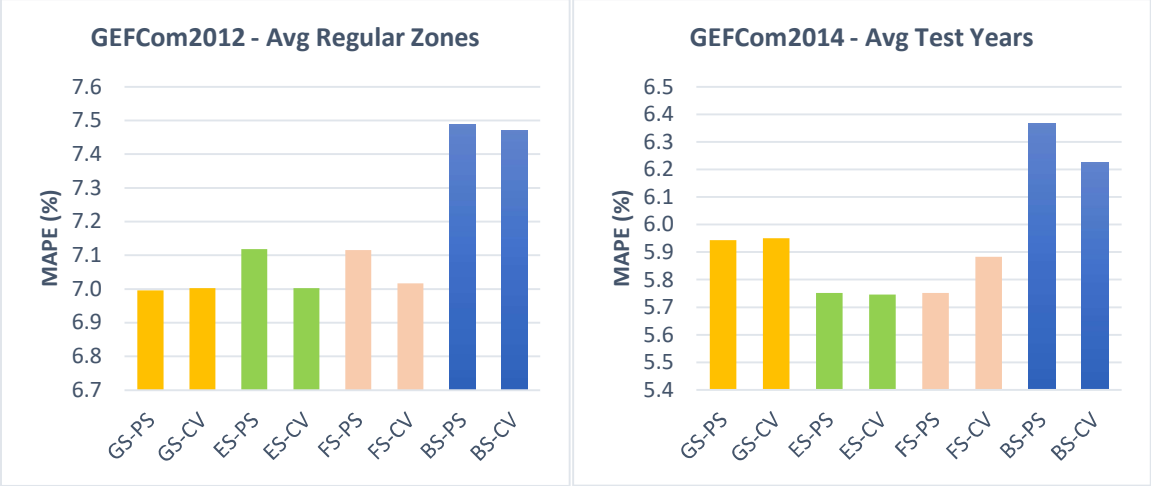

FIGURE 13: Performance Comparison of CV and PS Based Methods: GEFCom2012 (left) and GEFCom2014 (right)

TABLE 19: Comparison Summary Between CV and PS Based Methods (GEFCom2012)

Framework	No. of zones CV is better	No. of zones PS is better	No. of zones two methods lead to the same selection
GS	**3**	2	14
ES	**14**	1	4
FS	**13**	0	6
BS	**8**	4	7

TABLE 20: Comparison Summary Between CV and PS Based Methods (GEFCom2014)

Framework	No. of years CV is better	No. of years PS is better
GS	**2**	1
ES	**2**	1
FS	**2**	1
BS	**2**	1

5.4. In-sample vs. Post-sample based selections

It is widely known that a good in-sample fit does not necessarily lead to a good post-sample forecast. A prevailing issue of overfitting sometimes occurs when a complex model leading to very small in-sample fit errors generates a terrible forecast on the post-

sample data. On the other side, the post-sample fit requires the history data to be divided into two parts, one of which is held out for validation/testing purposes and the rest is used as the training set. One advantage the post-sample fit can have is to avoid the overfitting issue on the training sample set. The downside is obvious – it is difficult to implement when we have a lack of data; it could also result in another overfitting issue on the validation data, as indicated in Section 5.3.

For the GEFCom2012 case study, we compare the MAPE results from in-sample (IS) methods and the post-sample (PS) methods under the four frameworks presented in this thesis. In *TABLE 21* and *TABLE 22*, the bold MAPE values represent the statistical test method (either IS or PS) leading to a superior forecasting performance comparing to the other under each framework. For all four frameworks, the IS methods result in lower average MAPE values among the 18 regular zones. The IS methods under GS, ES and FS also result in lower MAPE at the top level (Z_{21}). The summary of results comparison shown in TABLE 23 concludes the IS methods lead to superior forecasting accuracy than the PS methods among most of the regular zones.

TABLE 21: Results Comparison Between IS and PS under GS and ES (GEFCom2012)

	Zone	GS-PS		GS-IS		ES-PS		ES-IS	
		No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)
	21	11	5.221	8	**5.029**	3	5.166	4	**4.968**
Regular Zones	1	3	7.008	3	7.008	2	**6.947**	3	7.008
	2	6	5.616	6	5.616	4	5.688	5	**5.498**
	3	6	5.616	6	5.616	4	5.688	5	**5.498**
	5	3	9.881	3	9.881	3	11.156	5	**10.170**
	6	7	**5.553**	6	5.559	3	5.739	5	**5.436**
	7	6	5.616	6	5.616	4	5.688	5	**5.498**
	8	4	7.500	3	**7.478**	3	**7.385**	3	7.478
	10	6	6.696	4	**6.509**	3	**6.535**	5	6.621
	11	4	**7.699**	3	7.813	3	**7.791**	3	7.813
	12	4	6.781	3	**6.741**	3	6.896	3	**6.741**
	13	4	7.391	3	**7.294**	5	7.595	3	**7.294**
	14	5	9.381	2	**9.242**	3	9.319	3	9.319
	15	2	7.438	1	**7.425**	3	7.551	1	**7.425**
	16	7	8.124	7	8.124	2	8.366	5	**7.995**
	17	6	5.262	6	5.262	4	**5.291**	3	5.299
	18	1	6.724	3	**6.674**	4	6.744	3	**6.700**
	19	3	7.900	3	7.900	4	8.112	3	**7.900**
	20	6	5.745	4	**5.643**	5	5.640	4	**5.622**
	Average (regular zones)	4.6	6.996	4.0	**6.967**	3.4	7.118	3.7	**6.962**

TABLE 22: Results Comparison Between IS and PS Methods under FS and BS (GEFCom2012)

	Zone	FS-PS		FS-IS		BS-PS		BS-IS	
		No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)
	21	3	5.166	6	**4.975**	6	**4.924**	7	4.945
Regular Zones	1	2	**6.947**	3	7.008	7	7.473	4	**7.245**
	2	4	5.688	5	**5.498**	10	5.880	5	**5.497**
	3	4	5.688	5	**5.498**	10	5.880	5	**5.497**
	5	3	11.156	5	**10.170**	6	11.024	5	**10.243**
	6	4	5.660	5	**5.436**	10	5.856	5	**5.517**
	7	4	5.688	5	**5.498**	10	5.880	5	**5.497**
	8	3	**7.385**	3	7.478	7	**7.601**	5	7.687
	10	3	**6.451**	5	6.621	9	7.139	5	**6.912**
	11	3	**7.791**	3	7.813	10	8.792	5	**8.151**
	12	3	6.896	3	**6.741**	7	7.915	7	**7.667**
	13	3	7.562	3	**7.294**	8	8.023	6	**7.476**
	14	3	9.319	3	9.319	9	9.972	9	**9.689**
	15	3	7.551	1	**7.425**	9	7.919	9	**7.773**
	16	2	8.366	3	**8.330**	8	8.802	10	**8.723**
	17	4	**5.291**	3	5.299	7	**5.374**	5	5.410
	18	3	6.902	3	**6.700**	6	6.841	5	**6.732**
	19	4	8.112	3	**7.900**	6	8.424	7	**8.105**
	20	5	5.640	4	**5.622**	10	5.968	7	**5.709**
	Average (regular zones)	3.3	7.116	3.6	**6.980**	8.3	7.487	6.1	**7.196**

TABLE 23: Comparison summary between IS and PS based methods (GEFCom2012)

Framework	No. of zones IS method is better	No. of zones PS method is better	No. of zones two methods lead to the same result
GS	**9**	2	8
ES	**13**	5	1
FS	**13**	5	1
BS	**16**	3	0

We extend the comparison using the GEFCom2014 data. In *TABLE 24* and *TABLE 25*, the bold MAPE values represent the statistical test method (either IS or PS) leading towards a superior forecasting performance comparing to the other under the same framework. As of ES, FS, and BS, the IS methods result in lower MAPE values among the average of three test years. The summary of results comparison shown in *TABLE 26* concludes the IS methods lead to superior forecasting accuracy than the PS methods among the majority of the test years.

TABLE 24: Results Comparison Between IS and PS Based Methods under GS and ES (GEFCom2014)

Test Year	GS-PS (Benchmark)		GS-IS		ES-PS		ES-IS	
	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)
2008	12	5.455	14	**5.435**	11	**5.307**	10	5.338
2009	17	**6.579**	12	6.838	9	6.302	11	**6.292**
2010	11	5.798	10	**5.790**	6	5.646	10	**5.401**
Average	13.3	**5.944**	12.0	6.021	8.7	5.752	10.3	**5.677**

TABLE 25: Results Comparison Between IS and PS Based Methods under FS and BS (GEFCom2014)

Test Year	FS-PS		FS-IS		BS-PS		BS-IS	
	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)	No.	MAPE (%)
2008	11	**5.309**	10	5.338	18	5.689	14	**5.656**
2009	9	6.302	11	**6.292**	16	6.796	15	**6.583**
2010	6	5.646	10	**5.401**	17	6.613	17	**6.170**
Average	8.7	5.752	10.3	**5.677**	17.0	6.366	15.3	**6.136**

TABLE 26: Comparison summary Between IS and PS Based methods (GEFCom2014)

Framework	No. of years IS is better	No. of years PS is better
GS	**2**	1
ES	**2**	1
FS	**2**	1
BS	**3**	0

FIGURE 14 gives a visual inspection of MAPE (%) comparison between the IS and PS based methods. Except for the GS-PS method in the GEFCom2014 case study, all the other methods indicate the performance superiority of IS methods. This can be explained in three-fold: 1) we use a relatively less complex model for the WSS and thus the IS fit is less likely to overfit the training data. 2) the PS methods require longer data history. Given our case study, we have three years of data in history to forecast a future year load. Using the PS fit requires us to hold out the most recent year in the data history for validation purposes. This gives us only 2 years for training and there may not be enough information to direct us to an optimal solution on the WSS. On the other hand, the IS fit is capable to incorporate the entire data history and come up with a proper WSS. 3) if the validation data we hold out could not well represent the behaviors in the test period, the PS fit could overfit on the validation period.

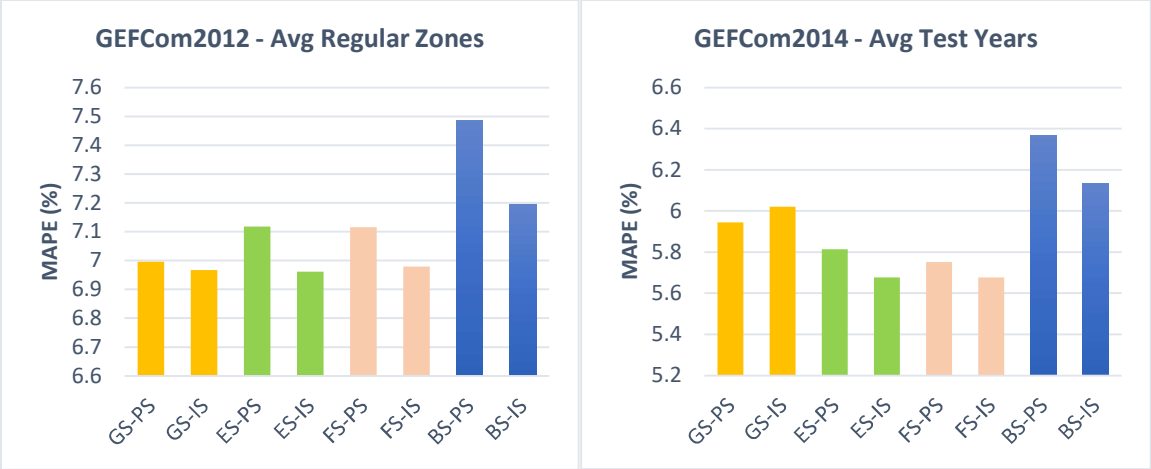

FIGURE 14: Performance Comparison of IS and PS Based Methods: GEFCom2012 (left) and GEFCom2014 (right)

5.5. In-sample vs. CV based selections

A natural extension of Section 5.3 and Section 5.4 would be to conduct a side-by-side comparison between the IS and CV based methods. *FIGURE 15* gives a visual inspection of MAPE (%) comparison between the IS and CV based methods. The graph indicates the superiority of IS based methods under almost all frameworks over the CV based methods. The only exception the GS framework under the GEFCom2014 data, where the MAPE difference between GS-IS and GS-CV is less than 0.07%.

More detail level comparisons are presented in *TABLE 27* and *TABLE 28*, while *TABLE 27* includes the aggregated zone. The above discussions (Section 5.3 – Section 5.5) may suggest a rule of thumb that when we lack data history (e.g., $\leq$ 3-4 years given the forecasting horizon as one year), the IS methods can be a better option for WSS. If we want to hold out a period for validation purposes and inspect the forecasting error on the validation period(s), the CV methods can be considered on top of IS and PS methods, since IS based methods do not provide out-of-sample error metrics and the PS based approaches tend to overfit on the validation data.

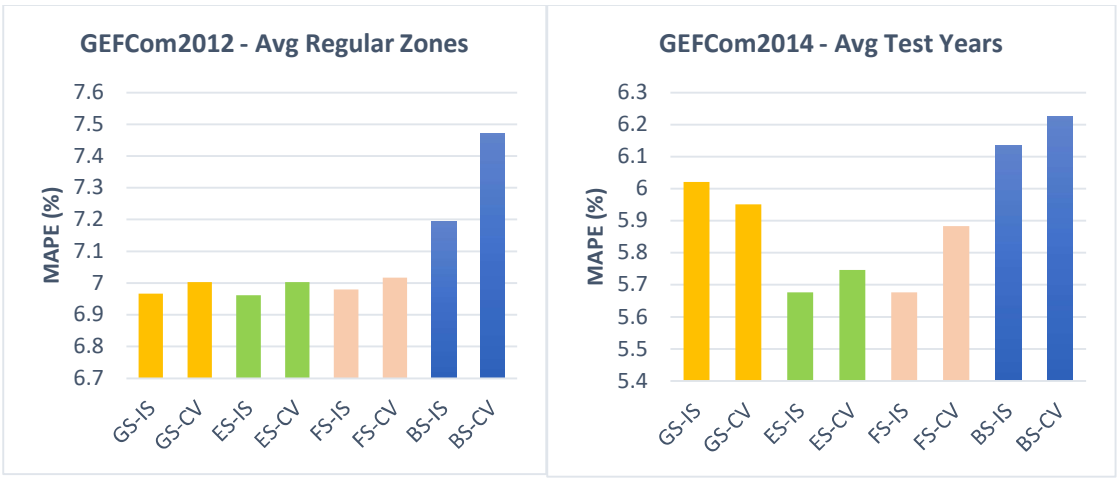

FIGURE 15: Performance Comparison of IS and CV Based Methods: GEFCom2012 (left) and GEFCom2014 (right)

TABLE 27: Comparison Summary Between IS and CV Based Methods (GEFCom2012)

Framework	No. of zones CV is better	No. of zones PS is better	No. of zones two methods lead to the same selection
GS	**8**	1	10
ES	**11**	5	3
FS	**11**	5	3
BS	**15**	4	0

TABLE 28: Comparison Summary Between IS and CV Based Methods (GEFCom2014)

Framework	No. of years CV is better	No. of years PS is better
GS	**2**	1
ES	**2**	1
FS	**2**	1
BS	**2**	1

5.6. Forward and Backward selection

Aiming at achieving good results while saving computational cost, both the forward and backward selection methods have been widely used as the heuristic methods in the fields. There hasn't been a formal study in the literature on using these two heuristic methods in the subject of WSS.

From the experiment results in both case studies, we recognize the BS methods tend to select the most amount of weather stations among the frameworks presented while the FS methods tend to select the fewest. The benchmark tends to lay in the middle.

For the GEFCom2012 case study, a boxplot of stations selected by each method among the regular zones is presented in *FIGURE 16*. The numbers of stations selected by each method under each load zone are gathered in *TABLE 29*, where the green highlights denote the selection is better than the benchmark method based on MAPE (%), while the grey buckets denote the selection is worse. The BS-PS and BS-CV select more weather stations among all 18 regular zones while all of which are inferior to the benchmark method. The BS-IS method selects more weather stations in 12 zones out of 18 regular zones and leads to worse MAPE comparing to the benchmark method in 13 out of 18 regular zones. However, the BS framework performs well at the top level (Z_{21}). As shown in *TABLE 8*, the three BS methods rank the top three among the 12 methods presented in this thesis while BS-CV and BS-PS result in the same WSS.

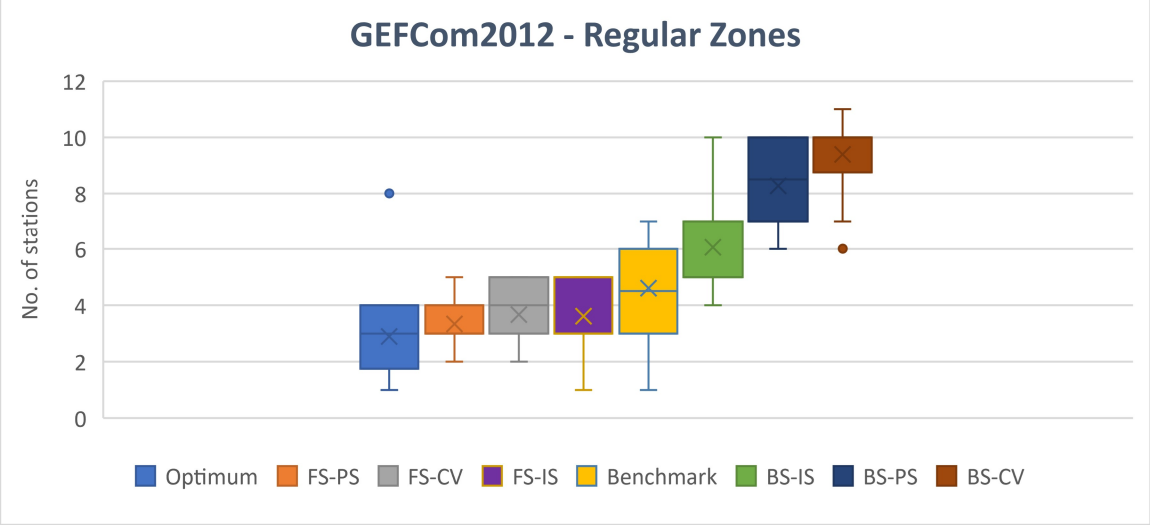

FIGURE 16: Boxplot of Stations Selected by Each Method under Regular Zones (GEFCom2012)

On the other side, the FS framework selects fewer weather stations in the majority of the regular zones compared to the benchmark method. The FS-PS selects fewer stations than the benchmark method in 14 regular zones and FS-CV selects fewer stations in 11. FS-PS is inferior to benchmark in 13 out of the 18 regular zones and FS-CV shows some improvements on top of FS-PS. However, the average MAPE of FS-CV is slightly inferior to the benchmark (*TABLE 8*). The FS-IS method selects fewer weather stations than the benchmark in 14 regular zones out of the 18, while in 12 out of the 18 regular zones, FS-IS leads to better forecasting performance. The FS-IS, in turn, results in better average MAPE among the 18 regular zones. At the top level (Z_{21}), FS-PS selects only three weather stations, FS-CV selects five and FS-IS selects six. All three FS methods lead to a lower MAPE than the benchmark, while their forecasting performance are not as good as the three BS methods.

TABLE 29: No. of Stations Selected by the TO, GS-PS (benchmark), FS and BS Methods (GEFCom2012)

	Zone	*Theoretical Optimum*	GS-PS	FS-PS	FS-CV	FS-IS	BS-PS	BS-CV	BS-IS
	21	3	11	3	5	6	6	6	7
Regular Zones	1	2	3	2	2	3	7	7	4
	2	3	6	4	5	5	10	10	5
	3	3	6	4	5	5	10	10	5
	5	1	3	3	4	5	6	6	5
	6	3	7	4	5	5	10	10	5
	7	3	6	4	5	5	10	10	5
	8	3	4	3	3	3	7	10	5
	10	2	6	3	3	5	9	10	5
	11	4	4	3	4	3	10	8	5
	12	1	4	3	4	3	7	9	7
	13	1	4	3	4	3	8	8	6
	14	2	5	3	3	3	9	10	9
	15	1	2	3	3	1	9	10	9
	16	4	7	2	2	3	8	10	10
	17	8	6	4	5	3	7	10	5
	18	4	1	3	2	3	6	10	5
	19	4	3	4	3	3	6	10	7
	20	3	6	5	4	4	10	11	7
	Average (regular zones)	2.9	4.6	3.3	3.7	3.6	8.3	9.4	6.1

The explanation for the findings above could be established from the "Theoretical Optimum" column in *TABLE 29*. At the top level, the load condition may be well represented by a smaller group of weather stations using the Vanilla model. (Moreno-Carbonell et al., 2019) also indicated that some temperature series in both GEFCom2012 and GEFCom2014 data are highly correlated and selecting highly correlated series can cause the over-parameterized combination of stations. In which case, The TO selects the most parsimonious combination which gives the lowest MAPE.

The benchmark (GS-PS) selects all 11 stations, while the FS and BS both select fewer weather stations and thus lead to better forecasting accuracy. When comparing FS to BS, we can look closer into the specific stations being selected by each method in *TABLE 30*. The BS-CV, BS-PS, BS-IS, and FS-IS lead to very close MAPE (%) and they all pick the station 5, which is part of the TO selection as well. Due to the overfitting issue on the validation period, the FS-PS fails to pick the station 5 and thus leads to a worse result. FS-CV improves the selection with station 5 included, while it misses the station 9 compared to BS-CV and BS-PS, which is likely due to the lack of training data in our case study.

At the bottom level, the TO selects fewer than 4 weather stations under the majority of the regular zones. Since the regular zones are the sub-regions of the aggregated zone, intuitively, the weather impacts on the load patterns cross these sub-regions can be well captured by smaller groups of weather stations. FS-IS tends to select fewer weather stations and thus leads to good forecasting accuracy among the regular zones. The FS-PS method selects fewer stations as well. Likely due to the overfitting issue on the validation period, the WSS out of FS-PS leads to worse forecasting performance than that of FS-IS. The FS-

CV improves the selection on top of FS-PS, but still is not as good as FS-IS, which is likely due to the lack of training data in our case study.

On the other side, both backward selection methods tend to select more weather stations among the regular zones and thus lead to worse results.

The GEFCom2012 case study may suggest a rule of thumb that that under the context of hierarchical load forecasting, one may prefer to use BS methods to forecast the top-level loads and use FS or GS methods to forecast the bottom-level loads.

TABLE 30: Weather stations selected for the aggregated zone (GEFCom2012)

Method	Station	MAPE (%)
TO	{2, 5, 7}	4.854
BS-CV	{1, 2, 5, 6, 7, 9}	4.924
BS-PS	{1, 2, 5, 6, 7, 9}	4.924
BS-IS	{3, 5, 6, 7, 9, 10, 11}	4.945
FS-IS	{2, 3, 5, 6, 7, 9}	4.975
FS-CV	{1, 2, 5, 6, 7}	5.072
FS-PS	{1, 2, 7}	5.166
GS-PS (Benchmark)	{1 - 11}	5.221

For the GEFCom2014, the average station number selected by each method among the three test years is presented in *FIGURE 17*. The bar chart is sorted in ascending MAPE values, indicating the FS methods result in the best forecasting accuracy among the test years, the BS methods lead to the worst, and the performance of the benchmark (GS-PS) sits in the middle.

More details on the WSS are gathered in *TABLE 31*. The average MAPEs across the three test years among the three FS methods are all lower than the benchmark (green highlighted), while the average MAPEs among the three BS methods are all higher (filled in gray). Looking at the "Theoretical Optimum" column again, the TO selects the most parsimonious combination which gives the lowest MAPE. The FS methods tend to select

fewer weather stations and thus lead to better forecasting accuracy. The benchmark selects more stations while the BS methods tend to select the most and thus lead to worse results.

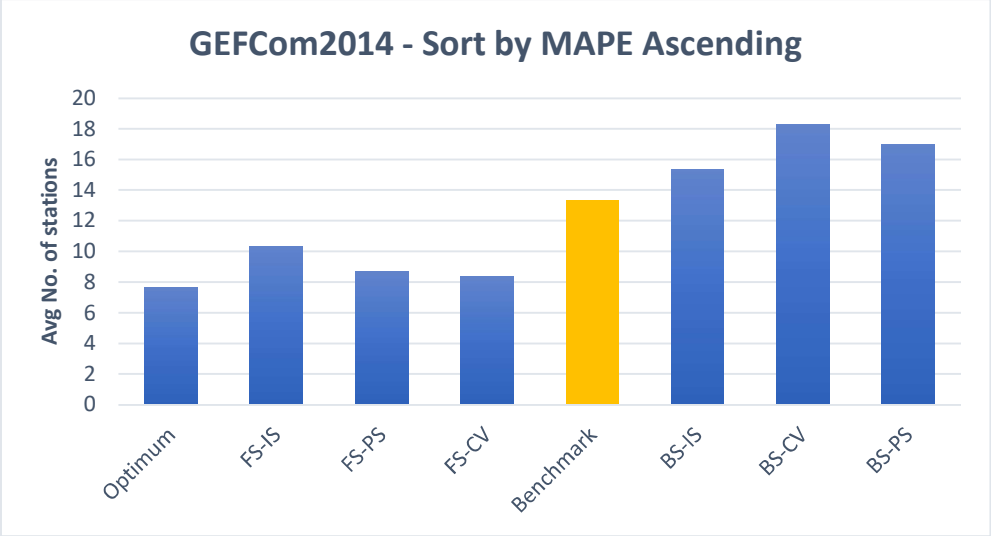

FIGURE 17: Bar Chart of Average Station No. Selected by Each Method Among the Test Years (GEFCom2014)

TABLE 31: No. of Stations Selected by TO, GS-PS (benchmark), FS and BS Methods (GEFCom2014)

Test Year	*Theoretical Optimum*	GS-PS	FS-PS	FS-CV	FS-IS	BS-PS	BS-CV	BS-IS
2008	9	12	11	10	10	18	12	14
2009	6	17	9	8	11	16	25	15
2010	8	11	6	7	10	17	18	17
Average	7.7	13.3	8.7	8.3	10.3	17.0	18.3	15.3

5.7. Comparison of Computational Cost

All computational experiments in the GEFCom2012 case study are performed using SAS (9.4) software on a personal laptop equipped with Intel Core i7 2.30 GHz CPU, 16GB usable RAM and Microsoft Windows 10 Professional. The computational experiments in the GEFCom2014 case study require much more computation resources

and are performed using SAS (9.4) software on a server with Intel Xeon E7 2.50 GHz CPU, 1TB usable RAM and Microsoft Windows Server 2012 R2 Datacenter.

The exhaustive search framework presented in this thesis goes through each possible WSS sequentially and records the selection based on the post-sample forecasting performance under the validation year (the PS method), across multiple validation years (the CV method), or based on the in-sample fit performance (the IS method). The IS and PS methods need about the same computational cost, while the CV methods could triple the cost given the implementation of 3-fold cross-validation in our case studies. *TABLE 32* reports the computation time under the ES methods. The computation time for each load zone is in minutes for the GEFCom2012 case study using a personal laptop, and in days for each test year in the GEFCom2014 case study using a Windows server. Noticeably, the ES framework is more applicable when we have a small amount of weather station candidates. As indicated in Section 3.4, one extra weather station candidate can almost double the computational cost of the selection process and as the number of station candidates grows, this framework tends to overburden the selection process.

TABLE 32: Computational Time under the ES framework

GEFCom2012 (*N*=11, on the laptop)		GEFCom2014 (*N*=25, on Windows server)	
Method	Minutes	Method	Days
ES-IS	9	ES-IS	4
ES-PS	9	ES-PS	4
ES-CV	27	ES-CV	12

Aside from the ES frameworks, the computation time required by the heuristic frameworks is much more manageable. Even under the case of 25 candidate stations, the selection processes can mostly be done within 4 mins using a personal laptop.

TABLE 33 and *TABLE 34* give the iterations required to evaluate the Vanilla model within each framework in the two case studies, sorted by their forecasting MAPEs. The benchmark method (GS-PS) and GS-IS require the fewest iterations and achieve workable results. GS-CV requires slightly more iterations and reaches very close results to the benchmark. Methods under ES require the most computational cost than the remaining approaches. ES-IS and ES-CV almost guarantee good results, while ES-PS is not as good due to the overfitting issue we have covered in Section 5.3.

TABLE 33: Computational Cost (in iterations) sorted by MAPE Ranking (GEFCom2012)

	Average of 18 Regular Zones			Aggregated Zone (Z21)		
Rank	Average MAPE (%)		Iterations (per zone)	MAPE (%)		Iterations
1	ES-IS	6.962	2047	BS-CV	4.924	165 (max)
2	GS-IS	6.967	22	BS-PS	4.924	55 (max)
3	FS-IS	6.980	55 (max)	BS-IS	4.945	55 (max)
4	GS-PS	6.996	22	ES-IS	4.968	2047
5	ES-CV	7.003	6141	FS-IS	4.975	55 (max)
6	GS-CV	7.003	44	ES-CV	5.027	6141
7	FS-CV	7.017	165 (max)	GS-IS	5.029	22
8	FS-PS	7.116	55 (max)	FS-CV	5.072	165 (max)
9	ES-PS	7.118	2047	ES-PS	5.166	2047
10	BS-IS	7.196	55 (max)	FS-PS	5.166	55 (max)
11	BS-CV	7.471	165 (max)	GS-CV	5.221	44
12	BS-PS	7.487	55 (max)	GS-PS	5.221	22

TABLE 34: Computational Cost (in iterations) sorted by MAPE Ranking (GEFCom2014)

Rank	Framework	MAPE (%)	Iterations
1	ES-IS	5.677	33,554,431
1	FS-IS	5.677	300 (max)
3	ES-CV	5.747	110,663,293
4	ES-PS	5.752	33,554,431
5	FS-PS	5.752	300 (max)
6	FS-CV	5.884	900 (max)
7	GS-PS	5.944	50
8	GS-CV	5.951	100
9	GS-IS	6.021	50
10	BS-IS	6.136	300 (max)
11	BS-CV	6.307	900 (max)
12	BS-PS	6.366	300 (max)

For GEFCom2012 case study, the ES-IS ranks top on the average forecasting performance at the bottom level and the 4th place at the top level (Z_{21}). In the GEFCom2014 case study, ES-IS ranks top again based on the average MAPE among test years. In which case, however, we have 25 candidate stations to choose from and it is unrealistic for a load forecaster to implement the ES framework because of the huge computational cost.

It is worth noting that in both case studies, the FS-IS requires reasonable computational cost and achieves good forecasting accuracy. In the GEFCom2014 case study, the FS-IS method leads to results as good as the ES-IS with a tiny fractional computational cost.

In summary, the above findings may suggest a rule of thumb that the ES-IS is approachable when we have fewer candidate stations; when dealing with more stations, the FS-IS stands out and should be considered for its computational simplicity.

CHAPTER 6: CONCLUSION

Weather factors are playing key roles to impact electricity load consumption. The selection of weather stations determines the key input to the electric load forecasting models which are reliant on weather variables. This thesis presents a comprehensive exhaustive search framework and three heuristic frameworks – forward selection framework, backward selection framework, and greedy selection framework, to solve the WSS problem. Under each framework, three types of statistical tests are performed and compared, namely the in-sample fit, the post-sample fit, and the out-of-sample cross validation test. We conduct case studies using GEFCom2012 and GEFCom2014 data and compare the results based on the MAPE and the distance to the theoretical optimum. Our experimental results show that the forecasting accuracy can be significantly improved by several proposed selection frameworks. Meanwhile, several heuristic methods have been applied to cut down the computational cost. Finally, we extend our discussion on several practical data fitting issues on the WSS subject and suggest actionable rules of thumb that load forecasting practitioners can follow.

The contribution made by this thesis to the WSS of electric load forecasting literature is obvious: (1) this is the first time that the exhaustive search method has been explored and its effectiveness has been evaluated; (2) this is the first time that the theoretical optimum selection is introduced to unveil important insight on why some WSS frameworks outperform the others; (3) this is the first time that a group of heuristic methods are implemented and compared on their selection behavior with transparency; (4) it covers the first formal comparison and extensive discussion on the model selection methods in terms of in-sample fit, post-sample fit and cross-validation on the WSS subject and leads

to actionable rules of thumb; (5) publicly available data in our case study and transparent implementations allow future researchers to reproduce our results.

The below bullet points summarize the key takeaways from this thesis work:

- Among the 12 methods presented in this thesis, the ES-IS and FS-IS show superior forecasting accuracy in both case studies comparing to the WSS benchmark. The ES-IS is approachable when we have fewer candidate stations; when dealing with more stations, the FS-IS stands out and should be considered for its computational simplicity.
- Among the three types of statistical tests, PS methods may lead to overfitting issues on the validation data and results have shown the CV methods help to remediate the overfitting issue discovered under the ES and BS frameworks. Given limited data history, the IS based methods show better forecasting performance than the PS methods under the Vanilla model. Further comparison shows the IS based methods, in general, outperform the CV based methods.
- An extensive discussion on the WSS outcome of FS and BS methods has been made and results may suggest that under the context of hierarchical load forecasting, one may prefer the BS methods for the prediction of the top-level loads, while FS and GS methods are recommended for the prediction of the bottom-level loads.
- Regarding the computational cost, GS-PS and GS-IS require the least computational time and achieve workable forecasting accuracy. The FS-IS heuristic method requires reasonable computational cost and achieves fairly good forecasting performance.

This thesis uses simple averages to combine temperature series and create virtual weather stations. As extra weather station combination methods have been discussed in (Sobhani et al., 2019) and (Moreno-Carbonell et al., 2019), further research directions may combine WSS methods in this thesis along with the combination techniques to refine the multiple weather station solutions. Aside from the above direction, further research may also incorporate other factors such as anomaly detection of weather station data and location information into the selection method to improve the accuracy, transparency and interpretability of the WSS.

REFERENCES

1. Andersen, F. M., Baldini, M., Hansen, L. G., & Jensen, C. L. (2017). Households' hourly electricity consumption and peak demand in Denmark. *Applied Energy*, *208*, 607–619. https://doi.org/10.1016/j.apenergy.2017.09.094

2. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(0), 40–79. https://doi.org/10.1214/09-SS054

3. Ben Taieb, S., Huser, R., Hyndman, R. J., & Genton, M. G. (2016). Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, *7*(5), 2448–2455. https://doi.org/10.1109/TSG.2016.2527820

4. Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394. https://doi.org/10.1016/j.ijforecast.2013.07.005

5. Charlton, N., & Singleton, C. (2014). A refined parametric model for short term load forecasting. *International Journal of Forecasting*, *30*(2), 364–368. https://doi.org/10.1016/j.ijforecast.2013.07.003

6. Chen, B. J., Chang, M. W., & Lin, C. J. (2004). Load forecasting using support vector machines: A study on EUNITE Competition 2001. *IEEE Transactions on Power Systems*, *19*(4), 1821–1830. https://doi.org/10.1109/TPWRS.2004.835679

7. Chen, K., Chen, K., Wang, Q., He, Z., Hu, J., & He, J. (2019). Short-Term Load Forecasting with Deep Residual Networks. *IEEE Transactions on Smart Grid*, *10*(4), 3943–3952. https://doi.org/10.1109/TSG.2018.2844307

8. Chen, Y., Luh, P. B., Guan, C., Zhao, Y., Michel, L. D., Coolbeth, M. A., Friedland, P. B., & Rourke, S. J. (2010). Short-term load forecasting: Similar day-based wavelet neural networks. *IEEE Transactions on Power Systems*, *25*(1), 322–330. https://doi.org/10.1109/TPWRS.2009.2030426

9. Dordonnat, V., Pichavant, A., & Pierrot, A. (2016). GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models. *International Journal of Forecasting*, *32*(3), 1005–1011. https://doi.org/10.1016/j.ijforecast.2015.11.010

10. Fahad, M. U., & Arbab, N. (2014). Factor Affecting Short Term Load Forecasting. *Journal of Clean Energy Technologies*, *2*(4), 305–309. https://doi.org/10.7763/jocet.2014.v2.145

11. Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, *27*(1), 134–141. https://doi.org/10.1109/TPWRS.2011.2162082

12. Fan, S., Methaprayoon, K., & Lee, W. J. (2009). Multiregion load forecasting for system with large geographical area. *IEEE Transactions on Industry Applications*, *45*(4), 1452–1459. https://doi.org/10.1109/TIA.2009.2023569

13. Gaillard, P., Goude, Y., & Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, *32*(3), 1038–1050. https://doi.org/10.1016/j.ijforecast.2015.12.001

14. Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*(2), 411–447. https://doi.org/10.1007/s12667-016-0203-y

15. Ghedamsi, R., Settou, N., Gouareh, A., Khamouli, A., Saifi, N., Recioui, B., & Dokkar, B. (2016). Modeling and forecasting energy consumption for residential buildings in Algeria using bottom-up approach. *Energy and Buildings*, *121*, 309–317. https://doi.org/10.1016/j.enbuild.2015.12.030

16. Haben, S., & Giasemidis, G. (2016). A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting. *International Journal of Forecasting*, *32*(3), 1017–1022. https://doi.org/10.1016/j.ijforecast.2015.11.004

17. Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55. https://doi.org/10.1109/59.910780

18. Hong, T. (2010). Short Term Electric Load Forecasting dissertation. *3442639*, 175. https://doi.org/10.1017/CBO9781107415324.004

19. Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011

20. Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363. https://doi.org/10.1016/j.ijforecast.2013.07.001

21. Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913. https://doi.org/10.1016/j.ijforecast.2016.02.001

22. Hong, T., & Wang, P. (2014). Fuzzy interaction regression for short term load forecasting. *Fuzzy Optimization and Decision Making*, *13*(1), 91–103. https://doi.org/10.1007/s10700-013-9166-9

23. Hong, T., Wang, P., & White, L. (2015). Weather station selection for electric

load forecasting. *International Journal of Forecasting*, *31*(2), 286–295. https://doi.org/10.1016/j.ijforecast.2014.07.001

24. Hong, T., Wang, P., & Willis, H. L. (2011). A naïve multiple linear regression benchmark for short term load forecasting. *IEEE Power and Energy Society General Meeting*, 1–6. https://doi.org/10.1109/PES.2011.6038881

25. Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.02.006

26. Islam, B., Baharudin, Z., & Nallagownden, P. (2017). Modified meta heuristics and improved backpropagation neural network-based electrical load demand prediction technique for smart grid. *IEEJ Transactions on Electrical and Electronic Engineering*, *12*, S20–S32. https://doi.org/10.1002/tee.22420

27. Kanda, I., & Veguillas, J. M. Q. (2019). Data preprocessing and quantile regression for probabilistic load forecasting in the GEFCom2017 final match. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.02.005

28. Khuntia, S. R., Rueda, J. L., & van der Meijden, M. A. M. M. (2016). Forecasting the load of electrical power systems in mid- and long-term horizons: A review. *IET Generation, Transmission and Distribution*, *10*(16), 3971–3977. https://doi.org/10.1049/iet-gtd.2016.0340

29. Lai, S., & Hong, T. (2013). When one size no longer fits all - electric load forecasting with a geographic hierarchy. *SAS*, 1–14. http://assets.fiercemarkets.com/public/sites/energy/reports/electricloadforecasting.pdf

30. Laković, M., Pavlović, I., Banjac, M., Jović, M., & Mančić, M. (2017). Numerical computation and prediction of electricity consumption in tobacco industry. *Facta Universitatis, Series: Mechanical Engineering*, *15*(3), 457–465. https://doi.org/10.22190/FUME170927025L

31. Li, P., Li, Y., Xiong, Q., Chai, Y., & Zhang, Y. (2014). Application of a hybrid quantized Elman neural network in short-term load forecasting. *International Journal of Electrical Power and Energy Systems*, *55*, 749–759. https://doi.org/10.1016/j.ijepes.2013.10.020

32. Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2017). Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts. *IEEE Transactions on Smart Grid*, *8*(2), 730–737. https://doi.org/10.1109/TSG.2015.2437877

33. Lloyd, J. R. (2014). GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *International Journal of Forecasting*,

30(2), 369–374. https://doi.org/10.1016/j.ijforecast.2013.07.002

34. Luo, J., Hong, T., & Yue, M. (2018). Real-time anomaly detection for very short-term load forecasting. *Journal of Modern Power Systems and Clean Energy*, *6*(2), 235–243. https://doi.org/10.1007/s40565-017-0351-7

35. Mangalova, E., & Shesterneva, O. (2016). Sequence of nonparametric models for GEFCom2014 probabilistic electric load forecasting. *International Journal of Forecasting*, *32*(3), 1023–1028. https://doi.org/10.1016/j.ijforecast.2015.11.001

36. Moreno-Carbonell, S., Sánchez-Úbeda, E. F., & Muñoz, A. (2019). Rethinking weather station selection for electric load forecasting using genetic algorithms. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.08.008

37. Nedellec, R., Cugliari, J., & Goude, Y. (2014). GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, *30*(2), 375–381. https://doi.org/10.1016/j.ijforecast.2013.07.004

38. Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*, 1535–1547. https://doi.org/10.1109/59.99410

39. Senjyu, T., Mandal, P., Uezato, K., & Funabashi, T. (2005). Next day load curve forecasting using hybrid correction method. *IEEE Transactions on Power Systems*, *20*(1), 102–109. https://doi.org/10.1109/TPWRS.2004.831256

40. Senjyu, T., Takara, H., Uezato, K., & Funabashi, T. (2002). One-hour-ahead load forecasting using neural network. *IEEE Transactions on Power Systems*, *17*(1), 113–118. https://doi.org/10.1109/59.982201

41. Sobhani, M., Campbell, A., Sangamwar, S., Li, C., & Hong, T. (2019). Combining weather stations for electric load forecasting. *Energies*, *12*(8), 1510. https://doi.org/10.3390/en12081510

42. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

43. Taylor, J. W. (2008). An evaluation of methods for very short-term load forecasting using minute-by-minute British data. *International Journal of Forecasting*, *24*(4), 645–658. https://doi.org/10.1016/j.ijforecast.2008.07.007

44. Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting*, *32*(3), 585–597. https://doi.org/10.1016/j.ijforecast.2015.09.006

45. Wang, Y., Chen, Q., Hong, T., & Kang, C. (2019). Review of Smart Meter Data

Analytics: Applications, Methodologies, and Challenges. *IEEE Transactions on Smart Grid*, *10*(3), 3125–3148. https://doi.org/10.1109/TSG.2018.2818167

46. Weron, R. (2006). Modeling and forecasting electricity loads and prices: A statistical approach. In *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. John Wiley & Sons. https://doi.org/10.1002/9781118673362

47. Xie, J., Chen, Y., Hong, T., & Laing, T. D. (2018). Relative humidity for load forecasting models. *IEEE Transactions on Smart Grid*, *9*(1), 191–198. https://doi.org/10.1109/TSG.2016.2547964

48. Xie, J., & Hong, T. (2016). GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, *32*(3), 1012–1016. https://doi.org/10.1016/j.ijforecast.2015.11.005

49. Xie, J., & Hong, T. (2017). Wind speed for load forecasting models. *Sustainability (Switzerland)*, *9*(5), 795. https://doi.org/10.3390/su9050795

50. Xie, J., & Hong, T. (2018). Load forecasting using 24 solar terms. *Journal of Modern Power Systems and Clean Energy*, *6*(2), 208–214. https://doi.org/10.1007/s40565-017-0374-0

51. Xie, J., Hong, T., Laing, T., & Kang, C. (2017). On Normality Assumption in Residual Simulation for Probabilistic Load Forecasting. *IEEE Transactions on Smart Grid*, *8*(3), 1046–1053. https://doi.org/10.1109/TSG.2015.2447007

52. Xie, J., Hong, T., & Stroud, J. (2015). Long-term retail energy forecasting with consideration of residential customer attrition. *IEEE Transactions on Smart Grid*, *6*(5), 2245–2252. https://doi.org/10.1109/TSG.2014.2388078

53. Zhang, P., Wu, X., Wang, X., & Bi, S. (2015). Short-term load forecasting based on big data technologies. *CSEE Journal of Power and Energy Systems*, *1*(3), 59–67. https://doi.org/10.17775/cseejpes.2015.00036

54. Zheng, H., Yuan, J., & Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation. *Energies*, *10*(8). https://doi.org/10.3390/en10081168

55. Ziel, F., & Liu, B. (2016). Lasso estimation for GEFCom2014 probabilistic electric load forecasting. *International Journal of Forecasting*, *32*(3), 1029–1037. https://doi.org/10.1016/j.ijforecast.2016.01.001

56. Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. otexts.com/fpp2/regression-intro.html. Accessed on 1/28/2020.